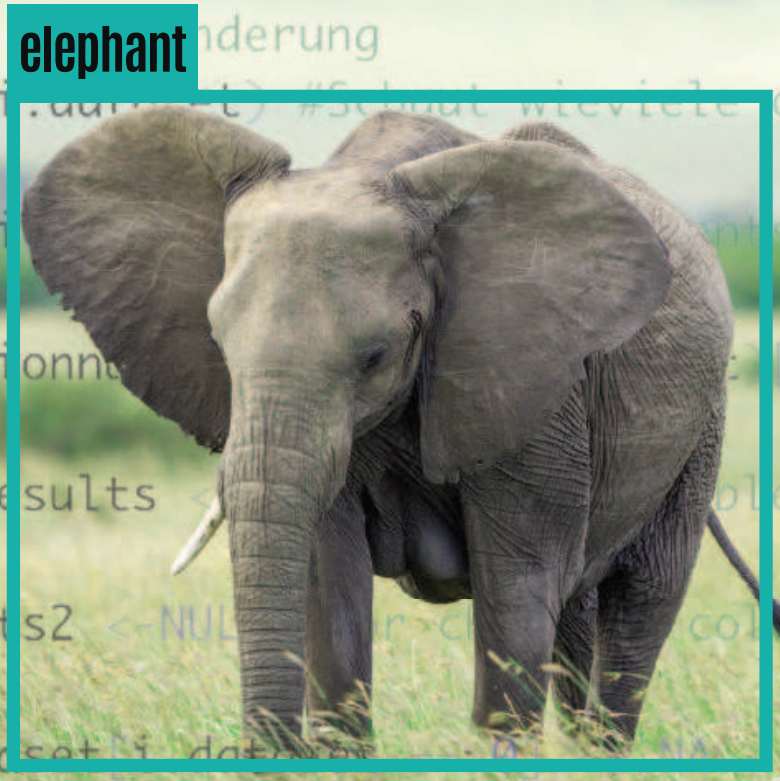
Katarzyna Biernacka und Sandra Schulz

FORSCHUNGSDATENMANAGEMENT IN DER INFORMATIK



λογος

Forschungsdatenmanagement in der Informatik

Katarzyna Biernacka Sandra Schulz

Katarzyna Biernacka (Humboldt-Universität zu Berlin) 0000-0002-6363-0064
Sandra Schulz (Universität Hamburg) 0000-0002-2254-6579

Forschungsdatenmanagement in der Informatik
1. Auflage
Erschienen 2022

Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10,
12681 Berlin
Tel.: +49 030 42 85 10 90
Fax: +49 030 42 85 10 92
INTERNET: http://www.logos-verlag.de

ISBN PRINT: 978-3-8325-5490-3
DOI 10.30819/5490

Die Veröffentlichung wurde gefördert aus dem Open-Access-Publikationsfonds der Humboldt-Universität zu Berlin.

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

Zitationsvorschlag
Biernacka, K. & Schulz, S. (2022). *Forschungsdatenmanagement in der Informatik*. Logos Verlag Berlin. https://doi.org/10.30819/5490

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ArbnErfG	Arbeitnehmererfindergesetz
BDSG	Bundesdatenschutzgesetz
BMBF	Bundesministerium für Bildung und Forschung
CC	Creative Commons
DFG	Deutsche Forschungsgemeinschaft
DMP	Datenmanagementplan
DOI	Digital Object Identifier
DSGVO	Datenschutzgrundverordnung
E-A	Ein- und Ausatmen
ELB	Elektronische Laborbücher
ELN	Electronic Lab Notebooks
FDM	Forschungsdatenmanagement
FDR	Forschungsdatenrepositorium
GI	Gesellschaft für Informatik
GND	Gemeinsame Normdatei
HU	Humboldt-Universität zu Berlin
ISNI	International Standard Name Identifier
LZA	Langzeitarchivierung
ODC	Open Data Commons
OpARA	Open Access Repository and Archive
ORCID iD	Open Researcher and Contributor ID
PatG	Patentgesetz

PID	Persistenter Identifikator
SMP	Softwaremanagementplan
TDM	Text- und Data-Mining
TGN	Getty Thesaurus of Geographic Names
TN	Teilnehmende
TSM	Tivoli Storage Manager
TU Dresden	Technische Universität Dresden
UHH	Universität Hamburg
UrhG	Urheberrechtsgesetz
VIAF	Virtual International Authority File
VVT	Verzeichnis von Verarbeitungstätigkeiten
WL	Workshopleitende

Vorwort

Das Buch verfolgt das Ziel, Forschungsdatenmanagement (FDM) im Informatikstudium zu verankern bzw. zu stärken. Es soll insbesondere die Lücke geschlossen werden, dass FDM oftmals disziplinunabhängig beschrieben und geschult wird, somit jedoch ein großer Transferaufwand notwendig ist, um es in die Praxis zu bringen. Dieses Buch soll die benötigte Brücke in die Praxis der Informatik bauen. Dafür werden im ersten Teil Möglichkeiten der curricularen Verankerung von FDM im Informatikstudium beschrieben. Sie dienen insbesondere für Lehrende als Orientierung dafür, wo in bereits bestehenden Grundlagen- und Fortgeschrittenenmodulen das Thema Forschungsdatenmanagement optimal adressiert werden kann.

Im zweiten Teil des Buches werden theoretische Grundlagen des FDM dargelegt. Um sie auf die Informatik zuzuschneiden, werden drei Szenarien beschrieben, die Informatikstudierende bzw. Promovierende im Fach Informatik skizzieren. Es wurde versucht, ein breites Spektrum an möglichen Anwendungskontexten von FDM in der Informatik aufzugreifen. Zunächst wird ein Abriss über Inhalte des FDM vorgestellt. Anschließend werden die Inhalte auf die zuvor skizzierten drei Szenarien übertragen, um einen direkten Praxistransfer zu gewährleisten. Kapitel 4.2 orientiert sich strukturell und inhaltlich stark an dem Train-the-Trainer-Konzept zum Thema Forschungsdatenmanagement (Biernacka et al., 2021a), wobei der Fokus auf die in der Informatik üblichen Forschungsdaten über das existierende Konzept hinausgeht.

Der dritte Teil des Buches umfasst konkretes Lehrmaterial, das im Informatikstudium verwendet werden kann. Direkt nach den Arbeitsblättern, die gern genau in dieser Form benutzt oder auch verändert werden dürfen (alle Materialien stehen unter der Creative Commons-Lizenz CC BY 4.0 International https://creativecommons.org/licenses/by/4.0/de/legalcode unter der DOI 10.5281/zenodo.6512432 zur Verfügung), sind entsprechende Musterlösungen angegeben. Es sind ebenfalls Vorschläge enthalten, in welche Module sie eingebracht werden können.

Wir hoffen, mit diesem Buch den Weg für das Forschungsdatenmangement im Informatikstudium zu ebnen und den Lehrenden ausreichende Unterstützung an die Hand zu geben, damit sie ihre Studierenden in den Qualifikationsphasen auf eine gute wissenschaftliche Praxis vorbereiten können.

Katarzyna Biernacka und Sandra Schulz

Teil I

Analyse von Informatikcurricula

Inzwischen gibt es verschiedene Bestrebungen, Data Literacy verstärkt in die universitäre Lehre einzubringen, da Data Literacy-Fähigkeiten als wichtige Kompetenzen des 21. Jahrhunderts betrachtet werden (Yang & Li, 2020). Es ist besonders ratsam, eine curriculare Verankerung von FDM zu forcieren, damit die Inhalte immer disziplin- und anwendungsbezogen sind. Des Weiteren müssen Studierende zum Zeitpunkt der Bachelorarbeit bereits über Kenntnisse und Kompetenzen im Umgang mit Forschungsdaten verfügen. Wie eine curriculare Verankerung in der Informatik gelingen kann, wird in diesem Teil dargestellt. Zuerst wird auf die Verankerung im Bachelor- und Masterstudium anhand von konkreten Modulen eingegangen. Im Folgenden werden Integrationsmöglichkeiten für die Promotionsphase bzw. die Weiterbildung von wissenschaftlichen Mitarbeitenden thematisiert.

Über den universitären Kontext hinaus, ist auch davon auszugehen, dass innerhalb der nächsten Jahre viele existierende Berufe zu Auslaufmodellen werden. Der Grund dafür ist die zunehmende Digitalisierung. Als Konsequenz daraus werden jedoch auch neue Berufsbilder entstehen, und die Konstruktion und Wartung von Technologien wird dabei eine zunehmende Rolle spielen, ebenso wie die dafür benötigten Data Literacy-Kompetenzen. FDM, als Teil von Data Literacy, trägt auch zur Förderung von Kompetenzen über die Forschung hinaus bei. Somit ist das Erlangen von Data Literacy-Kompetenzen auch für die berufliche Zukunft der Studierenden relevant, die keine akademische Laufbahn anstreben.

Kapitel 1

Anschlussfähige Module im Studium

Es existieren verschiedene Empfehlungen, welche Inhalte in einem Informatikstudium behandelt werden sollten. Insbesondere seit der Bologna-Reform sollen Informatikstudiengänge verschiedener Hochschulen inhaltlich miteinander vergleichbar sein, damit ein Wechsel des Studienorts erleichtert wird und die Studierenden ähnliche Kompetenzen erlangen. In den folgenden Abschnitten werden Modulbeschreibungen der folgenden Universitäten exemplarisch aufgegriffen: Humboldt-Universität zu Berlin (HU), Universität Hamburg (UHH) und Technische Universität Dresden (TU Dresden). Die Auswahl der Universitäten erfolgte aufgrund einer Onlinerecherche. In dieser Recherche wurde nach Universitäten gesucht, die ihre Studien- und Prüfungsordnungen auf den Webseiten publizieren und einen ähnlichen Aufbau aufweisen, damit ein Vergleich möglich ist. Es wurde sich bewusst dafür entschieden, Studienordnungen zu betrachten, da Modulbeschreibungen im aktuellen Vorlesungsverzeichnis regelmäßig wechseln können und zum Teil keinen festen inhaltlichen Aufbau aufweisen.

Es ist zu beachten, dass in der Studienordnung der UHH ausschließlich die Lernergebnisse der Studierenden festgehalten sind. Welche Fähigkeiten und Kompetenzen sollen die Studierenden also am Ende des Moduls erlangt haben. Bei der HU werden hingegen die Lernergebnisse getrennt von Themen und Inhalten aufgeschlüsselt. Da die Beschreibung von Themen und Inhalten für die Analyse hinsichtlich der Anschlussfähigkeit zum FDM mehr Details offenbart, werden in Bezug auf die HU die Themen und Inhalte beleuchtet. An der TU Dresden wird von Inhalten und Qualifikationszielen gesprochen. Dabei wird einerseits der erreichte Kompetenzstand am Ende der Veranstaltung beschrieben, andererseits werden auch Inhalte aufgeschlüsselt.

Die Empfehlungen der Gesellschaft für Informatik (GI) für Bachelor- und Masterprogramme im Studienfach Informatik an Hochschulen (Gesellschaft für Informatik e. V., 2016) werden ebenfalls betrachtet. Dieses Papier ist als Orientierung für die Ausgestaltung von Informatikstudiengängen zu verstehen und ist explizit auch auf die zukünftige Entwicklung der Informatik ausgerichtet. Für die folgende Analyse sind insbesondere die beschriebenen Kompetenzen wichtig, die Studierende im Studium erlangen sollten. Diese sind auf 17 Inhaltsbereiche verteilt, die große Ähnlichkeiten zu den Modulbeschreibungen von Universitäten

und Hochschulen aufweisen. Die GI-Empfehlungen betrachten verschiedene kognitive Kompetenzen für unterschiedliche inhaltliche Felder der Informatik und gliedern jedes einzelne in sechs Kompetenzdimensionen von „Stufe 1: Verstehen“ über „Stufe 2: Anwenden“, „Stufe 2a: Übertragen“, „Stufe 3: Analysieren“, „Stufe 3a: Bewerten“ bis hin zu „Stufe 4: Erzeugen“. Das zugrundeliegende Kompetenzmodell umfasst noch weitere Facetten, die jedoch für das Ziel der Analyse von Modulen und Empfehlungen an dieser Stelle vernachlässigt werden.

Die folgende Analyse der Module ist als Vorschlag für die Integration von FDM-Themen in das Informatikstudium anhand von einigen Beispielen zu verstehen. Es handelt sich dabei nicht um eine abgeschlossene, bindende Aufzählung. Sofern in der Studienordnung inhaltliche Überschneidungen zum FDM festgestellt werden konnten, wird das jeweilige Modul aufgeführt. Sollten Module nicht aufgelistet sein, dann hängt das weniger damit zusammen, dass FDM nicht mit dem Modul verknüpft werden kann, sondern mit weniger detaillierten Beschreibungen von Inhalten in der Modulbeschreibung. Die konkrete Differenzierung innerhalb der Module in Vorlesungen, Seminaren, Übungen und Praktika wird nur berücksichtigt, wenn sie für FDM-Themen relevant ist.

1.1 Einführung in die Programmierung

Im Folgenden wird das Pflichtmodul „Einführung in die Programmierung“ betrachtet. Es handelt sich um ein Grundlagenmodul, welches in den meisten Studienordnungen der Informatik zu finden ist und Grundlagen in der Programmierung legt. Vertiefungen werden zumeist durch Module wie „Softwareentwicklung“, „Algorithmik“ oder „Algorithmen und Datenstrukturen“ vorgenommen. Die Schwerpunktsetzung ist jedoch bei den einzelnen Modulen unterschiedlich. Teilweise werden auch andere Bezeichnungen dafür vergeben, was sich auch auf die Vertiefungsmodule auswirken kann. Beispielsweise wird das Modul auch als „Softwareentwicklung“ bezeichnet und das Vertiefungsmodul als „Softwareentwicklung II“. Bei dieser Betrachtung ist jedoch wichtig, dass es sich für die Studierenden um die erste Berührung mit Themen der Programmierung handelt. Im Folgenden werden die Beschreibungen von Inhalten und Kompetenzen der Studierenden in diesem Modul an den unterschiedlichen Standorten betrachtet. Dabei werden zuerst die FDM-relevanten Inhalte extrahiert, Verweise im Buch für die inhaltliche Erläuterung seitens FDM-Themen dargestellt und es wird auf konkretes Lehrmaterial hingewiesen.

Modul Einführung in die Programmierung (HU)

Themen und Inhalte:

- „Grundlagen: Algorithmus, von-Neumann-Rechner, Programmierparadigmen- Konzepte imperativer Programmiersprachen: Grundsätzlicher Programmaufbau; Variablen: Datentypen, Wertzuweisungen, Ausdrücke, Sichtbarkeit, Lebensdauer; Anweisungen: Bedingte Ausf., Zyklen, Iteration; Methoden: Parameterübergabe; Rekursion
- Konzepte der Objektorientierung: Objekte, Klassen, Abstrakte Datentypen; Objekt -Variablen/-Methoden, Klassen
- Variablen/-Methoden; Werte und Referenztypen; Vererbung, Sichtbarkeit, Überladung, Polymorphie; dynamisches Binden; Ausnahmebehandlung; Oberflächenprogrammierung; Nebenläufigkeit
- Einführung in eine konkrete objektorientierte Sprache (z. B. JAVA): Grundaufbau eines Programms, Entwicklungsumgebungen, ausgewählte Klassen der Bibliothek, Programmierrichtlinien für eigene Klassen, Techniken zur Fehlersuche (Debugging)
- Einfache Datenstrukturen und Algorithmen: Listen, Stack, Mengen, Bäume, Sortieren und Suchen
- Softwareentwicklung: **Softwarelebenszyklus**, Software-Qualitätsmerkmale- Alternative Konzepte: Zeiger, maschinennahe Programmierung, alternative Modularisierungstechniken“ (Humboldt-Universität zu Berlin, 2015)

Im Bachelorstudiengang Informatik bietet beispielsweise die Universität Hamburg für das erste Semester das Modul „Softwareentwicklung“ an. Die Inhalte werden wie folgt beschrieben:

Modul Softwareentwicklung (UHH)

Lernergebnisse:

„Die Studierenden können sicher mit einem Rechner umgehen, beherrschen das grundlegende Handwerkszeug der Programmierung im Kleinen und sind in der Lage, Lösungen zu rechtfertigen. Sie können **Programmierwerkzeuge** wie Compiler und Editoren nutzen sowie deren Grenzen einschätzen. Sie verstehen die Konzepte der Programmierung über eine konkrete Programmiersprache hinaus, kennen grundlegende Datenstrukturen, haben einen ersten Eindruck vom Komplexitätsbegriff und können die Tragweite von Tests abschätzen.“ (Universität Hamburg, 2019)

Von den oben abgebildeten Modulbeschreibungen können die folgenden Inhaltsbereiche abgeleitet werden, die in einem Zusammenhang mit FDM gebracht werden können. Die zugehörigen Inhalte des FDM sowie zugehörigen Lehrmaterialien werden in Tabelle 1.1 zugeordnet.

Tabelle 1.1: Verankerung FDM im Modul „Einführung in die Programmierung“

Inhaltsbereich	**FDM-Verweis**	**Lehrmaterial**
Versionierung	Kapitel 4.9.3	Kapitel 5.8
Metadaten	Kapitel 4.11.3	Kapitel 5.10
Dokumentation	Kapitel 4.11	Kapitel 5.10
Ordnung und Struktur	Kapitel 4.9	Kapitel 5.8
Softwarelebenszyklus	Kapitel 4.2.2	Kapitel 5.1

1.2 Software Engineering

Das Modul „Software Engineering/Softwareentwicklung“ ist ebenfalls ein verpflichtendes Grundlagenmodul im Bachelorstudium, bei dem Programmiergrundlagen vertieft werden und die Entwicklung von Software in Teams im Fokus steht. Das Ziel ist zumeist, Software für die Wirtschaft entwickeln zu können. Standards der Softwareentwicklung und des Softwarelebenszyklus sind zentrale Bestandteile. Es wird wieder zuerst das Beispiel der HU betrachtet.

Modul Softwareengineering (HU)

Themen und Inhalte:

- „Methoden der systematischen Entwicklung komplexer Software
- Vorgehensmodelle und **Software-Entwicklungsstandards**
- Qualitätskriterien, Metriken und Aufwandsabschätzung
- **Anforderungsanalyse**: **Pflichtenheft** und Produktmodell
- Objektorientierte (UML) und strukturierte Analyse-Software-Architekturen, Entwurfsmuster und Modularisierung
- Einsatz formaler Methoden-Validierung, Verifikation und Test
- **Produktzyklen**, Weiterentwicklung und Reverse Engineering
- Konfigurationsmanagement und **Entwicklungswerkzeuge**
- Einführung in die Software-Ergonomie“ (Humboldt-Universität zu Berlin, 2015)

An der TU Dresden ist in der Studienordnung eine Vorlesung zur Softwaretechnologie sowie ein spezielles Softwaretechnologie-Praktikum vorgesehen.

Softwaretechnologie (TU Dresden)

Inhalte und Qualifikationsziele:

„Die Studierenden beherrschen die Methoden zur Entwicklung von Softwaresystemen. Damit sind Studierende in die Lage versetzt, eine systematische ingenieurtechnische Vorgehensweise unter Verwendung der Konzepte der Objektorientierung anzuwenden, insbesondere den Einsatz der Modellierungssprache Unified Modeling Language (UML) in Analyse, Entwurf und Implementierung zu beherrschen. Zur praktischen Umsetzung der Systeme beherrschen die Studierenden den gezielten Einsatz der Programmiersprache Java, mit besonderer Betonung der Verwendung von Klassenbibliotheken und Entwurfsmustern. Grundinformationen zum **Projektmanagement und der Software-Qualitätssicherung** runden die Inhalte ab.“ (Technische Universität Dresden, 2016)

Softwaretechnologie-Projekt (TU Dresden)

Inhalte und Qualifikationsziele:

„Die Studierenden besitzen praktische ingenieursmäßige Kenntnisse in der **Durchführung von arbeitsteiligen Softwareprojekten**. Die Studierenden sind in der Lage, in Zusammenarbeit mit einem Kunden dessen Anforderungen zu analysieren sowie arbeitsteilig ein Softwaresystem zu entwerfen, zu implementieren, zu testen und vom Kunden abnehmen zu lassen.“ (Technische Universität Dresden, 2016)

An der UHH handelt es sich laut Modulbeschreibung bereits um das Modul „Softwareentwicklung II“, da wie zuvor beschrieben, bereits mit dem Modul „Softwareentwicklung I“ als erstes Grundlagenmodul gestartet wird.

Softwareentwicklung II (UHH)

Lernerergebnisse:

„Die Studierenden beherrschen die Grundlagen zur **Entwicklung kleiner, gebrauchstauglicher Anwendungen** mit Hilfe objektorientierter Konzepte und kennen zentrale Konzepte zur Abstraktion und Modularisierung. Weiterhin sind sie vertraut mit fortgeschrittenen Programmiersprachkonzepten, den Paradigmen der objektorientierten und funktionalen Programmierung sowie mit Konzepten von Entwurfsmustern und Refactorings und können mit integrierten **Entwicklungsumgebungen** umgehen.“ (Universität Hamburg, 2019)

Die GI beschreibt kognitive Kompetenzen im Bereich „Software-Engineering“ wie folgt.

Software-Engineering (GI)

Kognitive Prozessdimension
Stufe 1 Verstehen:

- „Grundkonzepte und Grundtechniken der **Softwareerstellung im Großen und in Teams** mit Fachbegriffen erklären.
- Verschiedene Prozess-/Vorgehensmodelle wie z. B. das Wasserfallmodell und iterative Modelle voneinander abgrenzen.
- Verschiedene Notationen wie z. B. UML für die Modellierung von Softwaresysteme erläutern.
- Aufgaben und typische Vorgehensweisen beim **Management und der Qualitätssicherung von Softwareprojekten** erläutern.“ (Gesellschaft für Informatik e. V., 2016)

In Tabelle 1.2 wird die Zuordnung von Inhaltsbereichen der Informatik zu den Themen des FDM sowie dem dazugehörigen Lehrmaterial vorgenommen.

Tabelle 1.2: Verankerung FDM im Modul „Software-Engineering“

Inhaltsbereich	**FDM-Verweis**	**Lehrmaterial**
Versionierung	Kapitel 4.9.3	Kapitel 5.8
Metadaten	Kapitel 4.11.3	Kapitel 5.10
Dokumentation	Kapitel 4.11	Kapitel 5.10
Ordnung und Struktur	Kapitel 4.9	Kapitel 5.8
Softwarelebenszyklus	Kapitel 4.2.2	Kapitel 5.1
Management von (Forschungs-)Software	Kapitel 4.2.4	Kapitel 5.1
Softwarelizenzen	Kapitel 4.14.5	Kapitel 5.13
Softwaremanagementplan	Kapitel 4.6.2	Kapitel 5.5

1.3 Informatik im Kontext

Das Modul „Informatik im Kontext“ ist in verschiedenen Facetten im Informatikstudium verankert. Inhaltlich wird darauf fokussiert, Informatik im historischen und gesellschaftlichen Zusammenhang zu begreifen. An der HU ist „Informatik im Kontext“ als Schlüsselqualifikation in der Studienordnung verankert. Das Modul wird wie folgt beschrieben.

Modul Informatik im Kontext (HU)

Themen und Inhalte:

„In dieser Veranstaltung wird die Wissenschaft Informatik mit ihrer Position im Gesamtgefüge der Wissenschaften und in ihrer historischen Entwicklung beschrieben. Die Informatik wird in ihrem **ökonomischen, politischen und rechtlichen, aber auch sozialen und kulturellen Kontext** betrachtet und sich daraus ableitende **Fragestellungen für beruflich im Bereich Informatik** tätige Personen werden diskutiert.“ (Humboldt-Universität zu Berlin, 2015)

An der UHH handelt es sich um ein Pflichtmodul, dass im Bachelorstudium der Informatik verortet ist, jedoch verschiedenen Bereichen zugeordnet werden kann. An der UHH wird das Modul wie folgt beschrieben.

Modul Informatik im Kontext (UHH)

Lernergebnisse:

„Die Studierenden sind in der Lage zu erkennen, dass Einsatzkontexte **Anforderungen an die Entwicklung von Informatiksystemen** stellen und dort **Wirkungen** entfalten. Sie besitzen das dafür erforderliche Faktenwissen zur menschlichen Informationsverarbeitung und verfügen über exemplarische Kenntnisse unterschiedlicher Aspekte des **Einsatzes von Informations- und Kommunikationstechnologie** (IKT) für Menschen, Organisationen, Märkte und Gesellschaft. Sie erwerben Methodenwissen für die Analyse von Anwendungskontexten und die Gestaltung von Informatiksystemen. Auf dieser Grundlage können sie auch entstehende **Wechselwirkungen** bewerten. Sie verfügen über ein tieferes Verständnis der **Berufspraxis von InformatikerInnen**. Ferner sind sie in der Lage, ein **gesellschaftliches und ethisches Bewusstsein** aufzubauen.“(Universität Hamburg, 2019)

Zum Datenschutz und dem Urheberrecht gibt es teilweise noch Vertiefungsmodule, die jedoch im Wahlbereich enthalten sind und somit nicht von allen Studierenden durchlaufen werden.

In den GI-Empfehlungen wird der Kompetenzbereich als „Informatik und Gesellschaft“ benannt, der inhaltlich ähnlich gelagert ist. Die folgenden inhaltlichen Aspekte werden aufgelistet (Auswahl).

Informatik und Gesellschaft (GI)

Kognitive Prozessdimension
Stufe 1: Verstehen

- „Grundkonzepte des **Datenschutzrechts** erklären; Maßnahmen zum Schutz **personenbezogener Daten** erklären.
- Grundkonzepte des geistigen Eigentums (**UrhG**, **PatG**) und der **Open Culture** erklären.
- Grundkonzepte des Computerstrafrechts erklären.
- Grundzüge der Informationsökonomie und der daraus folgenden Implikationen von Informatiksystemen erklären.
- Grundzüge der **Informatik-Berufsethik** erläutern; die Konzepte Verantwortung, Wert, Dilemma erklären.“

In der 2. Stufe (Anwenden) wird u.a: weiter konkretisiert:

- „**Lizenzformen in Softwaresystemen** identifizieren.“

In der 3. Stufe (Analysieren) heißt es auch:

- „Wechselwirkungen zwischen **rechtlichen Rahmenbedingungen** und Informatiksystemen analysieren.“ (Gesellschaft für Informatik e. V., 2016)

In den vorgestellten Studienordnungen sind relativ allgemeine Ziele formuliert, die Studierende erreichen sollen. Beispielsweise sollen sie ein ethisches und gesellschaftliches Bewusstsein entwickeln. Bei den GI-Empfehlungen hingegen werden konkrete Kompetenzen aufgezeigt, die eine Rolle in den universitären Modulen spielen sollten. Die aus allen aufgezeigten Modulen/Empfehlungen gesammelten Überschneidungen zu FDM sind in Tabelle 1.3 zu finden.

Tabelle 1.3: Verankerung FDM im Modul „Informatik im Kontext“

Inhaltsbereich	**FDM-Verweis**	**Lehrmaterial**
Datenschutz	Kapitel 4.8	Kapitel 5.7
Urheberrecht	Kapitel 4.14	Kapitel 5.13
Weitere rechtliche Aspekte	Kapitel 4.17	Kapitel 5.16
Lizenzierung von Software	Kapitel 4.14	Kapitel 5.13
Ethische Aspekte	Kapitel 4.7	Kapitel 5.6
Open Source	Verweis Literatur: Galetzka et al. (2021)	-
Open Access	Verweis Literatur: Biernacka et al. (2022)	-

1.4 Medieninformatik

Als Teilgebiet der Informatik wird hier auch die Medieninformatik betrachtet. Ein wichtiger Aspekt der Medieninformatik ist die digitale Verarbeitung von Texten, Bildern, Audio und Videos. An der TU Dresden wird das Modul „Einführung in die Medieninformatik“ mit der folgenden Modulbeschreibung angeboten.

Einführung in die Medieninformatik (TU Dresden)

Inhalte und Qualifikationsziele:

„Die Studierenden sind mit **grundlegenden Problemkreisen, die bei der Verarbeitung von digitalen Medien** mit dem Schwerpunkt auf audiovisuellen und dreidimensionalen Medien eine Rolle spielen, vertraut. Ausgehend von den physikalischen Reizen Schall und Licht können sie den Wahrnehmungsapparat des Menschen analysieren und so eine wahrnehmungsspezifische Digitalisierung festlegen. Darauf aufbauend kennen sie **digitale Repräsentationen und Speicherformate** der Medien sowie grundlegende Verfahren zur **Verarbeitung digitaler Medien**. Mit diesen Grundvoraussetzungen für die Behandlung multimedialer Dokumente besitzen die Studierenden notwendige Kompetenzen im Einsatz von digitalen Medien, die sie bei der praktischen Umsetzung in Form eines Projektes anwenden können.“ (Technische Universität Dresden, 2016)

Klar erkennbare FDM-Themen in diesem Modul sind insbesondere „Sicherung und Back-up“ sowie „Langzeitarchivierung“, da Speicherformate ein expliziter Bestandteil sind. Es wird auch von „grundlegenden Problemkreisen, die bei der Verarbeitung von digitalen Medien [...] eine Rolle spielen“ gesprochen. Das eröffnet einen großen thematischen Rahmen für Themen wie beispielsweise „weitere rechtliche Aspekte“ und „ethische Aspekte“ (vgl. Tabelle 1.4).

Tabelle 1.4: Verankerung FDM im Modul „Medieninformatik“

Inhaltsbereich	**FDM-Verweis**	**Lehrmaterial**
Sicherung und Back-up	Kapitel 4.10	Kapitel 5.9
Metadaten	Kapitel 4.11.3	Kapitel 5.10
Langzeitarchivierung	Kapitel 4.15	Kapitel 5.14
Nachnutzung	Kapitel 4.16	Kapitel 5.15
Datenschutz	Kapitel 4.8	Kapitel 5.7
Urheberrecht und Lizenzierung	Kapitel 4.14	Kapitel 5.13
Weitere rechtliche Aspekte	Kapitel 4.17	Kapitel 5.16
Lizenzierung von Software	Kapitel 4.14	Kapitel 5.13
Ethische Aspekte	Kapitel 4.7	Kapitel 5.6

1.5 Künstliche Intelligenz

Das Gebiet der Künstlichen Intelligenz ist facettenreich und weist auch einige Überschneidungen zu FDM-Themen auf. Beispielsweise beim Maschinellen Lernen können Videodaten analysiert und gelabelt werden, um anschließend mithilfe verschiedener Lernverfahren automatisiert zu werden. Module zur Künstlichen Intelligenz sind in den bisher herangezogenen Studienordnungen noch nicht verankert. Trotzdem werden Module zu diesem Thema angeboten und können beispielsweise im Wahlpflichtbereich angerechnet werden. Wegen dieses Sonderfalls werden im Folgenden nicht Inhalte aus den Studienordnungen vorgestellt, sondern Modulbeschreibungen aus dem Vorlesungsverzeichnis genutzt. Diese wurden zuvor nicht verwendet, da sie nicht rechtlich bindend für zukünftige Module dieser Art sind (im Gegensatz zu Studienordnungen). Dafür ist es jedoch möglich, sehr konkret aktuelle Themen für das Modul zu benennen. Im Folgenden wird ein Beispiel aus dem Lehrangebot zum Maschinellen Lernen, einer Spezialform der Künstlichen Intelligenz, vorgestellt.

Intelligente Systeme (TU Dresden)

„Inhalte und Qualifikationsziele:
Die Studierenden beherrschen die grundlegenden Methoden der Künstlichen Intelligenz und besitzen Kompetenzen im Bereich der Anwendung von mathematischen Verfahren und Algorithmen. Dies sind insbesondere Problemlösungsverfahren (z. B. Suchverfahren), Wissenspräsentation (z. B. probabilistische Graphische Modelle), sowie **Lernverfahren** (z. B. Entscheidungsbäume). Mit den erlernten Fähigkeiten können sie verschiedenste Methoden der Künstlichen Intelligenz einsetzen und diese spezifizieren.“

Modul Maschinelles Lernen (UHH)

„Lernziel:

- Vertiefte Kenntnisse der verschiedenen Ansätze zum **Lernen aus Daten** auch im Hinblick auf ihre jeweiligen Beschränkungen
- Fähigkeit zur vergleichenden Bewertung von **Lernverfahren** im Hinblick auf spezifische Anwendungsbedingungen
- Fähigkeit zur systematischen Einordnung neuer Verfahren
- Fähigkeit zur Konzeption, Umsetzung und Evaluation eines lernenden Systems für eine gegebene Aufgabenstellung
- Fähigkeit zur Präsentation von empirischen Befunden im Bereich des algorithmischen Lernens“

Von der aufgezeigten Modulbeschreibung werden die folgenden Inhalte extrahiert und mit Themen des FDM in Beziehung gesetzt. Die Ergebnisse sind in Tabelle 1.5 dargestellt.

Tabelle 1.5: Verankerung FDM im Modul „Künstliche Intelligenz“

Inhaltsbereich	**FDM-Verweis**	**Lehrmaterial**
Metadaten	Kapitel 4.11.3	Kapitel 5.10
Ethische Aspekte	Kapitel 4.7	Kapitel 5.6
Weitere rechtliche Aspekte	Kapitel 4.17	Kapitel 5.16
Lizenzierung von Software	Kapitel 4.14	Kapitel 5.13

1.6 Ethik in der Informatik

Das Modul „Ethik in der Informatik“ ist im Moment in wenigen Modulbeschreibungen als fester Bestandteil des Curriculums bzw. Studienangebots zu finden. Oftmals wird es stattdessen im freien Wahlbereich verortet, sodass es nicht von alle Studierenden der Informatik besucht werden muss. Teilweise werden einige Aspekte des Moduls auch in dem Modul „Informatik im Kontext“ integriert (vgl. Modul 1.3), was zu einer Verkürzung der Inhalte und ausgebildeter Kompetenzen führen muss. An der UHH ist das Thema wie folgt als Modul implementiert.

Modul Ethik in der Informatik (UHH)

Lernergebnisse:
Die Studierenden

- „kennen die wesentlichen Theorien und Konzepte der Ethik, welche für die kritische Reflexion **ethischer Herausforderungen im Kontext der Nutzung und der Entwicklung von Informationstechnologien** notwendig sind
- kennen die wichtigen Themen der ethischen Diskussion um Informationstechnologien
- können das erworbene Wissen anwenden, um die mit Informationstechnologien in Bezug stehenden **ethischen Herausforderungen** zu analysieren und Antworten auf diese zu entwickeln.“ (Universität Hamburg, 2019)

In den GI-Empfehlungen wird kein Modul verortet, dass explizit auf Ethik in der Informatik abzielt. In dem Modul „Informatik und Gesellschaft“ wird jedoch an zwei Stellen auf die Berufsethik in der Informatik verwiesen. Eine Zuordnung von informatischen Inhalten und FDM ist für diesen Themenbereich in Tabelle 1.6 zu finden.

Tabelle 1.6: Verankerung FDM im Modul „Ethik in der Informatik“

Inhaltsbereich	**FDM-Verweis**	**Lehrmaterial**
Ethische Aspekte	Kapitel 4.7	Kapitel 5.6
Datenschutz	Kapitel 4.8	Kapitel 5.7
Urheberrecht	Kapitel 4.9.1	Kapitel 5.13
Metadaten	Kapitel 4.11.3	Kapitel 5.10
Softwarelizenzen	Kapitel 4.14.5	Kapitel 5.13

1.7 Robotik

Die Robotik ist ein stark anwendungsbezogenes Forschungsfeld, wo die Aufnahme von Reizen durch Sensoren, die Verarbeitung und die Einwirkung auf die Umgebung durch Aktuatoren im Vordergrund stehen. An der TU Dresden wird das Modul „Einführungspraktikum RoboLab“ im Bachelor angeboten. Eine Zuordnung von informatischen Inhalten und FDM ist für diesen Themenbereich in Tabelle 1.7 zu finden.

Einführungspraktikum RoboLab (TU Dresden)

Inhalte und Qualifikationsziele:
„Nach Abschluss des Moduls sind die Studierenden in der Lage praktische Aufgaben der Informatik zu lösen. Sie kennen Grundlagen der **Team- und Projektbearbeitung**, sowie Vortrags- und Präsentationstechniken. Die Studierenden sind in der Lage, praktische Aufgaben der Roboterprogrammierung im Team zu lösen und anschließend vorzustellen.“ (Technische Universität Dresden, 2016)

In der Modulbeschreibung werden vor allem Aspekte der Teamarbeit benannt, was die FDM-Themen „Ordnung und Struktur“ bzw. im Speziellen „Versionierung“ adressiert. Im Bereich der Robotik sind jedoch weitere FDM-Themen wie „Ethische Aspekte“ unverzichtbar. Beispielsweise ist die Roboterethik ein wichtiger Bereich der Robotik und lässt sich gut um weitere ethische Aspekte (insbesondere forschungsethische Fragen) erweitern und sollte eine Rolle spielen bei der Anforderungsanalyse an einen Roboter und die Implementation. Auch „Metadaten“ und „Personenbezogene Daten“ sind zu berücksichtigen, da diese Daten beispielsweise bei der Interaktion mit Menschen gesammelt und verarbeitet werden können. Identifizierte Überschneidungen zu FDM sind in Tabelle 1.7 dargestellt.

Tabelle 1.7: Verankerung FDM im Modul „Robotik“

Inhaltsbereich	**FDM-Verweis**	**Lehrmaterial**
Versionierung	Kapitel 4.9.3	Kapitel 5.8
Datenschutz	Kapitel 4.8	Kapitel 5.7
Metadaten	Kapitel 4.11.3	Kapitel 5.10
Ethische Aspekte	Kapitel 4.7	Kapitel 5.6

1.8 Softskills/Schlüsselqualifikationen

Je nach Aufbau des Studienangebots und der Studienordnung werden auch Module der Informatik im Bereich der „Softskills oder Schlüsselkompetenzen“ verortet. In den GI-Empfehlungen ist beispielsweise der Kompetenzbereich „Projekt- und Teamkompetenz“ zu finden, der insbesondere auf das gemeinsame Programmieren und Arbeiten in Projekten abzielt.

Projekt- und Teamkompetenz (GI)

Kognitive Prozessdimension
Stufe 1: Verstehen

- „Die Bedeutung grundsätzlicher Begriffe (Projektplan, Arbeitspaket, Abhängigkeiten zwischen Arbeitspaketen) des **Projektmanagements** erläutern.
- Die Artefakte (u.a. **Pflichtenheft**, Entwurf, **Handbuch**) und typischen Abläufe bei der Bearbeitung von IT-Projekten beschreiben und erläutern.
- Mechanismen zur **Qualitätssicherung** beschreiben und erläutern.“

Stufe 2: Anwenden

- „Arbeitspakete selbständig planen, termingerecht bearbeiten und **dokumentieren**.
- Mit einem **Repository zum Versionsmanagement** umgehen.
- Fremden Quelltext lesen, darin Entwurfskonzepte erkennen sowie Änderungen durchführen.“ (Gesellschaft für Informatik e. V., 2016)

Die Zuordnung von informatischen Inhalten zu konkreten Themen des FDM sowie zu Lehrmaterialien sind in Tabelle 1.8 zu finden.

Tabelle 1.8: Verankerung FDM im Modul „Softskills/Schlüsselkompetenzen“

Inhaltsbereich	**FDM-Verweis**	**Lehrmaterial**
Versionierung	Kapitel 4.9.3	Kapitel 5.8
Dokumentation	Kapitel 4.11	Kapitel 5.10
Datenmanagementplan	Kapitel 4.6	Kapitel 5.5
Softwaremanagementplan	Kapitel 4.6.2	Kapitel 5.5

Kapitel 2

Forschungsdatenmanagement in der Promotion

Doktorand:innen bilden eine besondere Zielgruppe in Bezug auf die Ausbildung zum Thema Forschungsdatenmanagement: Sie besuchen keine regelmäßigen Vorlesungen, befinden sich jedoch selbst schon mitten in der Forschung. Um diese Gruppe abzuholen, sind andere Lösungen und Formate notwendig. In diesem Kapitel werden verschiedene Optionen vorgestellt.

2.1 Graduiertenkollegs

Im Rahmen von Graduiertenkollegs werden Fähigkeiten im Bereich der Softskills sowie fachübergreifender Fertigkeiten vermittelt. Dazu können disziplinspezifische Workshops zum Thema Forschungsdatenmanagement für Doktorand:innen hinzugezählt werden, da FDM sich mittlerweile zu einer Kernkompetenz bei Forschenden etabliert hat. Es handelt sich dabei meist um ganztägige interaktive Workshops, in denen die Teilnehmenden aktiv das neu gewonnene Wissen anwenden und somit direkt auf ihre eigene Forschung übertragen können. Für die Durchführung der Workshops kann eine Zusammenarbeit mit den FDM-Koordinationsstellen der Hochschulen angestrebt oder externe Dienste in Anspruch genommen werden.

In den meisten Fällen werden die angebotenen Workshops zum Thema FDM generisch aufgebaut, um die Zielgruppe der Promovierenden aus verschiedenen Bereichen abzudecken. Es empfiehlt sich jedoch, auch angepasste Angebote zu erstellen. Das Train-the-Trainer-Konzept zum Thema Forschungsdatenmanagement von Biernacka et al. (2021a) bietet eine gute Basis, um einen disziplinspezifischen Workshop zu konzipieren. Die Tabelle 2.1 stellt ein potenzielles Lehrdrehbuch für diese Art von Veranstaltungen dar. Alle hier verwendeten Methoden, die Idee von *Ein- und Ausatmen (E-A)*[1], sowie das Konzept des *Stimmenklingens*[2] werden detail-

[1] Lernen ist laut Döring (2008) ein Prozess, der aus zwei Phasen besteht: „Einatmen“ und „Ausatmen“. Abwechselnd nehmen die Lernenden Wissen auf und geben es wieder. In dem vorgestellten Lehrdrehbuch wird das Aktivieren des Vorwissens auch als Ausatmen betrachtet, da es das Lernen erleichtert und das neue Wissen an das alte Wissen angeknüpft wird. Die rezeptive Phase (das Einatmen) sollte dabei nicht länger als 20 Minuten dauern.

[2] Das Konzept *Die Stimmen zum Klingen bringen* zielt darauf ab, die Teilnehmenden zum Sprechen zu bringen, um so einen besseren Austausch und die Übertragung des neu gewonnenen Wissens zu gewährleisten.

liert in dem Train-the-Trainer-Konzept zum Thema Forschungsdatenmanagement (Biernacka et al., 2021a) erläutert. Hervorgehoben sind die Elemente, die sich von dem generischen Ablauf unterscheiden und die fachspezifischen Bedürfnisse adressieren. Diese leiten sich direkt aus den Inhalten des Kapitels 4 und den Arbeitsmaterialien aus Kapitel 5 ab. Der vorgeschlagene Workshop ist für 12 Teilnehmende (TN) und bis zu zwei Workshopleitende (WL) als Präsenzveranstaltung konzipiert.

2.2 Weitere Angebote

2.2.1 Berufliche Weiterbildung der Forschungseinrichtungen

Viele Institutionen bieten für ihre Forschenden eine Reihe von Lehrveranstaltungen über berufliche Weiterbildungen an, die der Vertiefung und Erweiterung von Wissen dienen sollen. Diese Kurse können auch das Thema Forschungsdatenmanagement abdecken und von Promovierenden (meist kostenfrei) besucht werden. Häufig werden in diesem Fall diese Veranstaltungen zwar generisch gehalten, können den Forschenden jedoch eine gute Orientierung bieten.

2.2.2 Zertifikatskurse

Eine weitere Möglichkeit für Doktorand:innen, Kenntnisse im Bereich FDM aufzubauen, bieten unterschiedliche Zertifikatskurse. Diese strecken sich über eine längere Zeit und geben einen tieferen Einblick in die Aspekte des FDM als gewöhnliche Workshops. Darüber hinaus erlangt man nach dem erfolgreichen Abschluss ein Zertifikat.

Tabelle 2.1: Lehrdrehbuch für einen Informatik-spezifischen Workshop für Doktorand:innen ausgehend von Biernacka et al. (2021a).

Uhrzeit	**Thema**	**Baustein**	**Ziel der Phase/Lernziel**	**Dauer in Min.**	**Inhalt**	**Arbeitsform**	**Materialien**	**E-A**	**Stimmenklingen**
09:00 – 09:20	Begrüßen und Kennenlernen	Vorstellung	TN überwinden die Hemmschwelle des Sprechens	3	TN stellen sich vor	Zuruf	-	-	ja
		Begrüßung	TN lernen die WL kennen	1	WL begrüßen die TN und stellen sich vor	Vortrag	-	-	nein
		Kennenlernen	TN lernen sich gegenseitig kennen	8	Methode: Wir und ich	Gruppenarbeit	Flipchart, Stifte	-	ja
			TN sprechen vor der Gruppe	6	Gruppen stellen ihre Ergebnisse vor	Gruppenarbeit/ Forum	-	-	ja
09:20 – 09:25	Orientierung	-	TN erhalten einen Überblick über den Tag	5	WL stellt Seminarlandkarte vor (Schwäbischer Sparplan)	Vortrag	Seminarlandkarte	Ein	nein
09:25 – 10:10	Forschungsdatenmanagement	Forschungsdaten	TN aktivieren ihr Vorwissen	4	Methode: Frage-Ball; „Mit welchen Forschungsdaten arbeiten Sie?“	Aktivierende Übung	Weicher Ball	Aus	ja

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
			TN können den Begriff der digitalen Forschungsdaten in der Informatik erklären	2	Erklärung des Begriffs der digitalen Forschungsdaten in der Informatik	Vortrag	Folien	Ein	nein
		Forschungsdatenlebenszyklus	-	10	Methode: Drehen und Wenden	Gruppenarbeit	Vorbereitete Karten	Aus	ja
			-	5	TN stellen ihre Ergebnisse vor und begründen ihre Wahl	Gruppenarbeit/ Plenum	-	Aus	ja
			TN können den Forschunsdatenlebenszyklus erkennen	2	WL löst die Übung auf	Vortrag	Folien	Ein	nein
		Softwarelebenszyklus	TN können den Softwarelebenszyklus erklären; TN können zwischen Forschungsdatenlebenszyklus und Softwarelebenszyklus differenzieren	5	WL erläutert den Softwarelebenszyklus im Vergleich zum Forschunsdatenlebenszyklus	Vortrag	Folien	Ein	nein

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
		FDM	TN können den Begriff FDM erklären	2	WL erklärt den Begriff FDM	Vortrag	Folien	Ein	-
			TN können die Aspekte des FDM erklären	3	WL erläutert die Aspekte des FDM	Vortrag	Folien	Ein	nein
			TN können die FAIR-Prinzipien erklären	5	WL stellt die FAIR-Prinzipien vor	Vortrag	Folien; Handout 5.4	Ein	nein
			TN können die FAIR4RS-Prinzipien erklären	5	WL stellt die FAIR4RS-Prinzipien vor	Vortrag	Folien; Handout 5.4	Ein	nein
10:10 – 10:25	Forschungsdaten-Policy	Forschungsdaten-Policy	TN können die verschiedenen Policy-Typen benennen	5	Die verschiedenen Policy-Typen werden vorgestellt	Vortrag	Folien	Ein	nein
		Institutionelle Forschungsdaten-Policy	TN können die Anforderungen an den Umgang mit Forschungsdaten an ihrer Einrichtung beschreiben	5	TN lesen die Forschungsdaten-Policy ihrer Einrichtung (oder ein Beispiel einer anderen Einrichtung)	Einzelarbeit	Forschungsdaten-Policy der Einrichtung	Ein	nein

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
			-	5	TN tauschen sich über die Policy aus	Methode: Schnattern	-	Aus	ja
10:25 – 10:35					Kaffeepause				
10:35 – 10:45	Datenmanagementplan (DMP)	Definition	TN können den Begriff DMP definieren	2	Erläuterung des Begriffs DMP	Vortrag	Folien	Ein	nein
		Anforderungen der Forschungsförderer	TN können die externen Anforderungen der Forschungsförderer vergleichen	3	Tabellarischer Vergleich der wichtigsten Förderer in Deutschland	Vortrag	Folien	Ein	nein
		Bestandteile eines DMP	TN können die wichtigsten Bestandteile eines DMP benennen	3	Der Umfang und die Bestandteile eines DMP werden besprochen	Vortrag	Folien	Ein	nein
		DMP-Werkzeuge	TN können die unterschiedlichen Werkzeuge und Hilfestellungen benennen	2	Benennung der verschiedenen Werkzeuge (RDMO, DMPOnline, Checklisten)	Vortrag	Folien	Ein	nein

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
10:45 – 10:55	Softwaremanagementplan	Definition	TN können einen Softwaremanagementplan beschreiben	5	Besprechung des Softwaremanagementplans	Vortrag	Folien	Ein	nein
10:55 – 11:15	Ordnung und Struktur	Grundlagen	TN können den Begriff der Ordnerstruktur erklären; TN können grundlegende Verhaltensempfehlungen anwenden	3	Verhaltensempfehlungen werden anhand von Negativbeispielen aufgezeigt	Vortrag; Plenum	Folien	Ein und Aus	ja
		Dateibenennung	TN können Datenbenennung bewerten	3	Hinweise für gute Namenskonventionen werden vermittelt	Vortrag	Folien	Ein und Aus	ja
			TN können Werkzeuge zur gleichzeitigen Umbenennung von Dateien anwenden	2	Werkzeuge zur gleichzeitigen Umbenennung von Dateien werden vorgestellt	Vortrag	Folien	Ein	nein

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
		Versionierung	TN können Möglichkeiten der Versionierung beschreiben	1	Erläuterung der Notwendigkeit von Versionierung und der Hilfestellung bei der Versionskontrolle	Vortrag	Folien	Ein	nein
			TN können Versionierung mit Git anwenden	10	Git wird anhand einer praktischen Übung vorgestellt	Miniübung	Git-Account	Ein und Aus	ja
11:15 – 12:00	Dokumentation	Dokumentationsformen	TN können den Sinn von Dokumentation erklären	15	TN erarbeiten in Kleingruppen Tipps für die Verbesserung der Datendokumentation	Methode: Tippsuche	Vorbereitete Beispieldatentabelle	Aus	ja
			-	5	TN stellen die erarbeiteten Tipps der Gruppe vor		Whiteboard	Aus	ja

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
			TN können unterschiedliche Dokumentationsformen benennen	5	WL stellt die Dokumentationsformen vor	Vortrag	Folien	Ein	nein
			TN können Besonderheiten der Softwaredokumentation benennen	10	Besondere Formen der Softwaredokumentation werden vorgestellt	Vortrag; Methode: Zuruf	Folien	Ein und Aus	ja
		Metadaten	TN können den Begriff Metadaten erläutern	5	Erläuterung von Metadaten und kontrolliertem Vokabular	Vortrag	Folien	Ein	nein
		Metadatenstandards	TN können Metadatenstandards benennen	5	Erläuterung der Metadatenstandards	Vortrag	Folien	Ein	nein
			TN können die fachspezifischen Metadatenstandards benennen	5	TN suchen nach den fachspezifischen Metadatenstandards	Gruppenarbeit	Link zu bartoc.org	Aus	ja
12:00 – 13:00					Mittagspause				

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
13:00 – 13:10	Speicherung und Back-up	Speichermedien	TN können Vor- und Nachteile verschiedener Speichermedien und Serviceangebote vergleichen	5	Unterschiedliche Speichermedien werden miteinander verglichen und deren Vor- und Nachteile hervorgehoben	Einzelarbeit	Arbeitsblatt 5.9	Aus	ja
		Back-up	TN können Argumente für Back-ups benennen	2	Abschreckungsbeispiele werden den TN vorgestellt	Vortrag	Folien	Ein	nein
			TN können Strategien und Kriterien für ein sicheres Back-up benennen	3	Die Kriterien und Strategien für ein sicheres Back-up werden vorgestellt	Vortrag	Folien	Ein	ja
13:10 – 13:20	Langzeitarchivierung (LZA)	Grundlagen	-	3	Abgrenzung von Archivierung und Back-up	Methode: Zuruf	-	Aus	ja
			TN können den Begriff LZA erläutern	2	Erläuterung des Begriffs LZA	Vortrag	Folien	Ein	nein

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
		Nachhaltige Dateiformate	TN können geeignete Formate für die Archivierung der Daten benennen	5	Erläuterung des Unterschiedes zwischen offenen und proprietären Formaten	Vortrag	Folien	Ein	nein
13:20 – 14:10	Publikation	Publikationswege	TN können verschiedene Publikationswege benennen	5	Die drei Publikationswege für Forschungsdaten werden vorgestellt	Vortrag	Folien	Ein	nein
			-	10	TN suchen nach einem Repositorium auf re3data.org	Einzelarbeit	Notebooks; Internet	Aus	nein
			-	5	Besprechung der Ergebnisse	Plenum	-	Aus	ja
			TN können Kriterien für die Auswahl eines Repositoriums benennen	5	TN erarbeiten Kriterien für die Auswahl eines Repositoriums	Methode: Zuruf	Whiteboard	Aus	ja

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
		Lizenzen	TN können verschiedene Lizenzmodelle (CC/ODC) erklären	10	WL stellt die Lizenzmodelle vor	Vortrag	Folien	Ein	nein
			TN können offene Softwarelizenzen erklären	10	WL stellt offene Softwarelizenzen vor	Vortrag	Folien	Ein	nein
		Persistente Identifikatoren	TN können den Begriff DOI definieren	2	DOIs werden vorgestellt und deren Nutzen dargestellt	Vortrag	Folien	Ein	nein
			TN können den Begriff ORCID iD definieren	3	ORCID iD wird vorgestellt und TN ermutigt, eine anzulegen	Vortrag	Folien	Ein	nein
14:10 – 14:50	Nachnutzung	Recherchieren	TN können verschiedene Informationsquellen für das Recherchieren von Forschungsdaten angeben	2	Einführung und Vorstellung von Rechercheмöglichkeiten	Vortrag	Folien	Ein	ja
			-	5	TN suchen nach Forschungsdaten	Miniübung	Notebooks; Internet	Aus	nein

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
			-	5	TN tauschen sich über ihre Erfahrungen aus	Methode: Zuruf	-	Aus	ja
		Zitieren	TN können Angaben der Datenzitation bei verbreiteten Standards benennen	3	Vorstellung zweier Beispiele	Vortrag	Folien	Ein	nein
			TN können Datenzitationen entsprechend behandelter Standards anwenden	5	TN formulieren Zitationen	Einzelarbeit	Arbeitsblatt 5.15	Aus	nein
			TN können die Nachnutzungsregeln von Lizenzen anwenden	10	TN beantworten eine Umfrage zu den CC-Lizenzen	Einzelarbeit /Plenum	Umfrage bzw. Arbeitsblatt 5.13	Aus	ja
			-	10	TN wenden das Wissen zu Softwarelizenzen an	Einzelarbeit /Plenum	Arbeitsblatt 5.13	Aus	nein

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
14:50 – 15:30	Datenschutz	Einführung	TN können gesetzliche Regelungen benennen, die bei der Publikation von Forschungsdaten relevant sind	5	Übersicht über die verschiedenen Rechtsgebiete	Vortrag	Folien	Ein	nein
		Datenschutz	TN können den Begriff der personenbezogenen Daten definieren	10	TN lernen die Grundlagen der Datenschutzgrundverordnung (DSGVO)	Vortrag	Folien	Ein	nein
		Anonymisierung	TN können Anonymisierungsverfahren anwenden	15	TN führen eine Anonymisierung eines Textes durch	Gruppenarbeit	Arbeitsblatt 5.7	Aus	nein
			-	10	Besprechung der Ergebnisse	Methode: Zuruf	Musterlösung 5.7	Ein und Aus	ja
15:30 – 15:50	Zusammenfassung	DMP	TN können einen stichpunktartigen DMP entwickeln	20	TN schreiben einen DMP	Einzelarbeit	Vorlage 5.5	Aus	nein

Uhrzeit	Thema	Baustein	Ziel der Phase/Lernziel	Dauer in Min.	Inhalt	Arbeitsform	Materialien	E-A	Stimmenklingen
15:50 – 16:00	Abschluss	Verabschiedung	-	10	-	-	-	-	-

Teil II

Forschungsdatenmanagement anhand der Informatikszenarien

Kapitel 3

Personas und Szenarien aus Studium und Forschung

In diesem Abschnitt werden drei verschiedene Szenarien vorgestellt, an denen Forschungsdatenmanagement (FDM) erläutert und zum Leben erweckt wird. Diese Szenarien werden in Form von Personas formuliert, um sie besonders greifbar und übertragbar für Lesende zu machen. Es werden Szenarien gewählt, die sich an realistischen Projekten von Studierenden und Promovierenden orientieren. Die Namen und Szenarien sind jedoch frei erfunden. Es wurden bewusst unterschiedliche Szenarien gewählt, um eine große Breite des FDM in der Informatik abzudecken. Die Szenarien unterscheiden sich im Wesentlichen in den folgenden Punkten: die Personen

- befinden sich auf verschiedenen Abschlussniveaus (Bachelor, Master, Promotion),
- verwenden unterschiedliche Forschungsmethoden und erheben unterschiedliche Forschungsdaten,
- arbeiten entweder allein an ihrer Abschlussarbeit oder mit weiteren Personen,
- arbeiten innerhalb einer Disziplin oder interdisziplinär,
- arbeiten national oder international,
- arbeiten mit personenbezogenen Daten oder nicht.

Mit der zunehmenden Komplexität der Szenarien sind auch verschiedene Aspekte beim FDM zu berücksichtigen. Daraus folgt, dass auch die Beschreibung des Umgangs mit Forschungsdaten immer genauer wird.

Nachdem die drei Szenarien vorgestellt wurden, wird an dieser Stelle analysiert, welche Konsequenzen sich für das FDM ableiten lassen. Es wird zunächst von der Forschungsmethodik ausgegangen, um die benötigten Instrumente abzuleiten. Daraus folgt, welche konkreten Daten erhoben werden und wie sie verarbeitet und gespeichert werden sollen. An dieser Stelle wird grob betrachtet, welche Aspekte aus Sicht des FDM wichtig sind. Eine genauere Differenzierung folgt in den entsprechenden Kapiteln.

3.1 Szenario 1 (Interviewstudie)

Name	Carla (Ruiz)
Alter	21
Geschlecht	weiblich
Studiengang	Informatik
Angestrebter Abschluss	Bachelor
Universität	Universität Hamburg
Herkunft	deutsch; kubanische Vorfahren

Carla plant, ihre Bachelorarbeit im Bereich Human-Computer-Interaktion zum Thema Kollaborationstools zu schreiben. Sie interessiert sich insbesondere dafür, wie Studierende besser zusammenarbeiten können und inwieweit Studierende der Informatik für diese Tools aufgeschlossen sind. Kollaborationstools sind ein wichtiges Medium, um eine erfolgreiche Zusammenarbeit zu ermöglichen, indem auch Distanzen und zeitliche Hürden der Kollaborateur:innen überbrückt werden können. Häufig genutzte Tools für die Zusammenarbeit sind Moodle, Overleaf oder Git. Die Nutzung und Auswahl dieser Tools ist domänenspezifisch. Es ist naheliegend, dass Studierende der Informatik gut über aktuelle Tools informiert sind, diese auch gern für die Zusammenarbeit ausprobieren und gegebenenfalls nach Alternativen suchen. Da eine Zusammenarbeit der Studierende auch in gemeinsamen Arbeitsräumen (beispielsweise nach der Vorlesung) als direkten Austausch möglich wäre, sollen zunächst die Vor- und Nachteile der Zusammenarbeit mit digitalen Kollaborationstools erhoben werden. In späteren Untersuchungen kann davon abgeleitet werden, wie die Tools auf die Nutzenden angepasst werden sollten. In dieser Bachelorarbeit soll evaluiert werden, inwieweit die Tools tatsächlich einen Mehrwert für die Studierenden bieten.

Carla ist in Deutschland aufgewachsen, jedoch stammen ihre Eltern aus Kuba. Deswegen möchte sie ihre Bachelorarbeit nutzen, um gleichzeitig mehr über die Studiensituation in Kuba zu erfahren und möglicherweise erste Erkenntnisse zu erzeugen, um die Studiensituation in Kuba zu verbessern. Um dieses Ziel zu erreichen, hat sie sich vorgenommen die Ergebnisse von Studierenden ihrer Universität in Deutschland mit Studierenden der Informatik in Kuba zu vergleichen.

Carla formuliert zunächst die folgenden Forschungsfragen:

1. Welche Kollaborationstools nutzen Studierende der Informatik, um ihre Studienleistungen in Teams zu realisieren?

2. Welche Vor- und Nachteile des gemeinsamen Arbeitens mit Kollaborationstools beschreiben die Studierenden der Informatik?

3. Inwieweit unterscheiden sich die beschriebenen Vor- und Nachteile der Kollaborationstools bei Studierenden an einer deutschen und einer kubanischen Universität?

Carla entscheidet sich für die Durchführung einer Interviewstudie, um ihre Forschungsfragen beantworten zu können. Die Forschungsergebnisse sollen dabei nicht nur Gemeinsamkeiten

und Unterschiede aufzeigen, sondern auch einen Beitrag zur Entwicklungshilfe in Kuba leisten. Dafür sollen die Ergebnisse auf der Webseite des Lateinamerika-Forums veröffentlicht werden, und anderen Forschenden zur Verfügung stehen.

Bezug zum Forschungsdatenmanagement im ersten Szenario

Damit Carla die ersten beiden Forschungsfragen ihrer Bachelorarbeit erfolgreich beantworten kann, plant sie die Durchführung einer Interviewstudie mit Studierenden. Diese Interviews werden an der Universität Hamburg mit Studierenden der Universität durchgeführt. Die kubanischen Interviews verbindet sie mit dem Besuch ihrer Familie und führt auch diese in Kuba vor Ort durch. Dafür wird die Methode des leitfadengestüzten Interviews gewählt. Bei dieser Methode werden vorstrukturierte und standardisierte Fragen entwickelt, jedoch sind Abweichungen durchaus möglich, wenn die Teilnehmenden zum Beispiel besonders interessante Aspekte ansprechen. Um die gewählte Kohorte beschreiben zu können, muss sie auch personenbezogene Daten der zu interviewenden Personen erheben (z. B. Alter, Geschlecht, Studiensemester etc.). Aus diesem Grund wird eine **informierte Einwilligungserklärung** benötigt, die laut DSGVO die Voraussetzung für die Arbeit mit personenbezogenen Daten ist. Der **Fragebogen**, der die Grundlage für das leitfadengestützte Interview bildet, muss ebenfalls vorliegen. Somit wird transparent nachvollziehbar, welche konkreten Fragen gestellt wurden, ohne in einzelne Interviews schauen zu müssen.

Für die dritte Forschungsfrage soll ein Vergleich zu einer kubanischen Universität gezogen werden. Auch hier werden die gleichen Methoden wie für die ersten beiden Forschungsfragen verwendet. Es muss weiterhin die informierte Einwilligung eingeholt werden, da die Daten in Deutschland verarbeitet werden und somit der DSGVO unterliegen. Zusätzlich muss geprüft werden, welche kubanischen Richtlinien für die Datenerhebung existieren. Da die Ergebnisse auf der Webseite des Lateinamerika-Forums publiziert werden sollen, muss zusätzlich beachtet werden, wie die Ergebnisse kommuniziert werden. Es sollte insbesondere darauf geachtet werden, dass politische Vorgaben und mögliche Spannungsfelder berücksichtigt werden. Auch eine gründliche **Anonymisierung** der Daten ist entscheidend, damit beispielsweise die interviewten Personen keine schlechtere Bewertung ihrer Studienleistungen zu erwarten haben, weil sie zum Beispiel die Studienbedingungen als mangelhaft beschrieben haben.

3.2 Szenario 2 (SW-Entwicklung)

Name	Timo (Ryan)
Alter	25
Geschlecht	männlich
Studiengang	Informatik
Universität	TU Dresden
Angestrebter Abschluss	Master
Herkunft	irisch

Timo möchte seine Masterarbeit in der Informatik schreiben und hat sich dafür ein Projekt überlegt. Er hat das Open-Source-Projekt *Digital Public Health for All* gefunden und möchte es bei der Entwicklung unterstützen. Das Projekt hat das Ziel, Software im Gesundheitswesen zu entwickeln, um Patient:innendaten zu sichern und für Fachärzt:innen und Krankenkassen verfügbar zu machen. Es soll einerseits möglich sein, dass Patient:innen Zugriff auf ihre Daten erhalten und diese komplikationslos verschiedenen Fachärzt:innen digital zur Verfügung stellen können. Auch ein Wechsel zwischen Ärzt:innen soll damit erleichtert werden. Andererseits sollen die Krankenkassen angeschlossen werden, damit der Arztbesuch unkompliziert abgerechnet werden kann und den Patient:innen dargestellt wird, welche Leistungen des Arztes von der Krankenkasse übernommen werden können oder wo ggf. Selbstbeteiligungen notwendig sind. Da es sich bei Gesundheitsdaten um sensible Daten handelt, ist es besonders wichtig, eine transparente und sichere Software zu entwickeln. Da der Source Code offen ist, kann von Nutzenden die Verarbeitungsart ihrer Daten nachvollzogen werden. In diesem Szenario werden ausschließlich die Krankenkassen als Nutzende betrachtet.

Im Rahmen des Projekts möchte Timo eine existierende App weiterentwickeln, die die Anbindung an die Krankenkassen ermöglicht. Er möchte die folgenden Funktionen integrieren: 1.) Mitglieder der privaten Krankenkasse können ihre Abrechnungen sofort einscannen und bekommen ihr Geld erstattet, 2.) wenn Patient:innen zwischen Ärzt:innen wechseln oder zu Fachärzt:innen gehen, sollen sie die Möglichkeit haben, ihre Gesundheitsdaten „zentral“[1] speichern zu lassen, um sie bei anderen Ärzt:innen abrufen zu können.

Laut Studienordnung hat Timo für seine Masterarbeit nur sechs Monate Zeit. Ihm fällt schnell auf, dass er die Implementierung des Teilprojekts in dieser Zeit nicht allein schaffen kann. Während des Projekts hat er Kayla kennengelernt, sie muss ebenfalls eine Masterarbeit schreiben und findet Timos Teilprojekt interessant. Sie überlegen gemeinsam, wie sie sich das Teilprojekt aufteilen können. Timo möchte sich der ersten Funktion widmen und Kayla der zweiten. Beide Schnittstellen werden viele Ähnlichkeiten aufweisen, sodass sich die beiden gegenseitig unterstützen können. Trotzdem können beide eine eigenständige Masterarbeit schreiben und der Erfolg des einen ist nicht abhängig vom Erfolg des anderen. Im Zusammenhang mit der Entwicklung der Schnittstelle untersucht Timo in seiner Masterarbeit die folgenden Forschungsfragen:

[1] Eine echte zentrale Speicherung ist aufgrund dieser sensiblen Daten nicht ohne spezielle Sicherheitsmaßnahmen und verteilte Speicherorte zu empfehlen.

1. Wie können sensible Daten unter der Verwendung von Smartphone-Apps sicher übermittelt werden?

2. Welche Hürden existieren beim Anschluss von Open-Source-Software an das deutsche Gesundheitswesen?

Um diese Fragen beantworten zu können, muss Timo sich zunächst mit existierender Fachliteratur auseinandersetzen. Anschließend wird die eigene Implementation vorgestellt und die Güte der Software beurteilt. Im letzten Teil seiner Masterarbeit wird er sein Vorgehen und die Erkenntnisse aus der Fachliteratur reflektieren.

Bezug zum Forschungsdatenmanagement im zweiten Szenario

In diesem Szenario wird Software entwickelt und ein Softwareentwicklungsprozess durchlaufen, um die beiden Forschungsfragen beantworten zu können. Diese Software soll danach als Open-Source-Software frei zur Verfügung gestellt werden, wofür eine **Lizenz** für die Nutzung der Software vergeben werden muss. Da die Studierenden an einigen Teilen gemeinsam arbeiten, ihre Eigenleistung jedoch identifizierbar bleiben sollen, muss die **Autor:innenschaft** hinreichend kenntlich gemacht werden.

Ein weiterer wichtiger Aspekt in diesem Szenario ist die **Interdisziplinarität** des Projekts. Es wird die Informatik mit der Medizin zusammengebracht, weshalb ein gutes Verständnis darüber geschaffen werden muss, welche Daten in der App anfallen werden und welche besonders schützenswert sind. Eine Zusammenarbeit mit den Krankenkassen ist ebenfalls wichtig, damit die Schnittstellen anschlussfähig programmiert werden können. Da Timo dafür Einblicke in interne Prozesse der Krankenkasse bekommt, muss er eine **Verschwiegenheitserklärung** unterschreiben. Die Testung der App mit Proband:innen wird Timo nicht mehr innerhalb der Masterarbeit umsetzen können. Deswegen fallen **keine personenbezogenen Daten** im Rahmen der Masterarbeit an.

3.3 Szenario 3 (Drittmittelprojekt)

Name	Alex (Müller)
Alter	30
Geschlecht	divers
Studiengang	Informatik
Angestrebter Abschluss	Promotion
Universität	Humboldt-Universität zu Berlin
Herkunft	deutsch

Alex ist Doktorand:in in der Informatik und forscht in dem BMBF-Projekt „Learning Analytics to Improve Computer Science Teaching (LA2ICST)“. Das Projekt verfolgt das Ziel, die Abbruchquote im Informatikstudium zu reduzieren und die universitäre Lehre zu unterstützen. Dafür wird für ausgewählte Kurse ein eigener Kurs-Hashtag bei Twitter angelegt, und die Twitter-Daten der Teilnehmenden werden während des Kurses analysiert und ausgewertet. Das interdisziplinäre Verbundprojekt wird an mehreren deutschen Universitäten durchgeführt und beschäftigt sich ausschließlich mit der Zielgruppe von Informatikstudierenden. Da in der Informatik viele Studierende das Studium bereits im ersten oder zweiten Semester abbrechen, werden insbesondere Grundlagenveranstaltungen im Rahmen des Projekts untersucht. Aus der Literatur konnten die folgenden Faktoren identifiziert werden, die im Zusammenhang mit den hohen Abbruchquoten zu stehen scheinen: 1.) abbruchgefährdete Studierende werden nicht frühzeitig identifiziert und 2.) die Studierenden arbeiten nicht ausreichend kollaborativ. Diese Faktoren bilden das Zentrum des Projekts und sollen anhand von Indikatoren in den Twitter-Daten ermittelt werden. Indikatoren können beispielsweise eine deprimierte Stimmung in den Tweets, eine geringe Anzahl an Nachrichten auf Twitter oder die Abwesenheit von Studierenden sein. Einerseits soll frühzeitig erkannt werden, welche Studierenden den Inhalten nicht mehr folgen können und sich unzureichend beteiligen. Inwieweit diese Indikatoren auf andere Fächer übertragbar sind, soll in Anschlussprojekten untersucht werden. Andererseits wird ermittelt, bei welchen Lernaktivitäten die Studierenden Hilfestellungen für eine erfolgreiche Kollaboration benötigen. Die Forschungsfragen des Projekts sind:

1. Welche Indikatoren können aus Twitter-Daten expliziert werden, um Studierenden mit einem hohen Abbruchrisiko im Informatikstudium zu identifizieren?

2. Welche dieser Indikatoren scheinen domänenspezifisch für die Informatik zu sein? (Diese Indikatoren sollen später als Ausgangspunkt für die Vergleichsstudie im Anschlussprojekt sein.)

3. Welche Lernaktivitäten setzen besonders gute kollaborative Fähigkeiten von den Studierenden voraus?

4. Wie können die Studierenden beim Erlangen kollaborativer Fähigkeiten von den Dozierenden unterstützt werden?

Alex beschäftigt sich ausschließlich mit der ersten und zweiten Forschungsfrage des Projekts. Dafür werden quantitative Daten in Form von Twitter-Daten herangezogen, die mit einem selbst entwickelten Algorithmus analysiert werden.

Bezug zum Forschungsdatenmanagement im dritten Szenario

Um Alex' Forschungsfragen zu beantworten, ist das Messen von **Twitter-Datenströmen** notwendig. Dafür werden Kurs-Hashtags angelegt, unter denen die Studierenden Kommentare hinterlassen können. Da bei der Auswertung auch auf die deprimierte Stimmung geachtet werden soll, muss eine **Ethikprüfung** stattfinden. Obwohl es sich bei den Tweets um öffentlich gestellte Nachrichten handelt, muss geklärt werden, wer der Eigentümer dieser Daten ist und wie die **Nutzungsrechte** geregelt sind. Beim Speichern der Daten werden diese automatisch **pseudonymisiert**. Trotzdem unterliegen pseudonymisierte Daten der DSGVO, da ein Rückbezug auf die Person noch herstellbar ist. Alex muss daher eine **informierte Einwilligungserklärung** der Kursteilnehmenden einholen.

Kapitel 4

Forschungsdatenmanagement in der Informatik

Dieses Kapitel des Buches basiert strukturell und inhaltlich auf dem Train-the-Trainer-Konzept zum Thema Forschungsdatenmanagement von Biernacka et al., 2021a, wobei der Fokus auf die in der Informatik üblichen Forschungsdaten über das existierende Konzept hinausgeht. Da es eine große inhaltliche Überschneidung der Basisinformationen zum Thema Forschungsdatenmanagement gibt, wird zugunsten der Lesbarkeit auf wiederholte Zitierung der einzelnen Passagen verzichtet, da das ganze Kapitel als eine erweiterte Zitation angesehen werden kann.

4.1 Forschungsdaten in der Informatik

Digitale Forschungsdaten

Unter digitalen Forschungsdaten verstehen wir [...] alle digital vorliegenden Daten, die während des Forschungsprozesses entstehen oder ihr Ergebnis sind. (Kindling & Schirmbacher, 2013)

Informatiker:innen haben schon immer mit Daten gearbeitet. Das Konzept von Daten als Informationen, die in eine binäre digitale Form umgewandelt werden, ist daher in dieser Disziplin nicht unbekannt. Da heutzutage die sogenannte Bindestrich-Informatik kaum noch wegzudenken ist (z. B. Wirtschafts-Informatik, Bio-Informatik, Geo-Informatik u. v. m.), ist die Vielfalt der Forschungsdaten in der Disziplin umso größer geworden. Sie umfassen somit den geschriebenen Quellcode, Software-Artefakte, Datenbanken, Knowledge Graphs oder andere unstrukturierte Daten. Die Datendomäne schließt sowohl Text-, Audio-, Video- als auch Bilddaten mit ein. Es kann sich dabei um Messdaten, Laborwerte, audiovisuelle Informationen, Texte, Umfragedaten oder Beobachtungsdaten, methodische Testverfahren, Algorithmen, formale Beschreibungen von Beweisen, Simulationen sowie Fragebögen handeln (vgl. Abbildung 4.1). Die Heterogenität der Forschungsdaten in der Informatik macht eine detailliertere Kategorisierung unmöglich und führt zu einer Fülle an unterschiedlichen Standards und Methoden, die zur Anwendung kommen.

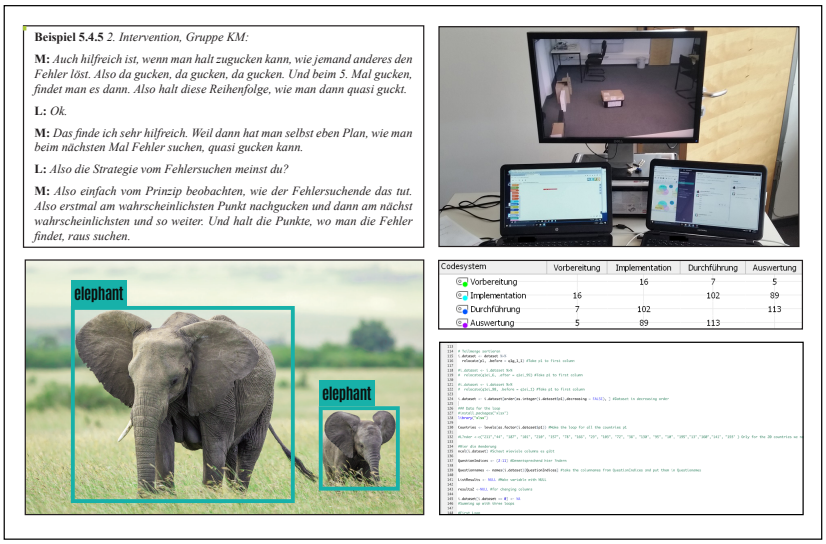


Abbildung 4.1: Beispiele von Forschungsdaten in der Informatik; Quelle der Abbildungen links und rechts oben sowie rechte Mitte: Schulz (2019)

4.1.1 Software als Forschungsdatum

Software hat in der Forschung eine wichtige Funktion, aber dennoch wird ihre Rolle als Forschungsdatum stark diskutiert. Nach Clément-Fontaine et al. (2019) kann man die Funktion von Software in der Forschung dreierlei aufteilen:

1. Software fundiert als **Werkzeug** zur Verarbeitung, Erstellung und zum Testen von Hypothesen anhand verschiedener Forschungsdaten.

2. Software kann ein eigenständiges **Forschungsergebnis** sein, das als eine effektive algorithmische Lösung eines Problems dient.

3. Software kann selbst ein **Forschungsgegenstand** sein.

Software lässt sich daher nicht auf eine Reihe von Codezeilen (oder „Commits“) reduzieren. Im ersten Fall sprechen wir von Forschungssoftware, die als Funktion die Unterstützung der Forschung und Wissenschaft hat. Gute Forschungssoftware kann den Unterschied zwischen gültigen, nachhaltigen, reproduzierbaren Forschungsergebnissen und kurzlebigen, unzuverlässigen oder fehlerhaften Ergebnissen ausmachen.

In der zweiten und dritten Funktion kann Software als Forschungsdatum betrachtet werden.

4.1.2 Anwendung auf die Szenarien

Forschungsdaten bei Szenario 1

In der Arbeit von Carla fallen aufgezeichnete Audiodaten und Transkripte der Audiodaten als Forschungsdaten an.
Darüber hinaus werden im Rahmen der Forschungsergebnisse weitere Daten extrahiert und erzeugt. Dazu gehören beispielsweise: Codemanuals, Codesysteme und Anzahl der

Codings, statistische Auswertungen, demografische Daten, Daten zur Infrastruktur und verwendeter Tools, sowie die erstellten Fragebögen.

Forschungsdaten bei Szenario 2

In dem Fall von Timo stellt die Software einerseits einen Forschungsgegenstand dar, andererseits aber auch ein Forschungsergebnis. Darüber hinaus werden Testdatensätze für Funktionstests (Daten über Patient:innen, von Ärzt:innen etc.) erstellt und verwendet.

Forschungsdaten bei Szenario 3

In diesem Projekt wird Software als Werkzeug genutzt, um Twitterdaten zu extrahieren, um sie anschließend analysieren und auswerten zu können. Da dieser Algorithmus von Alex entwickelt wird, handelt es sich dabei auch um einen Forschungsgegenstand. Darüber hinaus fallen verschiedene Forschungsdaten an, wie Twitternachrichten, Anzahl an Tweets und statistische Auswertung.

4.2 Definition von Forschungsdatenmanagement

4.2.1 Forschungsdatenlebenszyklus

Forschungsdaten haben häufig eine längere Lebensdauer als die Forschungsprojekte selbst und durchleben somit einen eigenen Lebenszyklus.

Forschungsdatenlebenszyklus

Der Forschungsdatenlebenszyklus stellt die Phasen vor, die die Forschungsdaten durchlaufen (können).

Der Forschungsdatenlebenszyklus ist dabei ein Hilfsmittel, das modellhaft die unterschiedlichen Stadien, die die Forschungsdaten durchlaufen, darstellt. Die Granularität dieser Darstellung kann je nach Disziplin und Forschungsvorhaben stark variieren, wobei sechs Stationen in jedem Forschungsdatenlebenszyklus vorkommen: Planung, Erhebung, Verarbeitung und Analyse, Zugang, Archivierung und Nachnutzung (vgl. Abbildung 4.2). Weitere Schleifen zwischen den einzelnen Stationen sind möglich und in größeren Projekten sogar geläufig. Ein Beispiel für einen komplexeren Forschungsdatenlebenszyklus ist Abbildung 4.3 zu entnehmen.

Die zirkuläre Darstellung des Forschungsdatenlebenszyklus unterstreicht den Gedanken der offenen Wissenschaft (Open Science/Open Research) mit dem stark vertretenen Aspekt der Publikation und Wiederverwendbarkeit von Daten (Open Data) sowie der Transparenz der Forschung.

Abbildung 4.2: Forschungsdatenlebenszyklus nach UK Data Service (2020)

Abbildung 4.3: Forschungsdatenlebenszyklus nach UK Data Service (2020) erweitert um die beispielhaften Schleifen

4.2.2 Softwarelebenszyklus

Der Softwarelebenszyklus (siehe Abbildung 4.4) stellt, ähnlich dem Forschungsdatenlebenszyklus, die unterschiedlichen Phasen dar, die die Software durchlaufen muss, um am Ende

ausgeliefert zu werden. Der Zyklus beginnt mit der Erhebung der Anforderungen und dem Erkennen des (Forschungs-)Problems. Diese wird nachfolgend analysiert und umgesetzt. Nach dem Implementieren und Testen wird die Software produktiv genutzt. Als nächstes sollte die Software nachhaltig abgelegt werden, um eine Nachnutzung zu ermöglichen. Leider wird dieser Schritt bei Forschungssoftware immer noch häufig vernachlässigt. Im Sinne von Open Source sollte Software (sowohl als Forschungssoftware als auch als Forschungsdatum) offen publiziert und archiviert werden.

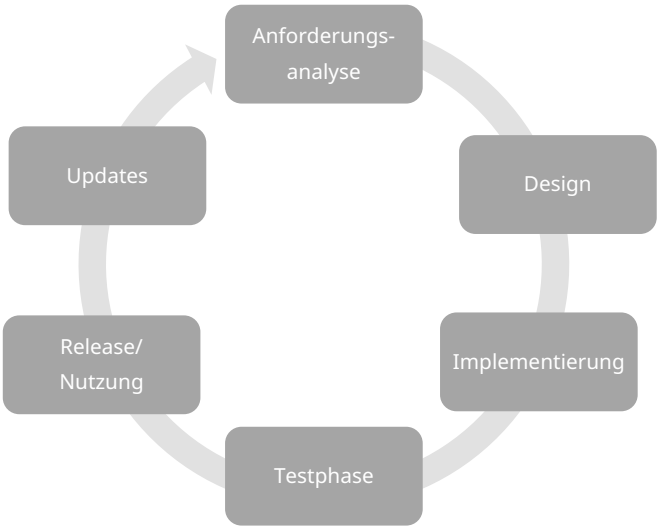


Abbildung 4.4: Beispielhafter schematischer Softwarelebenszyklus

Da bei der Entwicklung von (Forschungs-)Software auch Daten eingelesen und genutzt werden, die wiederum selbst Forschungsdaten sind bzw. sein können, kann es Überschneidungen von mehreren Forschungsdaten- und Softwarelebenszyklen geben, wie in Abbildung 4.5 dargestellt.

4.2.3 Forschungsdatenmanagement

Anhand des Forschungsdatenlebenszyklus lässt sich erläutern, was FDM ist:

Forschungsdatenmanagement

FDM umfasst alle Aktivitäten, die mit der Planung, Erhebung, Verarbeitung, Analyse, Zugang, Archivierung, Nachnutzung bis zur Löschung der Daten in Zusammenhang stehen. Mit anderen Worten ist das Forschungsdatenmanagement der professionelle Umgang mit den Forschungsdaten entlang des Forschungsdatenlebenszyklus.

Forschungsdaten sind eine wertvolle Ressource in der Forschung und sollten mit Bedacht und Verantwortung verwaltet werden. FDM begleitet den Forschungsprozess von den ersten Planungen bis zur Archivierung, Nachnutzung und Löschung der Daten.

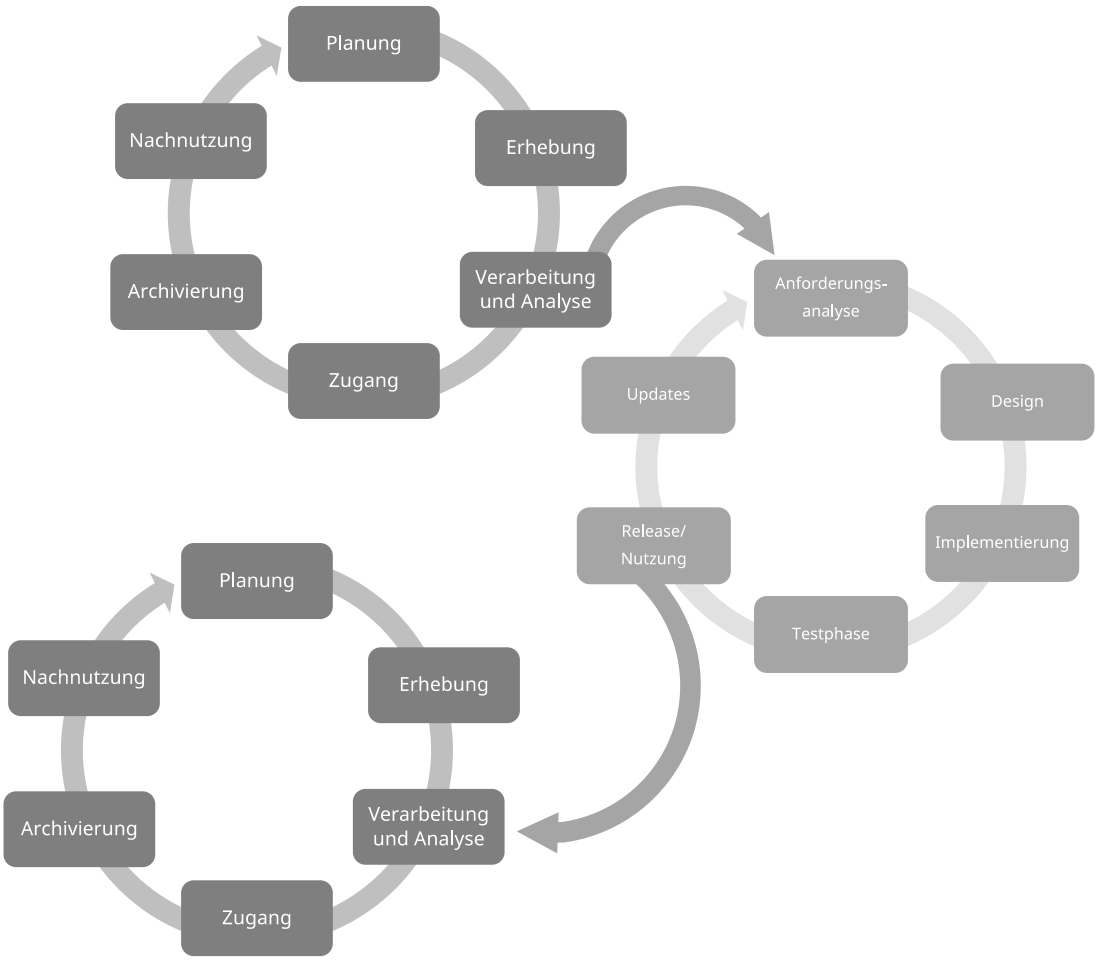


Abbildung 4.5: Beispiel von Überschneidungen von mehreren Forschungsdaten- und Softwarelebenszyklen

Die Aspekte des Forschungsdatenmanagements lassen sich einfach entlang der Phasen des Forschungsdatenlebenszyklus aufzeigen (vgl. Abbildung 4.6):

- Planung

  In der Planungsphase des Forschungsvorhabens sollte ein Untersuchungsdesign unter dem Aspekt der Datenerhebung und -verarbeitung erstellt werden. Dafür eignet sich die Erstellung eines Datenmanagementplans (vgl. Kapitel 4.6). Darüber hinaus sollten schon vor Beginn des Forschungsvorhabens bereits existierende Forschungsdaten lokalisiert werden (vgl. Kapitel 4.16). Bei der Stellung von Drittmittelanträgen sollte FDM mitgedacht und finanzielle Mittel beantragt werden.

- Erhebung

  Die Erstellung bzw. Erhebung von Forschungsdaten kann auf unterschiedliche Weise geschehen. Es können sowohl Beobachtungen, Messungen, Experimente, Simulation, Interviews und/oder Fragebögen durchgeführt als auch bereits existierende Daten für die eigene Forschung einbezogen werden. Bereits bei der Erhebung sollten die Daten mit

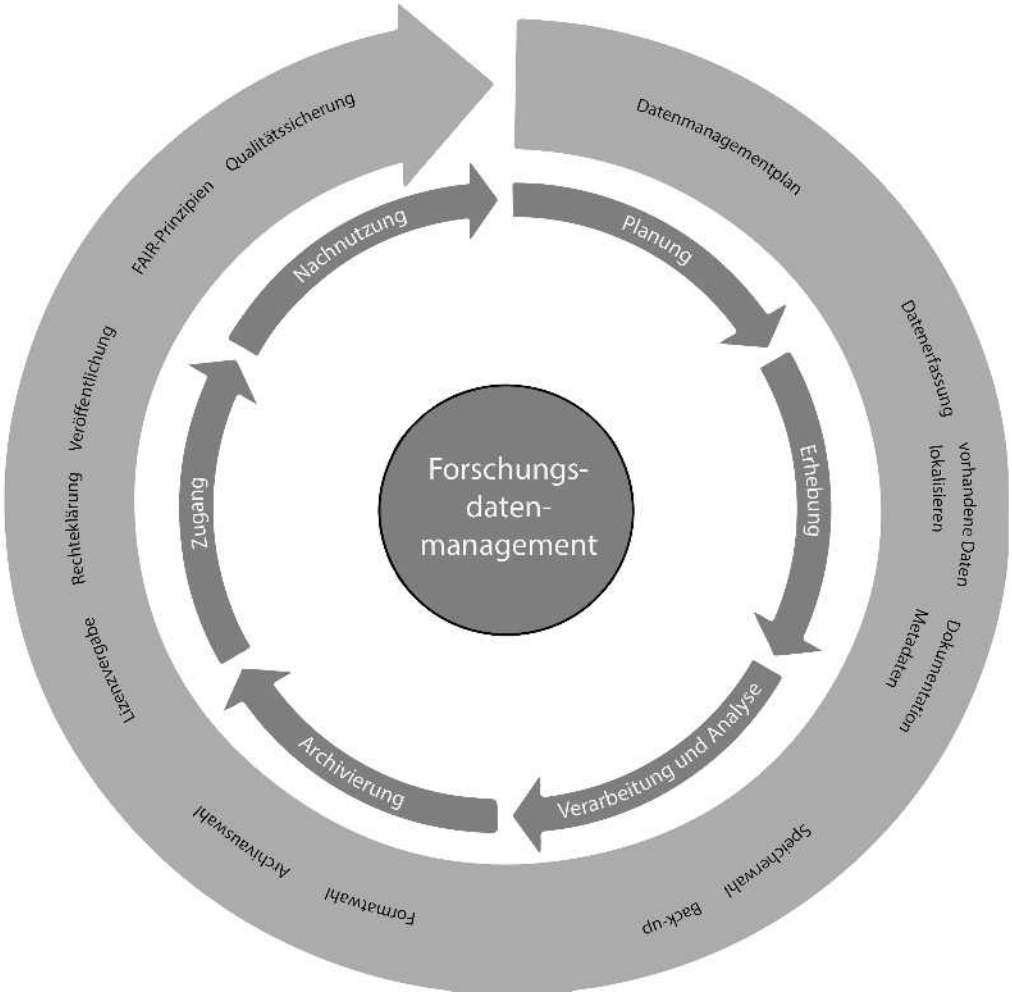


Abbildung 4.6: Aspekte des Forschungsdatenmanagements anhand des Forschungsdatenlebenszyklus

Metadaten beschrieben werden (vgl. Kapitel 4.11). Bei Erhebung von personenbezogenen oder sensiblen Daten ist es wichtig, bereits im Vorfeld informierte Einwilligungen von den Studienteilnehmenden einzuholen (vgl. Kapitel 4.8).

- Verarbeitung und Analyse

  In dieser Phase des FDM wird vor allem mit den Forschungsdaten selbst gearbeitet. Diese werden bereinigt, verarbeitet und analysiert. Währenddessen sollten die Daten gut dokumentiert (vgl. Kapitel 4.11), an geeigneten Speicherorten abgelegt, versioniert und gesichert werden (vgl. Kapitel 4.9 und 4.10).

- Zugang

  Der Zugang zu den Forschungsdaten kann auf verschiedenen Ebenen geschehen: unter den Projektmitarbeitenden, auf Anfrage für Externe oder – im Sinne von Open Science – ohne Einschränkungen für das breite Publikum (vgl. Kapitel 4.13). Um diesen Prozess zu vereinfachen, sollten an dieser Stelle die rechtlichen Aspekte geklärt (vgl. Kapitel 4.8 und 4.14) und Lizenzen vergeben werden (vgl. Kapitel 4.14.5).

- Archivierung

  Um einen langfristigen Zugriff auf die Forschungsdaten zu gewährleisten, ist es notwendig, diese in geeigneten Formaten zu speichern, zu dokumentieren und in Langzeitarchiven abzulegen (vgl. Kapitel 4.15).

- Nachnutzung

  Der sechste Aspekt des FDM kann aus zwei Perspektiven betrachtet werden: einerseits mit der Frage, wie man Forschungsdaten am besten zur Nachnutzung bereitstellt, und andererseits, wie man existierende Forschungsdaten nachnutzen und korrekt zitieren kann (vgl. Kapitel 4.16). Die letzte Betrachtung schließt den Zyklus, indem die Nachnutzung zugleich die Erhebung von Forschungsdaten darstellt.

4.2.4 Management von (Forschungs-)Software

Unabhängig davon, welche Perspektive man auf Software in der Forschung annimmt – ob als Forschungssoftware oder Forschungsdatum –, sollte Software verantwortungsvoll und nachhaltig verwaltet werden. Dies beinhaltet in beiden Fällen die gleichen Schritte entlang des Lebenszyklus. Neben der Planung, Erstellung und Verarbeitung beinhaltet es auch den Zugang, die Archivierung und Nachnutzung der Software. Software und Daten ähneln sich zudem bezüglich Credit und Metriken im Wissenschaftskontext.

4.2.5 Vorteile eines systematischen Forschungsdatenmanagements

Forschungsdatenmanagement ist Teil des Forschungsvorhabens, keine zusätzliche Aufgabe. Viele der Aspekte werden von Forschenden bereits unbewusst durchgeführt (z. B. Dateibenennung, Wahl des Speicherortes und -formats). Ein bewusstes Forschungsdatenmanagement bringt jedoch viele Vorteile mit sich, denn es vereinfacht nicht nur die Arbeit mit den Daten während des Foschungsprojektes, sondern auch danach. Obwohl es zu Beginn Zeit kostet, zahlt sich der Aufwand in den späteren Phasen aus. Das Wiederfinden und Nachvollziehen der Forschungsdaten nach vielen Jahren ist mit einem systematischen und dokumentierten FDM deutlich einfacher – sowohl für andere als auch für einen selbst. Die Daten sind eindeutig benannt und in offenen Formaten mit klar definierten Versionen abgelegt. Darüber hinaus werden die Forschungsdaten regelmäßig auf verschiedenen Speichermedien gesichert, wodurch Datenverlust vermieden werden kann.

Durch strukturiertes Forschungsdatenmanagement wird die ganze Forschung nachvollziehbarer und reproduzierbarer. Die Validierung der Ergebnisse im Sinne von Open Science und Guter Wissenschaftlicher Praxis wird somit gefördert. Dies kann zur besseren Sichtbarkeit und somit auch zu gesteigerter Reputation beitragen. Auch Verlage und Forschungsförderer betonen die Relevanz des FDM. Aus der Sicht der Förderer werden mit einem guten FDM die Geldmittel optimal eingesetzt, denn durch Nachnutzung von Daten können Kosten der erneuten Erhebung eingespart werden.

Die Publikation der Forschungsdaten unter offenen Lizenzen ermöglicht die problemlose Weitergabe der Daten. Diese können dann sowohl zitiert als auch referenziert werden, wodurch die eigene Arbeit an Relevanz und besserer Sichtbarkeit gewinnt.

4.2.6 Literaturempfehlungen

- Corti, L., Van den Eynden, V., Bishop, D. V. M. & Woollard, M. (2014). *Managing and Sharing Research Data: A Guide to Good Practice*. Los Angeles, CA: Sage.
- Kindling, M. & Schirmbacher, P. (2013). Die digitale Forschungswelt als Gegenstand der Forschung / Research on Digital Research / Recherche dans la domaine de la recherche numerique. *Information - Wissenschaft & Praxis, 64* (2-3), 127–136. https://doi.org/10.1515/iwp-2013-0017

4.2.7 Anwendung auf die Szenarien

Forschungsdatenmanagement bei Szenario 1

Auch im Rahmen ihrer Bachelorarbeit muss sich Carla bereits mit FDM auseinandersetzen. Sie erstellt ein Untersuchungsdesign und einen kurzen Datenmanagementplan, um ihr Vorhaben möglichst konkret zu planen. Vor der Datenerhebung ist die Einholung einer informierten Einwilligungserklärung für Carla notwendig, da sie auch personenbezogene Daten in den Interviews erhebt. In der Phase der Verarbeitung und Analyse arbeitet Carla mit den Daten direkt und dokumentiert ihr Vorgehen. Die Daten können auf ihrem eigenen PC liegen und sollten gesichert werden. Carla stellt sicher, dass nur sie Zugriff auf die erhobenen Daten hat und die Daten für eine Archivierung anonymisiert sowie ggf. in ein geeignetes Format für eine Langzeitarchivierung konvertiert werden. Carlas Daten haben großes Potenzial für eine Nachnutzung, weshalb sie sich intensiv mit der Datenbereitstellung beschäftigt.

Forschungsdatenmanagement bei Szenario 2

Obwohl Timos Projekt nicht drittmittelfinanziert ist, empfiehlt es sich dennoch, einen Datenmanagementplan zu schreiben, um sich auf diese Weise mit dem Projektpartner auf bestimmte Arbeitsweisen und Standards zu einigen. Da es sich bei dem Projekt um ein Open-Source-Projekt handelt, sollte Timo im Vorfeld nach potenziell nützlicher bereits existierender Software suchen. Timo würde voraussichtlich seinen privaten Computer als Speicherort wählen, wobei er also eine regelmäßige Datensicherung sicherstellen sollte. Darüber hinaus wäre eine weitere Speicheroption relevant, um die Daten mit Kayla zu teilen und zu prüfen, ob die Software und die Daten über die jeweiligen Forschungsfragen hinaus miteinander kompatibel sind. Die optimale Lösung würde die Nutzung des Clouddienstes der TU Dresden (Cloudstore[a]) darstellen, um die Daten zu speichern und ggf. mit Kayla zu teilen.

[a] https://tu-dresden.de/zih/dienste/service-katalog/zusammenarbeiten-und-forschen/datenaustausch/cloudstore. Archivierte Version: https://perma.cc/JSP2-6E5E.

Forschungsdatenmanagement bei Szenario 3

In Alex' Szenario wird ein Datenmanagementplan vom Drittmittelgeber erwartet. Dieser sollte bereits bei der Antragstellung vorbereitet werden. Bei Projektbeginn sollte Alex nach nachnutzbaren Daten suchen, die für das Projekt relevant sein könnten. Da es sich bei dem Projekt um ein Verbundprojekt handelt, ist es wichtig, einen Speicherort zu wählen, auf den alle Projektpartner:innen zugreifen können (z. B. die Cloud-Lösung der Humboldt-Universität zu Berlin (HU-Box[a]) mit gesponserten Accounts für die externen Projektpartner).

[a] https://box.hu-berlin.de/. Archivierte Version: https://perma.cc/GL7L-Y5SR.

4.3 Forschungsdaten-Policys

Eine Forschungsdaten-Policy ist ein Dokument, das die Vorgaben zum Umgang mit den Forschungsdaten beschreibt. Sie sind als Richtlinien für alle Mitarbeitenden und Studierenden einer Institution (z. B. Hochschule), wie Forschungsdatenmanagement durchgeführt werden soll, zu verstehen. Viele Einrichtungen greifen auf die Prinzipien des Open Access oder die Leitlinien der Guten Wissenschaftlichen Praxis (Deutsche Forschungsgemeinschaft, 2019) zurück (z. B. Universität Hamburg). Andere Institutionen haben hingegen spezielle Dokumente zum Umgang mit Forschungsdaten (z. B. Humboldt-Universität zu Berlin[1]). Drittmittelgeber wie die Deutsche Forschungsgemeinschaft (DFG)[2] oder Europäische Kommission[3] haben eigene Forschungsdaten-Policys formuliert.

Forschungsdaten-Policys können abhängig vom Ziel und der Zielgruppe in institutionelle, interdisziplinäre, disziplinspezifische, Zeitschriften- und Verlags-Policys sowie Förderrichtlinien aufgeteilt werden.

4.3.1 Institutionelle Forschungsdaten-Policys

Am 01. August 2019 traten die neuen „Leitlinien zur Sicherung guter wissenschaftlicher Praxis" der DFG in Kraft (Deutsche Forschungsgemeinschaft, 2019) und sind für alle Universitäten bindend. In diesem Dokument wird das Thema Forschungsdatenmanagement eindringlich betrachtet und fordert somit die Institutionen dazu auf, Richtlinien zum Umgang mit Forschungsdaten zu formulieren. Spätestens ab dem Zeitpunkt fingen die Universitäten an, den Umgang mit Forschungsdaten in ihren eigenen Universitätsrichtlinien[4] zu verankern.

[1] https://www.cms.hu-berlin.de/ueberblick/projekte/dataman/hu-fdt-policy/view. Archivierte Version: https://perma.cc/LC56-VTPU.

[2] https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/leitlinien_forschungsdaten.pdf. Archivierte Version: https://perma.cc/FHX8-J37U.

[3] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. Archivierte Version: https://perma.cc/UCD8-6CAJ.

[4] Liste der Forschungsdaten-Policys an deutschen Hochschulen: forschungsdaten.org: „Data Policys". https://www.forschungsdaten.org/index.php/Data_Policys#Institutionelle_Policys. Archivierte Version: https://perma.cc/5LK2-Y6ZT.

Neben der Regelung des offenen Zugangs von Forschungsdaten werden darin die personellen, organisatorischen und technischen Kapazitäten der Forschungseinrichtung für das FDM betrachtet (Hiemenz & Kuberek, 2018).

Institutionelle Forschungsdaten-Policys sollen einerseits den Studierenden und Forschenden der Einrichtung Anhaltspunkte geben, an welcher Stelle die Forschungsdaten gespeichert werden sollen, wie die Struktur und Sicherung der Daten aussehen soll oder auch wo und unter welchen Lizenzen die Forschungsdaten publiziert werden sollen. Andererseits soll es auch die Pflichten sowie die möglichen Dienste der Institution für das Forschungsdatenmanagement beschreiben und das Weiterbildungsangebot zu diesem Thema vorstellen. Eine Kontaktstelle bei Fragen rund um Forschungsdaten und Open Science sollte auch genannt sein.

Darüber hinaus bietet eine Forschungsdaten-Policy eine Reihe von Vorteilen, wie eine höhere Reputation der Universität, die Transparenz für Hochschulangehörigen, die sich klare Regelungen zum Forschungsdatenmanagement wünschen, oder Vorteile bei der Mitteleinwerbung durch die zunehmende Bedeutung der Sicherung, Aufbewahrung und nachhaltigen Verfügbarkeit von Forschungsdaten bei (inter-)nationalen Förderorganisationen (Hiemenz & Kuberek, 2018).

4.3.2 Disziplinspezifische Forschungsdaten-Policys

Forschungsdaten können in den verschiedenen Disziplinen sehr unterschiedlich sein und so auch deren Verwaltung. Aus diesem Grund gibt es in verschiedenen Fachgebieten bereits spezifische Richtlinien zum Umgang mit Forschungsdaten (z. B. in der Psychologie[5]). In den Sozialwissenschaften gibt es das Übereinkommen der Zusammenarbeit der europäischen Datenarchive, das vom Consortium of European Social Science Data Archives (CESSDA) erschaffen wurde. In den Lebenswissenschaften gibt es einerseits die Gute Klinische Praxis (GCP) und die Grundsätze der Guten Laborpraxis (GLP), die beide den Umgang mit Forschungsdaten beschreiben. Weitere fachspezifische Stellungnahmen, die in Fachkollegien der DFG verwendet werden, sind auf der Webseite der DFG[6] zu finden.

4.3.3 Policys von Forschungsförderern

Auch Forschungsförderer verfassen eigene Policys zum Umgang mit Forschungsdaten. Antragstellende werden darin mitunter dazu aufgefordert, Datenmanagementpläne zu schreiben und ihre Daten der breiten Öffentlichkeit zur Verfügung zu stellen.

4.3.4 Zeitschriften- und Verlags-Policys

Immer mehr Verlage und Zeitschriften setzen auf die Publikation von Forschungsdaten, die den wissenschaftlichen Artikeln zugrunde liegen (vgl. Tabelle 4.1). Auf diesem Weg soll auch ein Peer-Review der Daten selbst ermöglicht werden. Darüber hinaus ermöglicht die Publikation der Daten auch deren Nachnutzung und gestaltet den ganzen Forschungsprozess transparenter und reproduzierbarer.

[5] https://doi.org/10.31234/osf.io/hcxtm

[6] https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/empfehlungen/index.html. Archivierte Version: https://perma.cc/NL9C-QPPX.

Tabelle 4.1: Beispiele von Verlags-Policys

Zeitschriften-verlag	**Forschungsdaten-Policy**
Springer Nature	**Typ 1** – Es wird dazu ermutigt, Daten zu teilen und zu zitieren. **Typ 2** – Es wird dazu ermutigt, Daten zu teilen und Datenzugänglichmachung nachzuweisen. **Typ 3** – Es wird dazu ermutigt, Daten zu teilen; Aussagen zur Datenverfügbarkeit sind verpflichtend. **Typ 4** – Es ist eine Voraussetzung, Daten zu teilen, dies nachzuweisen und ein Peer-Review der Daten zu ermöglichen.
Taylor & Francis	**Grundlegende Richtlinien zum Teilen von Daten** – Es wird dazu ermutigt, dort wo möglich, Daten zu teilen und zu publizieren, sie zu zitieren und Aussagen zur Datenverfügbarkeit zu machen. **Teilen auf Anfrage** – Es ist verpflichtend, Daten auf begründete Anfrage hin zur Verfügung zu stellen. **Öffentlich zugänglich** – Es ist eine Voraussetzung, Daten unter einer frei gewählten Lizenz zu teilen. **Open Data** – Es ist verpflichtend, Daten unter einer Lizenz, die die Nachnutzung durch Dritte für jeden rechtmäßigen Zweck erlaubt, zu teilen. Die Daten müssen auffindbar und vollständig zugänglich sein. **Open und FAIR Data** – Es ist verpflichtend, Daten unter einer Lizenz, die die Nachnutzung durch Dritte für jeden rechtmäßigen Zweck erlaubt, zu teilen. Die Daten müssen den FAIR-Prinzipien entsprechen (vgl. Kapitel 4.5).
Elsevier	Es wird dazu ermutigt, Daten zu teilen, dort, wo angemessen und zum frühestmöglichen Zeitpunkt.
PLOS	Es ist verpflichtend, Daten, die zur Replikation der Ergebnisse einer Studie notwendig sind, zum Zeitpunkt der Veröffentlichung uneingeschränkt öffentlich zugänglich zu machen. Wenn bestimmte rechtliche oder ethische Einschränkungen die öffentliche Freigabe eines Datensatzes verbieten, muss angegeben werden, wie Dritte Zugang zu den Daten erhalten können.

4.3.5 Literaturempfehlungen

- Hiemenz, B. M. & Kuberek, M. (2018). Empfehlungen zur Erstellung institutioneller Forschungsdaten-Policies. DepositOnce. https://doi.org/10.14279/depositonce-7521
- Grasse, M., López, A. & Winter, N. (2018). Musterleitlinie für Forschungsdatenmanagement (FDM) an Hochschulen und Forschungseinrichtungen. https://doi.org/10.5281/zenodo.1149133

4.3.6 Anwendung auf die Szenarien

Forschungsdaten-Policys bei Szenario 1

In diesem Szenario muss sich an die Open-Access-Policy der Universität Hamburg[a] gehalten werden. Dabei werden die Angehörigen der Universität Hamburg ausdrücklich ermutigt und darin unterstützt, ihre Forschungsergebnisse, Forschungsdaten, Repositorien, virtuellen wissenschaftlichen Sammlungen sowie E-Learning- und E-Lecturing-Angebote über die institutionelle Infrastruktur zur Verfügung zu stellen

[a] https://www.openaccess.uni-hamburg.de/openaccess/oa-policy.html. Archivierte Version: https://perma.cc/4CGJ-QNS7.

Forschungsdaten-Policys bei Szenario 2

In diesem Szenario muss sich an die Leitlinien für den Umgang mit Forschungsdaten der Technischen Universität Dresden[a] gehalten werden. Hierbei wird ermutigt, einen Datenmanagementplan bereits bei Projektkonzeption bzw. -antragstellung zu erstellen, der den Umgang mit Forschungsdaten dokumentiert. Darüber hinaus wird der freie Zugang zu Forschungsdaten (Open Access) gefördert und unterstützt. Die Entscheidung für eine Veröffentlichung und deren rechtliche Bedingungen liegt jedoch in der Eigenverantwortung der Wissenschaftler:innen. Forschungsdaten, die Grundlage einer Publikation bilden, sollen langfristig in einem geeigneten vertrauenswürdigen Datenarchiv bzw. Repositorium archiviert und/oder geeignet veröffentlicht werden.

[a] http://tu-dresden.de/tu-dresden/qualitaetsmanagement/ressourcen/dateien/wisprax/Leitlinien-fuer-den-Umgang-mit-Forschungsdaten-an-der-TU-Dresden.pdf?lang=de. Archivierte Version: https://perma.cc/95NW-NN8Z.

Forschungsdaten-Policys bei Szenario 3

In diesem Szenario muss sich an die Forschungsdaten-Policy der Humboldt-Universität zu Berlin[a] gehalten werden. Demnach sind HU-Angehörige dazu verpflichtet, die Forschungsdaten sicher zu speichern, angemessen aufzubereiten und zu dokumentieren sowie langfristig aufzubewahren. Es gibt nach diesen Richtlinien keine konkreten Angaben, zu welchem Zeitpunkt und zu welchen rechtlichen Bedingungen Forschungsdaten zugänglich gemacht werden. Darüber hinaus muss sich über Policys des Bundesministeriums für Bildung und Forschung informiert werden und diese müssen eben-

falls berücksichtigt werden. Diese erwartet in der Regel einen sogenannten Verwertungsplan oder Angaben zur Verwertung der Ergebnisse als Teil des Förderantrags. In Abhängigkeit vom Programm wird auch eine Datenablage in einem Repositorium erwartet.

[a] https://www.cms.hu-berlin.de/de/dl/dataman/infos/policy. Archivierte Version: https://perma.cc/388W-9SMK.

4.4 Institutionelle Infrastruktur

Die Aufgabe der Hochschule und Bildungseinrichtungen im Allgemeinen ist es, ihre Forschenden in ihren Forschungsvorhaben zu unterstützen. Dazu gehört auch die Bereitstellung der notwendigen Infrastruktur und Dienstleistungen sowie die Entwicklung der Kompetenzstandards für Forschende und wissenschaftsunterstützendes Personal.

Forschungsdatenmanagement ist im institutionellen Kontext eng mit organisationalen Prozessen verzahnt. Daher sollten verschiedene Akteur:innen eingebunden werden, um die Forschenden bei der Betreuung und beim Betrieb von Repositorien zu unterstützen, sie bei der Erstellung von einem DMP (vgl. Kapitel 4.6) mit Expertise und Werkzeugen zu beraten sowie bei der Erhebung, Nachnutzung, Verwaltung, Publikation, Archivierung bis zur Löschung der Forschungsdaten zu begleiten. Im Idealfall gehören fünf Stakeholder zu der FDM-Service-Landschaft: die FDM-Kontaktstelle, das IT/Rechenzentrum, die Bibliothek, die Forschungsförderung und die rechtlichen Anlaufstellen.

Je nach Möglichkeiten der Einrichtungen könnten folgende Services zur Verfügung stehen:

- Beratung zum Umgang mit Forschungsdaten im Kontext von Projektanträgen bzw. Drittmittelbeantragungen
- Ggf. fachspezifische Schwerpunkte z. B. Beratung zu Digital Humanities-Anwendungen
- Vermittlung von internen und externen Diensten
- Unterstützung bei der Erstellung eines DMP
- Schulungen und/oder Workshops zum Thema FDM
- Informationsveranstaltungen
- Informationsmaterial
- Cloud-Services
- Rechtliche Beratung zum Thema Forschungsdaten (auch in Kooperation mit Expert:innen)
- Sync-and-Share-Lösungen
- Back-up-Service
- Datenbankservice
- Forschungsdatenrepositorium (auch integriert im Publikationsserver)
- Versionierungssoftware
- Langzeitarchivierung
- Vergabe von persistenten Identifikatoren
- DMP-Tool
- Umfrage-Tools
- Elektronische Laborbücher (ELB)

4.4.1 Anwendung auf die Szenarien

Institutionelle Infrastruktur bei Szenario 1

Carla findet an der Universität gute Bedingungen im Rahmen der institutionellen Infrastruktur vor. Beispielsweise kann sie durch das Zentrum für nachhaltiges Forschungsdatenmanagement[a] unterstützt werden und Tools vom Regionalen Rechenzentrum[b] nutzen.

[a] https://www.fdm.uni-hamburg.de/. Archivierte Version: https://perma.cc/FC9V-YLDH.
[b] https://www.rrz.uni-hamburg.de/. Archivierte Version: https://perma.cc/8Y7V-U586.

Institutionelle Infrastruktur bei Szenario 2

An der TU Dresden kann sich Timo insbesondere an das Zentrum für Informationsdienste und Hochleistungsrechner[a] wenden. Dort werden ihm auch Informationen und Services im Bereich des Forschungsdatenmanagements geboten.

[a] https://tu-dresden.de/zih. Archivierte Version: https://perma.cc/J78G-HKNY.

Institutionelle Infrastruktur bei Szenario 3

An der Humboldt-Universität zu Berlin werden Alex umfangreiche Möglichkeiten der Unterstützung geboten, wie diverse Tools und Informationen vom Computer- und Medienservice[a]. Darüber hinaus gibt es eine Forschungsdatenkoordinationsstelle[b], die mit vielen Hilfestellungen und Querverweisen auf der Webseite dient.

[a] https://www.cms.hu-berlin.de/de. Archivierte Version: https://perma.cc/A8Q7-CCSV.
[b] https://hu-berlin.de/dataman. Archivierte Version: https://perma.cc/4QQD-HJQF.

4.5 FAIR-Prinzipien

Im Jahr 2016 wurden im Journal *Scientific Data* die sogenannten FAIR-Prinzipien (Wilkinson et al., 2016) publiziert. Die Abkürzung steht dabei für *Findable* (auffindbar), *Accessible* (zugänglich), *Interoperable* (interoperabel) und *Reusable* (wiederverwendbar). Die Prinzipien wurden von FORCE11 – einer Gruppe aus Forschenden sowie Personen aus Bibliotheken, Archiven, Verlagen und Forschungsförderern – verfasst und geprägt. Auch wenn sie kein Standard sind, gewinnen sie zunehmend an Bedeutung, da sie im Sinne des Open Science die Nachnutzung der Daten fördern. Mittlerweile sind sie von vielen Forschungsförderern aufgenommen worden und bei einem Antrag verpflichtend (z. B. EC Horizon Europe oder Deutsche Forschungsgemeinschaft, 2019). Das Hauptziel der Prinzipien ist es, die Daten so aufbereitet zu wissen, dass sie sowohl von Menschen als auch von Maschinen lesbar und nachnutzbar sind. Um dieses Ziel zu erreichen, ist es wichtig, nicht erst bei der Publikation der Daten an die Prinzipien zu denken, sondern beim ganzen Forschungsdatenmanagement die FAIR-Prinzipien mitzudenken und umzusetzen (vgl. Engelhardt et al., 2022).

Die FAIR Prinzipien

Findable – Auffindbar

F1. (Meta-)Daten erhalten einen global eindeutigen und dauerhaften Identifikator.

F2. Daten werden mit umfangreichen Metadaten beschrieben (vgl. R1).

F3. Metadaten enthalten eindeutig und explizit den Identifikator der von ihnen beschriebenen Daten.

F4. (Meta-)Daten werden in einer durchsuchbaren Ressource registriert oder indiziert.

Accessible – Zugänglich

A1. (Meta-)Daten können anhand ihres Identifikators unter Verwendung eines standardisierten Kommunikationsprotokolls abgerufen werden.

A1.1 Das Protokoll ist offen, kostenlos und universell implementierbar.

A1.2 Das Protokoll ermöglicht bei Bedarf ein Authentifizierungs- und Autorisierungsverfahren.

A2. Auf Metadaten kann zugegriffen werden, auch wenn die Daten nicht (mehr) verfügbar sind.

Interoperable – Interoperabel

I1. (Meta-)Daten verwenden eine formale, zugängliche, gemeinsame und allgemein anwendbare Sprache für die Wissensrepräsentation.

I2. (Meta-)Daten verwenden Vokabulare, die den FAIR-Prinzipien folgen.

I3. (Meta-)Daten enthalten qualifizierte Verweise auf andere (Meta-)Daten.

Reusable – Wiederverwendbar

R1. (Meta-)Daten werden mit einer Vielzahl genauer und relevanter Attribute ausführlich beschrieben.

R1.1 (Meta-)Daten werden mit einer eindeutigen und zugänglichen Datennutzungslizenz veröffentlicht.

R1.2 (Meta-)Daten sind mit detaillierten Informationen über die Entstehung versehen.

R1.3 (Meta-)Daten entsprechen domänenrelevanten Community-Standards.

Die FAIR-Prinzipien betreffen demnach drei Ebenen: die Daten selbst, ihre Metadaten und die benötigte Infrastruktur. Dabei ist die Idee, die Prinzipien nicht nur auf Daten, sondern auch auf andere digitalen Ressourcen anzuwenden wie Algorithmen oder Tools.

Häufig werden die Konzepte FAIR Data und Open Data[7] als Synonyme verwendet. Auch wenn sie einige Überschneidungen haben, sind sie dennoch unterschiedlich. Nicht alle Daten, die FAIR sind, sind auch Open (siehe A2, z. B. bei Daten, die aus Datenschutzgründen nicht offen zur Verfügung gestellt werden können). Umgekehrt sind nicht alle Daten, die Open sind, auch FAIR (z. B. schreibt die Open Definition (Open Knowledge Foundation, 2021) keinen Identifikator vor).

4.5.1 FAIR-Prinzipien für Software

Auch wenn Software zu Forschungsdaten dazugehört, ist es eine besondere Art von Daten. Sie ist ausführbar, es gibt häufig neue Versionen, wird selten von Beginn an neu geschrieben u. v. m. Dennoch sollen die FAIR-Prinzipien auch auf Forschungssoftware angewendet werden. Lamprecht et al. (2020) haben die FAIR-Prinzipien umformuliert, um diese für Software besser anwendbar zu machen. Im Jahr 2022 wurden die Prinzipien als Ergebnis der Arbeitsgruppe „FAIR for Research Software“ (FAIR4RS WG) der Research Data Alliance (RDA) in einer überarbeiteten Form von N. P. Chue Hong et al. (2022) publiziert.

Die FAIR-Prinzipien für Forschungssoftware (FAIR4RS Principles) nach N. P. Chue Hong et al. (2022)

Findable – Auffindbar

F1. Software hat einen globalen, eindeutigen und dauerhaften Identifikator.

F1.1. Den verschiedenen Komponenten der Software müssen unterschiedliche Identifikatoren zugewiesen werden, die verschiedene Granularitätsebenen darstellen.

F1.2. Unterschiedliche Versionen der gleichen Software müssen unterschiedliche Identifikatoren erhalten.

F2. Software wird mit umfangreichen Metadaten beschrieben.

F3. Metadaten enthalten eindeutig und explizit die Identifikatoren der von ihnen beschriebenen Software.

F4. Metadaten sind FAIR und können durchsucht und indiziert werden.

Accessible – Zugänglich

A1. Software ist über ein standardisiertes Kommunikationsprotokoll über ihren Identifikator zugänglich.

A1.1 Das Protokoll ist offen, kostenlos und universell implementierbar.

A1.2 Das Protokoll ermöglicht bei Bedarf ein Authentifizierungs- und Autorisierungsverfahren.

A2. Metadaten sind zugänglich, auch wenn die Software nicht mehr verfügbar ist.

[7] *Open/offen* bedeutet nach der Open Knowledge Foundation (2021), dass jede:r darauf frei zugreifen, es nutzen, verändern und teilen kann.

Interoperable – Interoperabel

I1. Software liest, schreibt und tauscht Daten in einer Weise aus, die den für den Bereich relevanten Gemeinschaftsstandards entspricht.

I2. Software enthält qualifizierte Verweise auf andere Objekte.

Reusable – Wiederverwendbar

R1. Software ist mit einer Vielzahl genauer und relevanter Attribute ausführlich beschrieben.

R1.1 Software verfügt über klare und zugängliche Nutzungslizenzen.

R1.2 Software ist mit einer detaillierten Provenienz verbunden.

R2. Software enthält qualifizierte Verweise auf andere Software.

R3. Software entspricht den für den Bereich relevanten Community-Standards.

4.5.2 Literaturempfehlungen

- Biernacka, K., Halbherr, V., Lange, M., Martin, L., Mieck, C. & Reimer, N. (2022). *Open Access und wissenschaftliches Publizieren: Train-the-Trainer-Konzept*. https://doi.org/10.5281/zenodo.6034407
- Engelhardt, C., Biernacka, K., Coffey, A., Cornet, R., Danciu, A., Demchenko, Y., Downes, S., Erdmann, C., Garbuglia, F., Germer, K., Helbig, K., Hellström, M., Hettne, K., Hibbert, D., Jetten, M., Karimova, Y., Kryger Hansen, K., Kuusniemi, M. E., Letizia, V., ... Zhou, B. (2022). *How to be FAIR with your data. A teaching and training handbook for higher education institutions*. https://doi.org/10.5281/zenodo.5837500
- Higman, R., Bangert, D., & Jones, S. (2019). Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights the UKSG journal*, 32. https://doi.org/10.3233/DS-190026
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3* (1), 160018. https://doi.org/10.1038/sdata.2016.18.

4.5.3 Anwendung auf die Szenarien

FAIR-Prinzipien bei Szenario 1

Carla erhebt Interviewdaten und sollte sich für deren Publikation intensiv mit den FAIR-Prinzipien auseinandersetzen. Durch die Verwendung von Checklisten (vgl. Seite 182) kann diese Arbeit deutlich erleichtert werden. Sie sollte dabei vor allem auf nachhaltige und offene Formate achten (z. B. .txt für die transkribierten Interviews), es sollte eine Digital Object Identifier (DOI) bei der Publikation vergeben werden und sie sollte die Daten umfangreich mit Metadaten beschreiben.

FAIR-Prinzipien bei Szenario 2

Da Timo im Rahmen seiner Arbeit Software als Forschungsdatum erzeugt, muss er sich an die Richtlinien für FAIR-Prinzipien für Forschungssoftware orientieren. Da seine Software Teil eines Open Source-Projektes ist, sollte hinsichtlich der Wiederverwendbarkeit darauf geachtet werden, dass auch auf andere Software des Projektes verwiesen wird, die mit der eigenen Software in Beziehung steht. Darüber hinaus sollte Timo die Software mit umfangreichen Metadaten und einer ausführlichen Dokumentation beschreiben und ihr einen persistenten Identifikator vergeben. Durch die Verwendung von Checklisten (vgl. Seite 183) können die benötigten Schritte nachvollzogen werden.

FAIR-Prinzipien bei Szenario 3

Alex entwickelt Algorithmen und sollte sich somit an die FAIR-Prinzipien für Software halten. Dabei ist die Verwendung von Checklisten (vgl. Seite 183) behilflich. Um Persistenz und Zitierfähigkeit des Algorithmus zu gewährleisten, sollte sich Alex darum bemühen, dem Code einen DOI zu vergeben (z. B. durch die Publikation über GitLab/Zenodo). Der Algorithmus sollte mit klaren Nutzungslizenzen versehen und mit umfangreichen Metadaten sowie einer Dokumentation beschreiben werden. Da Alex' Forschung ein Bestandteil eines BMBF-Projekts ist, sollte auch diese Information in den Metadaten widergespiegelt werden.

4.6 Datenmanagementplan

Ein Datenmanagementplan (DMP) ist ein Dokument, das das beabsichtigte Forschungsdatenmanagement in einem Forschungsprojekt genau beschreibt. Es beinhaltet alle Informationen zur Erhebung, Aufbereitung, Aufbewahrung, Sicherung, Archivierung, Publikation bis hin zur Löschung der Forschungsdaten im Forschungsvorhaben. In anderen Worten ist es die „[. . .] Analyse des Workflows von der Erzeugung der Daten bis zu deren Nutzung“ (Ludwig & Enke, 2013).

Die Erstellung von DMP ist aus den Anforderungen der Fördermittelgeber entsprungen. Förderer wie die DFG oder das Bundesministerium für Bildung und Forschung (BMBF) erwarten von den Forschenden die Abgabe dieser Informationen entweder bereits bei der

Antragstellung oder in den ersten Monaten des bewilligten Projektes. Auch bei anderen Forschungsvorhaben (z. B. nicht drittmittelfinanzierten oder internen Projekten) oder Abschlussarbeiten ist ein DMP sinnvoll. Es bindet Ressourcen bei der Erstellung, fördert den Gedanken von Open Science und bietet gleichzeitig viele Vorteile. Denn es hilft beim Verständnis der eigenen Daten und ermöglicht, frühzeitig eventuelle Probleme zu erkennen und dafür Lösungen zu finden. Da es ein Dokument ist, in dem alle Verarbeitungsschritte mit den Forschungsdaten beschrieben sind, schafft es in Projekten mit mehreren Personen eine verbindliche Grundlage für einen einheitlichen Umgang mit den Daten, die in dem Projekt anfallen, und erleichtert somit auch die Abstimmung zwischen Projektpartner:innen.

Bei der Erstellung des DMP kann es hilfreich sein, den Prozess einmal rückwärts zu denken, d. h., wo und wie sollen die Daten archiviert bzw. publiziert werden? Aus diesen Überlegungen ergibt sich die Notwendigkeit, frühzeitig im Datenmanagement-Workflow Weichen zu stellen, z. B. bzgl. Formaten, Standards, Metadaten, Lizenzen etc.

4.6.1 Inhalte eines Datenmanagementplans

Der DMP kann unterschiedlich lang ausfallen. Je nachdem, ob das Vorhaben sehr datenreich ist, die Daten sehr hetero- oder homogen sind und wie viele Verarbeitungsschritte notwendig sind, kann die Länge eines DMP zwischen wenigen Absätzen und mehreren Seiten variieren. Grundsätzlich sollte das Dokument folgende Elemente enthalten:

1. Projekttitel, Laufzeit und Forschungsfrage(n)
2. Verantwortliche:r für das Forschungsdatenmanagement
3. Nachnutzung existierender Daten (vgl. Kapitel 4.16)
4. Zu erhebende Daten:
    - Beschreibung der zu erfassenden Daten, Datentypen und -formate
    - erwarteter Speicherbedarf
    - Methoden der Datenerhebung, verwendete Hard- und Software
5. Datenorganisation:
    - Datenspeicherung (vgl. Kapitel 4.10)
    - Back-up (vgl. Kapitel 4.10)
    - Ordnerstruktur (vgl. Kapitel 4.9.1)
    - Namenskonventionen für Ordner und Dateien (vgl. Kapitel 4.9.2)
    - Dokumentation und Metadaten (vgl. Kapitel 4.11)
        - Metadatenstandards
        - Normdaten
        - Kontrolliertes Vokabular
    - Versionierung (vgl. Kapitel 4.9.3)

6. Rechtliche Aspekte, z. B.:
    - Datenschutz (vgl. Kapitel 4.8)
    - Urheberrecht (vgl. Kapitel 4.14)
    - Patentrecht (vgl. Kapitel 4.17)
7. Datenaustausch und -zugang:
    - im Projekt
    - mit externen Partner:innen und Dienstleister:innen
8. Langzeitarchivierung (vgl. Kapitel 4.15)
9. Datenpublikation (vgl. Kapitel 4.13)
10. Qualitätssicherung
11. Zugriff und Nachnutzung (vgl. Kapitel 4.16)
12. Kosten des Datenmanagements

Ein unvollständiger DMP ist besser als gar keiner. Es handelt sich hierbei auch nicht um ein statisches Dokument. Veränderungen des Plans sind nicht ungewöhnlich und Aktualisierungen häufig notwendig. Idealerweise entwickelt sich ein Datenmanagementplan dynamisch: d. h., er wird im Verlauf des Forschungsvorhabens fortlaufend aktualisiert und ausgebaut und entwickelt sich somit von einer Skizze zu einer detaillierten Dokumentation (vgl. Kapitel 4.11).

4.6.2 Besonderheiten eines Softwaremanagementplans

Bei der Entwicklung von (Forschungs-)Software wird häufig vergessen, dass es sich auch dabei um ein Forschungsdatum handelt. Die Anforderungen an die Planung von (Forschungs-) Software werden dabei vernachlässigt. Während mithilfe eines DMP die Strukturen und Ziele der Datenerhebung und -verwaltung definiert werden, kann ein Softwaremanagementplan (SMP) beim Verständnis helfen, was genau entwickelt werden soll, für wen die Software bestimmt ist und wie sie zur Nutzung zur Verfügung gestellt werden kann (The Software Sustainability Institute, 2021).

Ein SMP ist ähnlich detailliert wie ein allgemeinerer DMP. Er enthält Informationen über die Art der zu erstellenden Ergebnisse, wie diese dokumentiert, gespeichert und veröffentlicht werden, und wer für die Verwaltung dieser Aufgaben verantwortlich ist. Ein SMP sollte sich auf die zu erwartenden Software-Outputs konzentrieren und darauf, wie diese während der Entwicklungs- und Lieferphasen verwaltet werden. Nach der Checkliste von The Software Sustainability Institute (2018) ist es empfohlen, unter anderem auf die folgenden Fragen in einem SMP zu antworten:

- Welche Software wird entwickelt?
- Wer sind die vorgesehenen Nutzer:innen der Software?
- Wie wird die Software den Nutzer:innen zur Verfügung gestellt?
- Wie ist der Support für die Software sichergestellt?
- Wie trägt die Software zur Forschung bei?
- Wie steht die Software in Beziehung zu anderen Forschungsobjekten?
- Wie wird der Beitrag der Software zur Forschung gemessen?
- Wo wird die Software hinterlegt, um die langfristige Verfügbarkeit zu gewährleisten?

Eine Hilfe bei der Erstellung eines SMP bietet DMPonline.[8]. Um das richtige Template zu erhalten, ist es notwendig, das Software Sustainability Institute als primäre Organisation auszuwählen und folgend das entsprechende SMP-Template (zur Auswahl stehen ein minimaler sowie vollständiger Plan). Der SMP sollte als lebendiges Dokument betrachtet werden, welches im Laufe der Softwareentwicklung überprüft und angepasst werden sollte.

4.6.3 Literaturempfehlungen

- Alves, R., Bampalikis, D., Castro, L. J., Fernández González, J. M., Harrow, J., Kuzak, M., ... Via, A. (2021). ELIXIR Software Management Plan for Life Sciences. BioHackrXiv. https://doi.org/10.37044/osf.io/k8znb
- CESSDA ERIC. Adapt your Data Management Plan. [Archivierte Version: https://perma.cc/6LDS-8UQ3]. https://www.cessda.eu/content/download/4302/48656/file/TTT_DO_DMPExpertGuide_v1.3.pdf
- Neuroth, H., Engelhardt, C., Klar, J., Ludwig, J. & Enke, H. (2018). Aktives Forschungsdatenmanagement. *ABI Technik*, 38(1): 55–64, 2018. https://doi.org/10.1515/abitech-2018-0008
- The Software Sustainability Institute. (2018). Checklist for a software management plan. https://doi.org/10.5281/zenodo.2159713
- The Software Sustainability Institute. (2021). Writing and using a software management plan. [Archivierte Version: https://perma.cc/4KWS-F3S9]. https://www.software.ac.uk/resources/guides/software-management-plans

[8] https://dmponline.dcc.ac.uk/ Archivierte Version: https://perma.cc/NM8M-UTJU.

4.6.4 Anwendung auf die Szenarien

Für Bachelorarbeiten gibt es bisher keine Vorgaben für die Erstellung eines Daten- oder Softwaremangementplans. Es empfiehlt sich jedoch, auch bei kleineren Projekten und allen Abschlussarbeiten, darüber nachzudenken, welche Daten man benötigt und wie man mit diesen umgeht. Auch die Gute Wissenschaftliche Praxis und die Open-Access-Erklärung ihrer Einrichtung empfehlen einen bewussten Umgang mit Forschungsdaten. Carla hat sich daher entschieden, ihre Überlegungen zu diesem Thema zu verschriftlichen und später ihrer Bachelorarbeit einen Absatz zum Forschungsdatenmangement hinzuzufügen.

Datenmanagementplan für Szenario 1

Geplanter Umgang mit den Forschungsdaten

Es werden circa fünf Interviews von deutschen Universitäten und fünf Interviews von kubanischen Universitäten erhoben. Diese werden als Audiodatei im mp3-Format gespeichert und anschließend transkribiert. Die Transkription erfolgt direkt mit dem Programm MaxQDA, in dem auch die Analyse der Transkripte vorgenommen wird. Da die Daten personenbezogen sind, werden sie auf Servern der Universität Hamburg gespeichert. Gemäß Richtlinien der Guten Wissenschaftlichen Praxis, werden die Daten zehn Jahre lang gespeichert. Der Interviewleitfaden sowie Notizen, die während der Interviews angefertigt werden (bspw. Vorkommnisse wie Unterbrechungen durch weitere Personen), werden gemeinsam mit den Daten gespeichert.

Wie bei Carla (Szenario 1) gibt es bei Timo für die Masterarbeit keine Vorgabe/Pflicht zur Erstellung eines Datenmanagementplans. Da er jedoch mit weiteren Personen und Einrichtungen zusammenarbeitet, empfiehlt es sich, auch hier ein Dokument zum Umgang mit Forschungsdaten zu erstellen, um einen reibungslosen und standardisierten Ablauf innerhalb der Kooperation zu gewährleisten. Da Timos Forschung auf der (Weiter-)Entwicklung einer App beruht, sollte er neben dem DMP auch einen SMP schreiben.

Datenmanagementplan für Szenario 2

Projektbeschreibung

Im Rahmen der Masterarbeit wird eine existierende App weiterentwickeln, die die Anbindung an die Krankenkassen ermöglicht. Die folgenden Funktionen sollen integriert werden: 1.) Mitglieder der Krankenkasse können ihre Abrechnungen sofort einscannen und bekommen ihr Geld erstattet, 2.) wenn Patient:innen die Ärzt:innen wechseln oder zu Fachärzt:innen gehen, sollen sie die Möglichkeit haben, ihre Gesundheitsdaten „zentral“ speichern zu lassen, um sie bei anderen Ärzt:innen abrufen zu können. Um dieses Vorhaben umzusetzen, müssen Schnittstellen zur Finanzabteilung und zur Datenbank mit Patient:innendaten der Krankenkasse geschaffen werden. Im Rahmen dieser Masterarbeit wird die erste Funktion, die Anbindung zur Krankenkasse, umgesetzt.

Primärforscher:in/Wissenschaftler:in

Timo Ryan

Kontakt

Hauptstraße 1, 01097 Dresden

Datenerhebung

Timo Ryan

Datenspeicherung

Die Software wird auf Git gespeichert und ist während der Betaphase als geschlossenes Projekt nicht öffentlich zugänglich.

Datendokumentation

Neben der allgemeinen ReadMe-Datei wird die Software innerhalb des Programmcodes kommentiert. Darüber hinaus wird ein Handbuch geschrieben, das die Nachnutzung und Weiterentwicklung ermöglicht.

Data Sharing

Falls der Austausch von Daten im Projekt unter den Projektpartner:innen erforderlich ist, wird das in einem separaten Vertrag geregelt.

Datenpublikation

Die Software wird nach der abschließenden Testung als Open Source Software auf Git veröffentlicht.

Datenerhalt

Die Software ist auf Zenodo (mit Verbindung zu Git) langzeitverfügbar.

Verantwortlichkeiten und Ressourcen

Timo Ryan ist für die Verfassung des DMP zuständig und publiziert den dokumentierten Sourcecode sowie das Handbuch. Verantwortlich für das Projekt ist Frau Prof. Dr. Ada Ecalevol, die derzeitige Leiterin des Arbeitsbereichs, in dem die Masterarbeit geschrieben wird.

Softwaremanagementplan für Szenario 2

Projektbeschreibung

Im Rahmen der Masterarbeit wird eine existierende App weiterentwickeln, die die Anbindung an die Krankenkassen ermöglicht. Die folgenden Funktionen sollen integriert werden: 1.) Mitglieder der Krankenkasse können ihre Abrechnungen sofort einscannen und bekommen ihr Geld erstattet, 2.) wenn Patient:innen die Ärzt:innen wechseln oder zu Fachärzt:innen gehen, sollen sie die Möglichkeit haben, ihre Gesundheitsdaten „zentral“ speichern zu lassen, um sie bei anderen Ärzt:innen abrufen zu können. Um dieses Vorhaben umzusetzen, müssen Schnittstellen zur Finanzabteilung und zur Datenbank

mit Patient:innendaten der Krankenkasse geschaffen werden. Im Rahmen dieser Masterarbeit wird die erste Funktion, die Anbindung zur Krankenkasse, umgesetzt.

Primärforscher:in/Wissenschaftler:in

Timo Ryan

Kontakt

Hauptstraße 1, 01097 Dresden

Softwareentwickler:in

Timo Ryan

Welche Software wird entwickelt?

In dem Open-Source-Projekt Digital Public Health for All wird die Software DigHealth entwickelt. Sie soll in Form einer App im Gesundheitswesen angewandt werden. Ziel ist, dass Patient:innendaten gesichert und abgerufen werden können. Einerseits soll es möglich sein, dass Patient:innen Zugriff auf ihre Daten erhalten und sie diese einfach verschiedenen Fachärzt:innen digital zur Verfügung stellen können. Auch ein Wechsel zwischen Ärzt:innen soll damit erleichtert werden. Andererseits sollen die Krankenkassen angeschlossen werden, damit der Ärzt:innenbesuch unkompliziert abgerechnet werden kann. Zusätzlich soll auch den Patient:innen dargestellt werden, welche Leistungen der Ärzt:innen von der Krankenkasse übernommen werden können oder wo ggf. Selbstbeteiligungen notwendig sind.

Wer sind die vorgesehenen Nutzer:innen der Software?

Die Software soll im Rahmen einer App von deutschen Krankenkassen genutzt werden.

Wie wird die Software den Nutzer:innen zur Verfügung gestellt?

Die Software wird auf GitHub veröffentlicht und kann von dort gemeinsam mit der Dokumentation heruntergeladen werden.

Wie ist der Support für die Software sichergestellt?

Innerhalb des Open Source-Projektes wird eine Version der Software entwickelt, die im Folgenden von einzelnen Krankenkassen auf ihre Bedürfnisse angepasst werden kann. Dementsprechend muss der Support für einzelne Weiterentwicklungen bei den jeweiligen Nutzer:innen sichergestellt werden.

Wie trägt die Software zur Forschung bei?

Die Software ist in diesem Fall der Forschungsgegenstand.

Wie steht die Software in Beziehung zu anderen Forschungsobjekten?

Es handelt sich um ein neues Open Source-Projekt, in dem bereits eine Software entwickelt wurde, die im Rahmen der App-Entwicklung genutzt und angepasst werden soll. Auch die Masterarbeit von Kayla steht in Beziehung zu Timos Software.

Wie wird der Beitrag der Software zur Forschung gemessen?
Die Software ist der Forschungsgegenstand und wird hinsichtlich der implementierten Features und der Sicherheit gemessen.
Wo wird die Software hinterlegt, um die langfristige Verfügbarkeit zu gewährleisten?
Die Software wird auf GitHub abgelegt und über Zenodo publiziert.

Da Alex an einem BMBF-Verbundprojekt arbeitet, muss er sich an die Anforderungen und Richtlinien des Förderers halten. Da es zum Zeitpunkt (18. August 2021) kein offizielles Template, nur grobe Richtlinien für die Erstellung eines DMP seitens des BMBF gibt, entscheidet sich Alex für die Nutzung des EU Horizon 2020 Template[9], da dies eine ausführliche Beschreibung des Projekts und der dabei anfallenden Daten ermöglicht. Darüber hinaus erstellt Alex einen SMP, um die Besonderheiten und den Nutzen der Software besser darzustellen.

Datenmanagementplan für Szenario 3
Projektname
Learning Analytics to Improve Computer Science Teaching (LA2ICST)
Forschungsförderer
Bundesministerium für Bildung und Forschung (BMBF)
Förderprogramm
Bekanntmachung vom 20. Dezember 2019 des Bundesministeriums für Bildung und Forschung nach der Richtlinie zur Förderung von Forschung über Studienerfolg und Studienabbruch vom 26. November 2019.
Primärforscher:in/Wissenschaftler:in
Alex Müller
ID Primärforscher:in/Wissenschaftler:in
ORCID 1234-5678-1234-5678
Kontakt
alex@hu-berlin.de Unter den Linden 6, 10099 Berlin
Kontaktperson Datenmanagement
forschungsdaten-kontakt@hu-berlin.de
Projektbeschreibung
Das Projekt verfolgt das Ziel, die Abbruchsquote im Informatikstudium zu reduzieren

[9] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template. Archivierte Version: https://perma.cc/G6K5-F7TN.

und die universitäre Lehre zu unterstützen. Dafür wird für ausgewählte Kurse ein eigener Kurs-Hashtag bei Twitter angelegt und die Twitter-Daten der Teilnehmenden werden während des Kurses analysiert und ausgewertet. Das Verbundprojekt wird an mehreren deutschen Universitäten durchgeführt und beschäftigt sich ausschließlich mit der Zielgruppe von Informatikstudierenden. Da in der Informatik viele Studierende das Studium bereits im ersten und zweiten Semester abbrechen, werden insbesondere Grundlagenveranstaltungen im Rahmen des Projekts untersucht.

Erstellungsdatum DMP

18. August 2021

Zu beachtende Vorgaben

Bei dem Umgang mit den Forschungsdaten wird neben den Richtlinien des BMBF, auch die Gute Wissenschaftliche Praxis, die Open-Access-Erklärung der Humboldt-Universität zu Berlin,[a] sowie die Leitlinie zum Umgang mit Forschungsdaten[b] der Universität berücksichtigt, laut der die Forschenden verpflichtet sind, die Forschungsdaten sicher zu speichern, angemessen aufzubereiten und zu dokumentieren sowie langfristig aufzubewahren. Die Verantwortung für die Gewährleistung dieser Prozesse liegt bei den HU-Angehörigen, die das Forschungsvorhaben leiten. Die Forschenden sind darüber hinaus dazu aufgefordert, die in ihrer wissenschaftlichen Tätigkeit entstehenden Forschungsdaten gemäß den im jeweiligen Fachgebiet etablierten Regelungen[c] bzw. Standards aufzubereiten. Gemäß der Open-Access-Erklärung der HU sollten Forschungsdaten ebenso wie die wissenschaftliche Publikation frühestmöglich öffentlich zugänglich gemacht werden.

Datenzusammenfassung

Geben Sie den Zweck der Datenerhebung/-erzeugung an
Anhand der erhobenen Daten soll ermittelt werden, ob ein erhöhtes Studienabbruchrisiko besteht. Es werden dabei Indikatoren in den Tweets identifiziert, die mit einem hohen Abbruchrisiko im Informatikstudium in Zusammenhang gebracht werden können.
Erläutern Sie den Bezug zu den Projektzielen
Anhand der identifizierten Indikatoren soll einerseits frühzeitig erkannt werden, welche Studierenden den Inhalten nicht mehr folgen können und sich unzureichend beteiligen. Andererseits wird ermittelt, bei welchen Lernaktivitäten die Studierenden Hilfestellungen für eine erfolgreiche Kollaboration benötigen.
Spezifizieren Sie die Typen und Formate der erzeugten/gesammelten Daten
Die gesammelten Daten sind Twitter-Datenströme, gesammelt von Informatikstudierenden an den Partneruniversitäten anhand eines vorher festgelegten Kurs-Hashtags. Die Tweets werden zuerst im JSON-Format gespeichert und dann, nach der Auswahl der relevanten Datenfelder, in CSV-Format exportiert.
Geben Sie an, ob vorhandene Daten nachgenutzt werden

In dem Projekt werden neben den eigenen Daten auch Daten aus einem anderen Projekt nachgenutzt, welches sich mit den Faktoren für Studienerfolg und -abbruch für digitale Studienformate befasst. Diese Daten stehen unter der Creative Commons Attribution 4.0 International Lizenz (CC BY) zur Verfügung. Darüber hinaus werden keine Daten nachgenutzt, da diese erst bei der Durchführung der aktuellen Kurse – und somit der Nutzung des Hashtags – generiert werden.

Geben Sie den Ursprung der Daten an

Die nachgenutzten Daten zum Testen des Algorithmus stammen von Twitter und wurden unter der doi: 10.17632/z9zw7nt5h2.1 publiziert. Die eigenen Daten werden über Twitter-Datenströme anhand des vergebenen Hashtags gesammelt.

Geben Sie den erwarteten Umfang der Daten an (falls bekannt)

Durch die Beschränkung auf wesentliche Datenfelder wird der Datenumfang bereits deutlich verringert, da einige Entities Informationen beinhalten, die für die Auswertung unerheblich sind, z. B. grafische Profildarstellung.

Beschreiben Sie den Nutzen der Daten: Für wen werden sie nützlich sein?

Da die Datensammlung aufgrund der rechtlichen Lage und der Bestimmungen von Twitter im Nachhinein nicht ohne Weiteres veröffentlicht werden kann, werden nur die anonymisierten Daten publiziert. Die vollständigen Daten können nur über einen Kooperationsvertrag zur Verfügung gestellt werden. Die Daten werden daher vorrangig den Wissenschaftler:innen in dem Projekt LA2ICST für die Klärung der Forschungsfragen und erst bei Folge- und Kooperationsprojekten für weitere Wissenschaftler:innen nützlich sein.

Auffindbarkeit von Daten, inklusive Metadatenvergabe (FAIR Data)

Darstellung der Auffindbarkeit von Daten (Bereitstellung von Metadaten)

Die Metadaten werden nach einem vorher festgelegten Metadatenschema vergeben. Dazu gehören unter anderem die folgenden Informationen: Universitätszugehörigkeit, Semester, Geschlecht, Disziplin, Grundlagenveranstaltung.

Beschreiben Sie die Identifizierbarkeit von Daten und verweisen Sie auf Standard-Identifizierungsmechanismen. Verwenden Sie dauerhafte und eindeutige Identifikatoren wie Digital Object Identifiers?

Die Daten werden bei VerbundFDB[d] geteilt. Dort erhalten sie einen Digital Object Identifier als persistenten Identifikator.

Skizzieren Sie die verwendeten Benennungskonventionen

Für die Verarbeitung und Analyse der Daten wird die Benennungskonvention einen Datums- und Zeitstempel des jeweiligen Tweets enthalten. Zusätzlich wird ein Datei-Präfix eingefügt mit dem Pseudonym der tweetenden Person.

Bei der Publikation der Forschungsdaten werden die Dateinamen anonymisiert und mit dem Namen des Projekts und einer zufallsgenerierten Nummer ersetzt.

Skizzieren Sie den Ansatz für mögliche Schlüsselwörter (Keywords)
Mögliche Schlüsselwörter wären: Learning Analytics, Studienabbruch, Informatikstudium.
Beschreiben Sie den Ansatz für eine klare Versionierung
Die Versionierung erfolgt durch die Vergabe von Versionsnummer und Änderungsdatum im Dateinamen. Darüber hinaus wird eine Versionskontrolltabelle geführt, in der zusätzlich zu der Information, wann und wer die Daten bearbeitet hat, auch eine kurze Notiz mit der Beschreibung enthalten ist, was verändert worden ist.
Geben Sie Standards für die Erstellung von Metadaten an (falls vorhanden). Wenn es in Ihrem Fachbereich keine Standards gibt, beschreiben Sie, welche Metadaten erstellt werden und wie
Da es für Learning-Analytics-Daten noch keinen Metadatenstandard gibt, werden Elemente von den Standards Dublin Core[e] und LOM[f] kombiniert, um eine hinreichend gute Beschreibung der Daten zu erhalten.

Datenzugänglichkeit (FAIR Data)

Geben Sie an, welche Daten öffentlich zugänglich gemacht werden? Falls einige Daten nicht zugänglich gemacht werden, begründen Sie dies
Es können nicht alle Twitter-Datensammlungen zugänglich gemacht werden, da dies den Nutzungsrechten von Twitter widersprechen würde. Es wird nur die anonymisierte Twitter-Datensammlung publiziert. Die vollständigen, pseudonymisierten Daten können auf Anfrage für Kooperationen und Folgeprojekte zur Verfügung gestellt werden.
Geben Sie an, wie die Daten zur Verfügung gestellt werden sollen
Die Daten werden bei VerbundFDB publiziert sowie über deren Webseite mit den notwendigen Autorisierungs- und Authentifizierungsschritten zur Verfügung gestellt.
Geben Sie an, welche Methoden oder Software für den Zugriff auf die Daten erforderlich sind. Ist eine Dokumentation über die für den Zugang zu den Daten erforderliche Software enthalten? Ist es möglich, die entsprechende Software einzubinden (z. B. als Open-Source-Code)?
Da die Daten in einem offenen Format (CSV) gespeichert werden, ist keine zusätzliche Software notwendig, um diese Daten lesen zu können.
Geben Sie an, wo die Daten und die zugehörigen Metadaten, die Dokumentation und der Code hinterlegt werden
Die Forschungsdaten, inklusive Metadaten und der Dokumentation, werden bei VerbundFDB abgelegt.
Geben Sie an, wie der Zugang im Falle von Einschränkungen gewährleistet wird
Der Zugang zu den nicht anonymisierten Daten ist eingeschränkt und erfolgt erst nach Abschluss eines Kooperationsvertrages.

Interoperabilität von Daten (FAIR Data)

Bewerten Sie die Interoperabilität Ihrer Daten. Legen Sie fest, welche Daten- und Metadatenvokabulare, Standards oder Methoden Sie anwenden werden, um die Interoperabilität zu erleichtern

Da es für Learning Analytics keine Metadatenstandards gibt, wird ein eigenes Metadatenschema anhand von Dublin Core und LOM erstellt. Das Vokabular basiert dabei sowohl auf dem kontrollierten Vokabular der Bildungswissenschaften als auch der Informatik.

Geben Sie an, ob Sie ein kontrolliertes Vokabular für alle in Ihrem Datensatz vorhandenen Datentypen verwenden werden, um eine interdisziplinäre Interoperabilität zu ermöglichen. Wenn nicht, werden Sie ein Mapping auf allgemeinere Ontologien bereitstellen?

Zur Gewährleistung der intradisziplinären Interoperabilität der Daten wird auf die in der Community üblichen Vokabulare und Standards zurückgegriffen, wie LOM aus den Bildungswissenschaften.

Nachnutzung von Daten (durch Lizenzvergabe) (FAIR Data)

Legen Sie fest, wie die Daten lizenziert werden sollen, um eine möglichst umfassende Nachnutzung zu ermöglichen

Aufgrund der Besonderheiten der Daten können keinen offenen Lizenzen, wie Creative Commons vergeben werden. Die Daten werden anhand individueller Verträge zur Nachnutzung übergeben.

Geben Sie an, wann die Daten zur Nachnutzung zur Verfügung gestellt werden. Geben Sie gegebenenfalls an, warum und für welchen Zeitraum ein Embargo erforderlich ist

Es können nicht alle Twitter-Datensammlungen zugänglich gemacht werden, da dies den Nutzungsrechten Twitters widersprechen würde. Deshalb wird nur die anonymisierten Twitter-Datensammlung nach Beendigung der Projektlaufzeit publiziert. Die Publikation wird erst zu diesem Zeitpunkt vorgenommen, weil die gründliche Aufbereitung und Anonymisierung der Daten einen großen Zeitraum beansprucht. Die vollständigen, pseudonymisierten Daten können auf Anfrage für Kooperationen und Folgeprojekte zur Verfügung gestellt werden.

Geben Sie an, ob die im Rahmen des Projekts erstellten und/oder verwendeten Daten von Dritten genutzt werden können, insbesondere nach Abschluss des Projekts. Falls die Weiterverwendung einiger Daten eingeschränkt ist, erläutern Sie, warum

Die im Projekt selbst erstellten Daten können bei Kooperationen nach Unterzeichnung eines Kooperationsvertrages wiederverwendet werden.

Beschreiben Sie den Prozess der Datenqualitätssicherung

Die Datenqualität wird auf zwei Wegen sichergestellt. Einerseits aus der technischen

Sicht, wo auf offene und gängige Formate geachtet wird, die eine spätere eventuell notwendige Migration einfacher gestalten und die Lesbarkeit der Daten gewährleisten. Darüber hinaus wird hierbei auch eine gute Dokumentation und Beschreibung mit Metadaten angestrebt. Andererseits wird bei der inhaltlichen Perspektive darauf geachtet, dass nur Tweets gesammelt werden, die den notwendigen Hashtag aufweisen und dessen Twitter-User Informatikstudierende an den teilnehmenden Universitäten sind. Diese werden in anonymisierter Form zur Verfügung gestellt.

Geben Sie an, wie lange die Daten weiterverwendet werden können

Die Dauer der Nachnutzung der Daten hängt mit der Dauer des Kooperationsprojektes zusammen.

Ressourcenzuweisung

Schätzen Sie die Kosten, die notwendig sind, um Ihre Daten FAIR zu machen. Beschreiben Sie, wie Sie dies finanzieren möchten

Die meisten anfallenden Aufgaben werden von den am Projekt arbeitenden Forschenden durchgeführt. Aufgaben, die vom Infrastrukturpersonal durchgeführt werden müssen, werden als Eigenanteil angerechnet. VerbundFDB bietet den Service der Sicherung und Bereitstellung von Forschungsdaten kostenfrei an. Es sind also keine zusätzlichen Mittel notwendig, um die Forschungsdaten FAIR zu machen.

Legen Sie die Verantwortlichkeiten für das Datenmanagement in Ihrem Projekt klar fest

Für das Datenmanagement in diesem Projekt sind mehrere Personen zuständig.
Für das Teilprojekt 1: Alex Müller
Für das Teilprojekt 2: Dr. Anna Gutmeier
Für das Teilprojekt 3: Klaus Mittermeier

Beschreiben Sie die Kosten und den potenziellen Wert einer Langzeitarchivierung

Die Forschungsdaten werden bei VerbundFDB für mindestens zehn Jahre aufbewahrt. Auch dieser Service ist gebührenfrei.

Datensicherheit

Beschreiben Sie den Prozess der Datenwiederherstellung sowie sicherer Speicherung und Übertragung von sensiblen Daten

Die Infrastruktur der Humboldt-Universität zu Berlin bietet eine Cloud-Lösung an (HU-Box[g]; basierend auf Seafile[h]), die für dieses Projekt genutzt wird. Die Daten werden dort in einer verschlüsselten Bibliothek abgelegt. Der Zugang wird nur den Personen gegeben, die unmittelbar mit diesen Daten arbeiten (müssen).

Die Cloud-Lösung ist an das institutionelle Back-up-System angebunden. Dieses wiederum basiert auf IBM Spectrum Protect™[i]. Das Back-up wird jede Nacht durchgeführt und 60 Tage lang gespeichert. Danach werden die Back-ups gelöscht. Die Wiederherstellung der Daten kann von den Admins angestoßen werden.

Ethische Aspekte

Gibt es ethische Fragen, die sich auf die gemeinsame Nutzung von Daten auswirken können? Diese können auch im Zusammenhang mit der Ethikprüfung erörtert werden. Gegebenenfalls sind in der Beschreibung der Maßnahme (DoA) Verweise auf die Ethikleistungen aufzunehmen.
Bei dieser Datenerhebung werden Stimmungen von Studierenden analysiert und es können beispielsweise Rückschlüsse auf psychische Erkrankungen gezogen werden. Aus diesem Grund sollte eine Absprache mit der Ethikkomission erfolgen.

Weitere Aspekte

Um die Daten für wissenschaftliche Zwecke sammeln zu können, wird einerseits die Einwilligung der Studierenden, andererseits ein Datennutzungsvertrag mit Twitter erforderlich sein. Die Einholung dieser Genehmigungen erfolgt zu Beginn des Forschungsprojektes.

[a] https://edoc.hu-berlin.de/e_info/oa-erklaerung.php. Archivierte Version: https://perma.cc/HUT9-J7M9.
[b] https://www.cms.hu-berlin.de/de/dl/dataman/infos/policy. Archivierte Version: https://perma.cc/388W-9SMK.
[c] http://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/. Archivierte Version: https://perma.cc/KK4Q-TD3Y.
[d] www.forschungsdaten-bildung.de. Archivierte Version: https://perma.cc/23RQ-8BZD.
[e] http://dublincore.org/documents/dcmi-terms/. Archivierte Version: https://perma.cc/WH42-9VDA.
[f] IEEE (2020).
[g] https://blogs.hu-berlin.de/hu-box/. Archivierte Version: https://perma.cc/YG4Q-ZT9P.
[h] https://www.seafile.com/en/home/. Archivierte Version: https://perma.cc/98QW-J988.
[i] https://www.ibm.com/de-de/products/data-protection-and-recovery. Archivierte Version: https://perma.cc/HHT5-AREY.

Softwaremanagementplan für Szenario 3

Projektbeschreibung

Das Projekt verfolgt das Ziel, die Abbruchquote im Informatikstudium zu reduzieren und die universitäre Lehre zu unterstützen. Dafür wird für ausgewählte Kurse ein eigener Kurs-Hashtag bei Twitter angelegt und es werden die Twitter-Daten der Teilnehmenden während des Kurses analysiert und ausgewertet. Das Verbundprojekt wird an mehreren deutschen Universitäten durchgeführt und beschäftigt sich ausschließlich mit der Zielgruppe von Informatikstudierenden. Da in der Informatik viele Studierende das Studium bereits im ersten und zweiten Semester abbrechen, werden insbesondere Grundlagenveranstaltungen im Rahmen des Projekts untersucht.

Primärforscher:in/Wissenschaftler:in

Alex Müller

Kontakt
alex@hu-berlin.de Unter den Linden 6, 10099 Berlin
Softwareentwickler:in
Alex Müller
Welche Software wird entwickelt?
In dem BMBF-Projekt „Learning Analytics to Improve Computer Science Teaching (LA2ICST)“ wird ein Algorithmus entwickelt, der Twitter-Datenströme auswertet. Verbunden mit Hashtags zur Lehrveranstaltung sollen die Tweets von Teilnehmenden hinsichtlich ihrer Stimmungen ausgewertet werden. Dieser Algorithmus soll genutzt werden, um Maßnahmen zu entwickeln, damit die Abbruchquote von Studierenden gesenkt werden kann.
Wer sind die vorgesehenen Nutzer:innen der Software?
Dieser Algorithmus soll in erster Linie für die Forscher:innen zur Analyse von Twitter-Datenströmen verwendet werden.
Wie wird die Software den Nutzer:innen zur Verfügung gestellt?
Der Algorithmus wird auf GitHub zusammen mit einer Dokumentation publiziert.
Wie ist der Support für die Software sichergestellt?
Der Algorithmus ist frei verfügbar und wird nach Ende der Projektlaufzeit nicht weiter supportet.
Wie trägt die Software zur Forschung bei?
Der Algorithmus ist der Forschungsgegenstand, aber auch das Mittel, um Forschungsdaten zu erheben. Twitter-Datenströme werden mithilfe des Algorithmus gesammelt und die Daten unter Nutzung von Machine-Learning-Verfahren ausgewertet.
Wie steht die Software in Beziehung zu anderen Forschungsobjekten?
Die Software wird in einem eigenständigen BMBF-Projekt entwickelt.
Wie wird der Beitrag der Software zur Forschung gemessen?
Es wird evaluiert, wie zuverlässig relevante Daten zur Beurteilung der Stimmungen von Studierenden aus den Datenströmen extrahiert werden.
Wo wird die Software hinterlegt, um die langfristige Verfügbarkeit zu gewährleisten?
Der Algorithmus wird auf GitHub abgelegt und über Zenodo publiziert.

4.7 Ethische Aspekte

Ein verantwortungsvoller Umgang mit Forschungsdaten wird von den Forschenden erwartet. Dazu gehört auch die Betrachtung, ob die Nutzung von Daten Schaden für Menschen,

Tiere oder die Welt mit sich zieht. Häufig werden diese ethischen Fragen gar nicht erst gestellt, sondern nur auf das geltende Recht verwiesen. Dabei ist nicht alles, was legal ist, auch ethisch akzeptabel. Ethik hat zum Ziel, Wertestandards zu etablieren und bei der Suche nach moralisch vertretbarem Handeln Orientierung zu verschaffen (Rösch, 2021).

Forschungsethische Fragen finden sich in allen Phasen des Forschungsdatenlebenszyklus wieder. Die Forschenden sind daher immer dazu aufgefordert, sich bereits vor der Erhebung der Daten diesen Fragen zu stellen und Risiken zu vermindern. Darüber hinaus muss der potenzielle Missbrauch der Daten unterbunden werden. Die Forschenden müssen prüfen, welche Schäden oder Gefährdungen eine missbräuchliche Verwendung der Daten mit sich bringen könnte. Das Votum der Ethikkomission ist dabei nicht allein ausschlaggebend. Sie betrachtet nur, ob es während des Prozesses zu Schäden kommt. Was nach der Publikation der Daten passiert, ist nicht in ihrem Fokus (Rösch, 2021, S. 3). Bei der Archivierung von personenbezogenen und sensiblen Daten müssen darüber hinaus auch die Regelungen zur datenschutzrechtlichen Zulässigkeit der Verarbeitung nach Art. 9 der DSGVO eingehalten werden (vgl. Kapitel 4.8).

4.7.1 Forschungsethische Fragen in der Informatik

Die moderne Technik bringt viele Chancen mit sich, jedoch verursacht sie auch völlig neue ethische Probleme, die ohne die Erfindung von Computern nicht bestanden hätten. Sie bringt vor allem die Frage hervor: Wie wirken sich Computertechnologien auf die Gesellschaft aus? Die Gesellschaft für Informatik e. V. (2004) sagt in ihren ethischen Leitlinien:

> *„Das Handeln von Informatikerinnen und Informatikern steht in Wechselwirkung mit unterschiedlichen Lebensweisen, deren besondere Vielfalt sie berücksichtigen sollen. Mehr noch sehen sie sich dazu verpflichtet, allgemeine moralische Prinzipien, wie sie in der Allgemeinen Deklaration der Menschenrechte formuliert sind, zu wahren. Diese Leitlinien sind Ausdruck des gemeinsamen Willens, diese Wechselwirkungen als wesentlichen Teil des eigenen individuellen und institutionellen beruflichen Handelns zu betrachten.“*

Informationsethik befasst sich im Speziellen mit dem Umgang mit Informationen sowie Informations- und Kommunikationstechnologien unter moralischen und ethischen Perspektiven. Dazu zählen:

- Netzethik,
- Medienethik,
- Computerethik.

Da das technische Handeln eines Individuums einen großen Einfluss auf die Gesellschaft haben kann (Weber-Wulff et al., 2009) und viele Aspekte des Alltags beeinflusst, muss es die Bestrebung geben, Werte und Grundrechte zu schützen, wie etwa Freiheit, Sicherheit und Bildung. Immer häufiger wird der Datenschutz, die Computerkriminalität, die Technikabhängigkeit, das geistiges Eigentum bzw. Urheberrecht, die Privatsphäre und Anonymität diskutiert.

Vor allem komplexe Systeme (z. B. KI-basierte Lösungen), deren Ergebniswege und Aktionen intransparent sind, führen zu großer Unsicherheit und Bedenken in der Gesellschaft. Ein KI-verursachter Schaden kann vorliegen, wenn eine erarbeitete Vorhersage oder ein Ergebnis die Fähigkeit eines Individuums, seine rechtmäßige Persönlichkeit zu etablieren, negativ beeinflusst und dies im Ergebnis dazu führt, dass seine Fähigkeit, auf Ressourcen zuzugreifen, beeinflusst oder beeinträchtigt wird (Crawford, 2017). Dabei kann es sich um Repräsentations- bzw. Allokationsschäden handeln.

Um gefährliche Fehlentwicklungen in der IT zu vermeiden, wurde das „Digitale Manifest“ von Helbing et al. (2017) publiziert, welche die folgenden Grundprinzipien definiert:

1. Die Funktion von Informationssystemen stärker zu dezentralisieren,
2. informationelle Selbstbestimmung und Partizipation zu unterstützen,
3. Transparenz für eine erhöhte Vertrauenswürdigkeit zu verbessern,
4. Informationsverzerrungen und -verschmutzung zu reduzieren,
5. von den Nutzer:innen gesteuerte Informationsfilter zu ermöglichen,
6. gesellschaftliche und ökonomische Vielfalt zu fördern,
7. die Fähigkeit technischer Systeme zur Zusammenarbeit zu verbessern,
8. digitale Assistenten und Koordinationswerkzeuge zu kreieren,
9. kollektive Intelligenz zu unterstützen, und
10. die Mündigkeit der Bürger:innen der digitalen Welt zu fördern – eine „digitale Aufklärung“.

Diese können als Richtlinie und Orientierung für zukünftige IT-Projekte genutzt werden.

4.7.2 CARE-Prinzipien

Während sich die FAIR-Prinzipien (vgl. Kapitel 4.5) auf Merkmale von Daten, die einen verstärkten Datenaustausch erleichtern sollen, konzentrieren, lassen sie gleichzeitig ethische Fragestellungen, Machtunterschiede und historische Kontexte außen vor. Aus diesem Grund wurden komplementär die CARE Principles for Indigenous Data Governance von der Research Data Alliance International Indigenous Data Sovereignty Interest Group (2019) veröffentlicht (vgl. Abbildung 4.7). Sie betrachten den Nutzen der Forschung für die Beforschten nicht lediglich als Abwägungsaspekt, sondern als wesentliches Motiv. Beforschte stehen dabei weniger als Individuen im Fokus, sondern als Kollektiv (Deppe, 2020).

Abbildung 4.7: FAIR and CARE (Research Data Alliance International Indigenous Data Sovereignty Interest Group, 2019)

Die CARE-Prinzipien für indigene Data Governance (Übersetzung von Carroll et al., 2019)

Kollektiver Nutzen (Collective Benefit)

C1. Für integrative Entwicklung und Innovation.

C2. Für eine bessere Steuerung und Bürger:innenbeteiligung.

C3. Für gerechte Ergebnisse.

Recht auf Kontrolle über die Daten (Authority Control)

A1. Anerkennung von Rechten und Anteilen.

A2. Daten zur (Selbst-)Verwaltung.

A3. Data Governance.

Verantwortung (Responsibility)

R1. Für positive Beziehungen.

R2. Für die Erweiterung von Fähigkeiten und Kompetenzen.

R3. Für indigene Sprachen und Weltanschauungen.

Ethik (Ethics)

E1. Für die Minimierung des Schadens und Maximierung des Nutzens.

E2. Für Gerechtigkeit.

E3. Für künftige Nutzung.

Daten, somit auch Forschungsdaten, haben wichtige Auswirkungen auf die Fähigkeit indigener Völker, ihre individuellen und kollektiven Rechte auf Selbstbestimmung auszuüben. Indigene Völker sind oft von der Entscheidungsfindung ausgeschlossen und ihr Wissen wird

an den Rand gedrängt, wenn dieses Wissen nur als Teil einer mündlichen Tradition existiert (Research Data Alliance International Indigenous Data Sovereignty Interest Group, 2019).

4.7.3 Literaturempfehlungen

- Imeri, & Rizzolli, M. (2022).CARE Principles for Indigenous Data Governance. Eine Leitlinie für ethische Fragen im Umgang mit Forschugnsdaten? *O-Bib. Das Offene Bibliotheksjournal, 9*(2), S. 1–14. VDB. https://doi.org/10.5282/o-bib/5815
- Rösch, H. (2021). Forschungsethik und Forschungsdaten. In M. Putnings, H. Neuroth & J. Neumann (Hrsg.), *Praxishandbuch Forschungsdatenmanagement* (S. 115—140). De Gruyter Saur. https://doi.org/10.1515/9783110657807-006
- Weber-Wulff, D., Class, C., Coy, W., Kurz, C. & Zehllhöfer, D. (2009). *Gewissensbisse. Ethische Probleme in der Informatik. Biometrie - Datenschutz - geistiges Eigentum.* transcript.

4.7.4 Anwendung auf die Szenarien

Ethische Aspekte bei Szenario 1

Bei diesem Szenario ist darauf zu achten, dass über Studierende der einzelnen Universitäten nicht geurteilt wird, wenn sie beispielsweise nicht in der Lage sind, kollaborativ zu arbeiten. Eine Verallgemeinerung von einzelnen Personen hin zu einer gesamten Universität ist auch auf Basis der Forschungsdaten unmöglich. Auch wenn in den Interviews Informationen zu konkreten Kommiliton:innen oder Dozierenden angesprochen werden, sind diese vertraulich zu behandeln. Es sollte prinzipiell verantwortungsvoll mit den Daten umgegangen werden, damit auch das Vertrauen in die Forschenden bestehen bleibt.
Insbesondere bei der Zusammenarbeit mit weiteren Ländern ist auf die politische Situation der Länder zu achten und darauf, wie beispielsweise mit der Meinungsfreiheit umgegangen wird. Kann eine öffentliche Äußerung von Regierungskritik Konsequenzen für die interviewten Personen nach sich ziehen, dann handelt es sich um besonders schützenswerte Daten. Die Verarbeitung, Anonymisierung und Speicherung muss sorgfältig abgewogen werden.
Bei der Zusammenarbeit mit Institutionen wie der deutschen und kubanischen Universität oder auch dem Lateinamerika-Forum ist darauf zu achten, dass die Veröffentlichungen den Ruf und die Außenwirkung der Institutionen prägen können.

Ethische Aspekte bei Szenario 2

Beim Schreiben vom Algorithmen muss besonders berücksichtigt werden, dass Personengruppen und Minderheiten durch den Algorithmus nicht diskriminiert werden. Ein mögliches Beispiel für eine Diskriminierung durch den Algorithmus ist, wenn die (passwortgeschützte) programmierte App auch via Gesichtserkennung entsperrt wer-

den kann, diese jedoch bei People of Color (PoC) oder verschiedenen Geschlechtern nur unzuverlässig funktioniert (abhängig von den Trainingsdaten für den Algorithmus). Auch könnte die App nur kompatibel für bestimmte Maße von Smartphones verfügbar sein, sodass finanziell schlechter gestellte Personen mit einem kleinen Display oder älterem Modell diskriminiert werden.

Ethische Aspekte bei Szenario 3

Die Auswertung von Twitter-Daten ist rein rechtlich betrachtet zunächst zulässig. Jedoch ist von ethischer Seite aus zu diskutieren, ob es vertretbar ist, da die Personen vermutlich keine Kenntnis davon haben, dass gemäß Twitters AGB diese Auswertung zulässig ist. Es sollte deshalb genau betrachtet werden, welche Daten von Twitter genutzt werden und ob die Nutzung anderen Personen schaden kann.

4.8 Datenschutz

Spätestens seit der Einführung der neuen DSGVO der Europäischen Union[10] im Mai 2018 ist das Thema Datenschutz[11] bei den Forschenden präsent. Die DSGVO führt weitgehend zu einer Vereinheitlichung des europäischen Datenschutzrechtes und hat es als Ziel, den Schutz der Grundrechte und Grundfreiheiten natürlicher Personen, insbesondere deren Recht auf Schutz personenbezogener Daten, zu gewährleisten. Das Datenschutzrecht ist darüber hinaus auch in den Datenschutzgesetzen der Länder und im Bundesdatenschutzgesetz (BDSG) niedergeschrieben.

4.8.1 Personenbezogene Daten

Die Vorgaben des Datenschutzrechts betreffen alle Aspekte des Forschungsdatenmanagements: von der Erhebung, Speicherung, Verarbeitung, Publikation bis hin zur Löschung personenbezogener Daten.

Personenbezogene Daten

Nach Art. 4 Nr. 1 DSGVO werden als personenbezogene Daten „alle Informationen bezeichnet, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen; als identifizierbar werden Personen angesehen, wenn sie direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, psychologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Personen sind, identifiziert werden können“.

[10] https://dsgvo-gesetz.de/. Archivierte Version: https://perma.cc/QQH8-T8QU.

[11] Der Begriff „Datenschutz“ ist von dem Begriff „Datensicherheit“ eindeutig zu trennen. Während sich Ersteres mit dem Schutz personenbezogener Daten vor etwaigem Missbrauch durch Dritte befasst, behandelt Letzteres den Schutz von Daten hinsichtlich gegebener Anforderungen an deren Vertraulichkeit, Verfügbarkeit und Integrität (Bundesamt für Sicherheit in der Informationstechnik, 2022).

Personenbezogene Daten sind demnach alle Informationen, die dazu dienen, eine natürliche Person zu identifizieren. Die Identifizierung kann entweder direkt oder indirekt erfolgen. Dabei handelt es sich bei den direkten Identifikatoren um Merkmale, deren Ausprägung einer Person entweder eindeutig oder nahezu eindeutig zuzuordnen sind (z. B. Name). Bei den indirekten Identifikatoren (auch Quasi-Identifikatoren genannt) handelt es sich um Merkmale, die zwar allein keine Identifikation zulassen, jedoch kombiniert mit anderen Daten die Identifikation ermöglichen (z. B. Geburtsdatum). Im Erwägungsgrund 26 der DSGVO werden die Voraussetzungen für die Identifizierbarkeit folgendermaßen beschrieben:

> *„Um festzustellen, ob eine natürliche Person identifizierbar ist, sollten alle Mittel berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren, wie beispielsweise das Aussondern. Bei der Feststellung, ob Mittel nach allgemeinem Ermessen wahrscheinlich zur Identifizierung der natürlichen Person genutzt werden, sollten alle objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, herangezogen werden, wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologischen Entwicklungen zu berücksichtigen sind.“*

In der Menge der personenbezogenen Daten ist eine besondere Kategorie von Daten enthalten, die eines besonderen Schutzes bedürfen – die sensiblen Daten.

Sensible Daten

Sensible Daten sind eine besondere Kategorie der personenbezogenen Daten, die eines erhöhten Schutzes bedürfen. Dazu zählen: die ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen, Gewerkschaftszugehörigkeit, genetische und biometrische Daten, Gesundheitsdaten, Daten zum Sexualleben oder der sexuellen Orientierung. Des Weiteren gehören dazu personenbezogene Daten über strafrechtliche Verurteilungen und Straftaten.

Angesichts der technischen Entwicklungen, vor allem durch leistungsstarke Analyseprogramme, die eine zunehmend leichte Verknüpfung einzelner Daten ermöglichen, die korreliert zur Identifizierung einer Person führen können, kann es unter Umständen schwierig sein, die Grenze zwischen personenbezogenen und nicht personenbezogenen Forschungsdaten zu ziehen (Baumann et al., 2021).

4.8.2 Informierte Einwilligung

Die DSGVO untersagt grundsätzlich die Verarbeitung von personenbezogenen Daten. Eine Ausnahme dazu ist, wenn die betroffene Person in die Verarbeitung der personenbezogenen Daten für einen oder mehrere festgelegte Zwecke ausdrücklich eingewilligt hat. Damit eine Einwilligung wirksam ist, muss sie vor allem zwei Voraussetzungen erfüllen:

- **Informiertheit** – die Einwilligung muss in einer „informierten Weise“ (Art. 4 Nr. 11 DSGVO) abgegeben werden. Darin wird die Person über ihre Rechte, die Verarbeitung

ihrer Daten, deren Verwendung sowie den Studienzweck aufgeklärt. Erst anhand dieser Informationen in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache willigt die Person ein, unter diesen Bedingungen an der Studie teilzunehmen. Die Information kann sowohl mündlich (z. B. bei Kindern oder Analphabeten) als auch schriftlich oder elektronisch erfolgen.

- **Freiwilligkeit** – die Einwilligung muss freiwillig erteilt worden sein (Art. 4 Nr. 11 DSGVO). Freiwilligkeit liegt auch dann vor, wenn die Personen für ihre Teilnahme an einer Studie eine angemessene finanzielle Aufwandsentschädigung erhalten.

Zudem muss eine Einwilligung bei sensiblen Daten „ausdrücklich“ erfolgen. Dies führt dazu, dass Einwilligungserklärungen mit Hinblick auf sensible Daten grundsätzlich schriftlich eingeholt bzw. elektronische Einwilligungen protokolliert werden sollten (Baumann et al., 2021). Eine schriftliche Einwilligung erfüllt darüber hinaus eine Warn- und Schutzfunktion für die Betroffenen.

Die informierte Einwilligung muss präzise Angaben zu den Verarbeitungsschritten der personenbezogenen Daten enthalten. Dies führt gerade bei wissenschaftlichen Projekten häufig zu Problem, da bestimmte Ergebnisse, Folgeprojekte oder Kooperationen im Allgemeinen nicht bereits bei der Datenerhebung geklärt sein können. Dafür sieht die DSGVO eine Ausnahme vor, das sogenannte „broad consent“ (Erwägungsgrund 33 DSGVO). Demnach sollte es

> *„betroffenen Personen erlaubt sein, ihre Einwilligung für bestimmte Bereiche wissenschaftlicher Forschung zu geben, wenn dies unter Einhaltung der anerkannten ethischen Standards der wissenschaftlichen Forschung geschieht. Die betroffenen Personen sollten Gelegenheit erhalten, ihre Einwilligung nur für bestimmte Forschungsbereiche oder Teile von Forschungsprojekten in dem vom verfolgten Zweck zugelassenen Maße zu erteilen.“*

Nach Art. 7 Abs. 3 S. 1 DSGVO kann die Einwilligungserklärung jederzeit und grundlos widerrufen werden. Das Widerrufsrecht wirkt zwar nicht rückwirkend, hat jedoch einen großen Einfluss auf alle zukünftigen Verarbeitungsprozesse. Denn bei Widerruf müssen alle Daten unverzüglich gelöscht werden. Dies kann z. B. errechnete Ergebnisse kurz vor der Publikation erheblich verändern. Es gibt zwar spezielle Erlaubsnisbestände für wissenschaftliche oder historische Forschungszwecke, der sichere Weg jedoch ist eine frühzeitige Anonymisierung der Daten, da anonymisierte Daten nicht der DSGVO unterliegen und somit unter Beachtung eventueller sonstiger rechtlicher Vorgaben (z. B. Patent- oder Urheberrecht) frei verwendet werden können.

4.8.3 Anonymisierung und Pseudonymisierung

Anonymisierung

Anonymisierung bedeutet die Daten so zu verändern, dass Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr (sog. absolute Anonymisierung) oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft (sog. faktische Anonymisierung) einer identifizierten oder identifizierbaren natürlichen Person zugeordnet werden können.

Es können unterschiedliche Methoden zur Anonymisierung genutzt werden, z. B. das Ersetzen der persönlichen Informationen mit einer Beschreibung oder Fantasiewerten, Schwärzen oder Aggregation der Angaben. Zur Gewährleistung absoluter Anonymität müssen alle sowohl direkten als auch indirekten Identifizierungsmerkmale gelöscht oder unkenntlich gemacht werden, sodass die Daten im Nachhinein auch nicht mehr zurückzuführen sind. Dabei müssen auch solche Informationen und solches Wissen mitgedacht werden, welches während der vorgesehenen Speicherdauer durch verfügbare Technologien oder technologische Entwicklungen beschafft werden können (Baumann et al., 2021).

Im Gegensatz dazu werden bei der faktischen Anonymisierung ausschließlich die direkten Identifizierungsmerkmale entfernt. Indirekte Merkmale werden erst dann verändert, wenn die Hinzunahme der identifizierenden Informationen für Dritte nicht mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft verbunden ist (Baumann et al., 2021).

Die DSGVO unterscheidet nicht zwischen diesen beiden Graden der Anonymisierung, daher ist davon auszugehen, dass immer eine absolute Anonymisierung gemeint ist (Baumann et al., 2021). Da dies kaum zu leisten ist, gilt als Mittelweg für die Datenverarbeitung in der Forschung, soweit möglich, zumindest die direkten Identifizierungsmerkmale unwiederbringlich durch absolute Anonymisierung zu löschen oder gar nicht erst zu erheben und für die indirekten Merkmale den für eine Identifizierung erforderlichen Aufwand zu betrachten (faktsiche Anonymisierung) (Baumann et al., 2021). Bei der Publikation anonymisierter Daten sollte sowohl für die direkten als auch indirekten Identifizierungsmerkmale eine absolute Anonymisierung angestrebt werden.

Pseudonymisierung

Bei einer Pseudonymisierung werden direkte und indirekte Identifizierungsmerkmale durch einen Code ersetzt. Das Codebuch wird dabei vom restlichen Datensatz getrennt und gesondert aufbewahrt.

Solange eine Re-Identifikation der Betroffenen möglich ist, unterliegen die Daten der DSGVO. Da pseudonymisierte Daten durch die Heranziehung des Codebuchs re-identifiziert werden können, unterliegen sie auch der DSGVO. Eine Pseudonymisierung allein reicht also nicht aus, um den datenschutzrechtlichen Vorgaben zu entgehen.

4.8.4 Verzeichnis von Verarbeitungstätigkeiten

Sobald personenbezogene Daten erhoben werden, soll nach Art. 30 der DSGVO ein Verzeichnis von Verarbeitungstätigkeiten (VVT) erstellt werden. Dieses Dokument dient als Nachweis für die Einhaltung des Gesetzes. Es betrifft sämtliche ganz oder teilweise automatisierte sowie nicht-automatisierte Verarbeitungen von personenbezogenen Daten, die in dem Forschungsvorhaben anstehen und gespeichert werden (sollen). Für jede einzelne dieser Verarbeitungen ist eine Beschreibung inkl. der folgenden Aspekte anzufertigen:

- „den Namen und die Kontaktdaten der Verantwortlichen und gegebenenfalls der gemeinsam Verantwortlichen, der Vertreter:innen der Verantwortlichen sowie der etwaigen Datenschutzbeauftragten,
- die Zwecke der Verarbeitung,
- eine Beschreibung der Kategorien betroffener Personen und der Kategorien personenbezogener Daten,
- die Kategorien von Empfänger:innen, gegenüber denen die personenbezogenen Daten offengelegt worden sind oder noch offengelegt werden, einschließlich Empfänger:innen in Drittländern oder internationalen Organisationen,
- gegebenenfalls Übermittlungen von personenbezogenen Daten an ein Drittland oder an eine internationale Organisation, einschließlich der Angabe des betreffenden Drittlands oder der betreffenden internationalen Organisation, sowie bei den in Art. 49 Abs. 1 Unterabs. 2 genannten Datenübermittlungen die Dokumentierung geeigneter Garantien,
- wenn möglich, die vorgesehenen Fristen für die Löschung der verschiedenen Datenkategorien,
- wenn möglich, eine allgemeine Beschreibung der technischen und organisatorischen Maßnahmen gemäß Art. 32 Abs. 1“.

Diese Regelung nach Art. 30 DSGVO verpflichtet nicht nur alle Verantwortlichen, sondern auch die Auftragsverarbeiter:innen, ein VVT zu erstellen und zu führen.

4.8.5 Datengrundsätze

Laut der DSGVO ist es notwendig, sechs Grundsätze zur Verarbeitung personenbezogener Daten bei jedem Verarbeitungsvorgang zu berücksichtigen. Art. 5 DSGVO beschreibt diese Grundsätze:

1. **Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz**
   Die Daten müssen „auf rechtmäßige Weise, nach Treu und Glauben und in einer für die betroffene Person nachvollziehbaren Weise verarbeitet werden“.
2. **Zweckbindung**
   Die Daten müssen „für festgelegte, eindeutige und legitime Zwecke erhoben werden und

dürfen nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden; eine Weiterverarbeitung für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke gilt gemäß Art. 89 Abs. 1 nicht als unvereinbar mit den ursprünglichen Zwecken“.

3. **Datenminimierung**
   Die Daten müssen „dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein“.

4. **Richtigkeit**
   Die Daten müssen „sachlich richtig und erforderlichenfalls auf dem neuesten Stand sein; es sind alle angemessenen Maßnahmen zu treffen, damit personenbezogene Daten, die im Hinblick auf die Zwecke ihrer Verarbeitung unrichtig sind, unverzüglich gelöscht oder berichtigt werden“.

5. **Speicherbegrenzung**
   Die Daten müssen „in einer Form gespeichert werden, die die Identifizierung der betroffenen Personen nur so lange ermöglicht, wie es für die Zwecke, für die sie verarbeitet werden, erforderlich ist; personenbezogene Daten dürfen länger gespeichert werden, soweit die personenbezogenen Daten vorbehaltlich der Durchführung geeigneter technischer und organisatorischer Maßnahmen, die von dieser Verordnung zum Schutz der Rechte und Freiheiten der betroffenen Person gefordert werden, ausschließlich für im öffentlichen Interesse liegende Archivzwecke oder für wissenschaftliche und historische Forschungszwecke oder für statistische Zwecke gemäß Art. 89 Abs. 1 verarbeitet werden“.

6. **Integrität und Vertraulichkeit**
   Die Daten müssen „in einer Weise verarbeitet werden, die eine angemessene Sicherheit der personenbezogenen Daten gewährleistet, einschließlich Schutz vor unbefugter oder unrechtmäßiger Verarbeitung und vor unbeabsichtigtem Verlust, unbeabsichtigter Zerstörung oder unbeabsichtigter Schädigung durch geeignete technische und organisatorische Maßnahmen“.

4.8.6 Literaturempfehlungen

- Baumann, P., Krahn, P. & Lauber-Rönsberg, A. (2021). *Forschungsdatenmanagement und Recht. Datenschutz-, Urheber- und Vertragsrecht.* W. Neugebauer.

- Brettschneider, P., Biernacka, K., Böker, E., Danker, S. A., Jacob, J., Perry, A., Wiljes, C. & Wuttke, U. (2021). Urheberrecht und Lizenzierung bei Forschungsdaten. https://doi.org/10.5281/zenodo.5243232

- Kreutzer, T. & Lahmann, H. (2021). *Rechtsfragen bei Open Science. Ein Leitfaden.* Hamburg University Press. https://doi.org/10.15460/HUP.211

4.8.7 Anwendung auf die Szenarien

Datenschutz bei Szenario 1

In diesem Szenario müssen spezielle Maßnahmen ergriffen werden, da mit personenbezogenen Daten gearbeitet wird. Somit muss zunächst eine informierte Einwilligungserklärung von den Befragten eingeholt werden, bevor Forschungsdaten erhoben werden dürfen. Die aufgenommenen Daten werden anschließend anonymisiert/pseudonymisiert (beispielsweise in Transkripten). Hier gilt der Grundsatz der Datenminimierung. Das heißt, es sollten nur die Daten erhoben werden, die für die Beantwortung der Forschungsfragen zwingend notwendig sind.

Datenschutz bei Szenario 2

In diesem Szenario wird zunächst nicht mit personenbezogenen Daten von existierenden Personen gearbeitet. Somit entfallen scheinbar Maßnahmen im Sinne des Datenschutzes, wie in Szenario 1. Jedoch wird in Szenario 2 Software programmiert, die in ihrer Anwendung sensible und personenbezogene Daten verarbeitet. Aus diesem Grund muss die Software als Forschungsdatum die Daten so verarbeiten, dass sie der DSGVO gerecht wird. Dafür sind beispielsweise folgende Maßnahmen notwendig: eine zweistufige Authentifizierung, Verschlüsselung der Patient:innendaten etc.

Datenschutz bei Szenario 3

In diesem Szenario wird es besonders schwierig. Eine informierte Einwilligung ist hier nicht notwendig, da die Personen ihre Daten freiwillig bei Twitter publizieren und entsprechend Twitters AGB eine Weiterverarbeitung rechtmäßig ist. Die eben genannte Freiwilligkeit muss jedoch unbedingt gegeben sein und die Nutzung des Kurshastags darf nicht verpflichtend sein, um beispielsweise den Kurs erfolgreich abzuschließen. Trotz der Freiwilligkeit müssen bei der Verarbeitung der Daten diese anonymisiert werden.

4.9 Ordnung und Struktur

Eine gute Ordnung und Struktur der eigenen Forschungsdaten führt zu effizienterem Arbeiten im Forschungsvorhaben, da Daten einfacher lokalisiert werden können und somit häufig Doppelarbeit oder Datenverlust, z. B. durch Überschreibung oder Löschung, vermieden werden kann. Bei der Arbeit in Gruppen hilft es auch anderen Gruppenmitgliedern, die Daten zu finden und sie auf Anhieb zuordnen zu können.

4.9.1 Verzeichnisstrukturen

Eine Verzeichnisstruktur (auch Verzeichnisbaum genannt) ist die hierarchische Anordnung, in der Ordner angelegt werden. Sie sollte klar ersichtlich und damit auch für andere Forschende verständlich sein. Je sorgfältiger man sie plant, desto einfacher findet man sich später darin

zurecht. Es ist dabei zu beachten, dass alle Forschende einer eigenen Logik folgen können und daher die Dokumentation der genutzten Verzeichnisstruktur auch hilfreich sein kann. Idealerweise folgen Verzeichnisstrukturen dem Workflow des jeweiligen Forschungsvorhabens.

4.9.2 Namenskonventionen

Eine gute Namenskonvention für Dateien und Ordner ist unverzichtbar, um diese Daten schnell und effizient wiederzufinden. Nicht nur von einem selbst, sondern auch von Projektpartner:innen und Mitarbeitenden. Die Benennung sollte den Kontext der Daten widerspiegeln, dabei intuitiv und objektiv sein. Bei den inhaltsspezifischen bzw. deskriptiven Informationen in Datei- bzw. Ordnernamen sollten diese nach Möglichkeit gängigen Abkürzungen folgen. Ein weiterer wichtiger Punkt ist die Konsistenz der Benennung. Dabei sind festgelegte Namenskonventionen hilfreich, die beispielsweise die Reihenfolge der Datumsangabe oder der benötigten Bestandteile der Benennung vorgeben (z. B. [YYYYMMDD]_[Umfrage_ID]_[Probanden_ID].csv). Sollte eine ID im Dateinamen verwendet werden, sollte die Skalierbarkeit beachtet werden. Beschränkt man die ID auf eine zweistellige Zahl, so kann die Anzahl der Dateien 99 nicht überschreiten. Die Namenskonventionen sollten dokumentiert werden (z. B. in der README-Datei; vgl. Kapitel 4.11) sowie in einem DMP erfasst werden. Insbesondere gewählte Abkürzungen oder Codes sollten dort erläutert werden, um die bestmögliche Nachnutzbarkeit zu ermöglichen.

Die Datei- und Ordnerbenennung sollte so lang wie nötig, aber so kurz wie möglich sein. Das gewährleistet eine bessere Lesbarkeit sowie bessere Darstellung unter allen Betriebssystemen. Auch wenn die meisten Betriebssysteme mittlerweile mit Sonderzeichen zurecht kommen, ist deren Vermeidung empfehlenswert. Zeichen wie { } [ ] ( ) * % # ; ’ , „“ : ? ! & $ § sowie Leerzeichen, Punkte, Umlautbuchstaben, Akzente, Ligaturen, Cédillen etc. sollten vermieden werden, um auch nachhaltig und langfristig die Lesbarkeit der Datei bzw. Ordner zu ermöglichen.

Alle Betriebssysteme können heutzutage Dateien in chronologischer Reihenfolge sortieren. Dennoch ist es empfehlenswert, übersichtshalber das Datum zum Dateinamen hinzuzufügen. Dies kann über die Sortierung hinaus, auch beim Verständnis der Daten behilflich sein. Dabei sollte auf das Format YYYYMMDD oder YYYY-MM-DD geachtet werden.

Viele Geräte generieren kryptische Dateinamen. Diese gilt es umzubenennen, um eine höhere Verständlichkeit für die Forschenden zu erreichen. Automatisch generierte Namen können darüber hinaus zu Konflikten durch Duplikate führen.

Müssen mehrere Dateien umbenannt werden, kann man sich an einer Reihe von Programmen bedienen, die dabei unterstützen können. Für alle Betriebssystem gibt es dafür Angebote (z. B. für Windows Ant Renamer[12] oder Rename-It[13], für MacOS Renamer[14] oder

[12] http://www.antp.be/software/renamer. Archivierte Version: https://perma.cc/VMC7-SLSD. Kostenfrei.
[13] https://sourceforge.net/projects/renameit/. Archivierte Version: https://perma.cc/M5ZX-R97H. Kostenfrei.
[14] https://renamer.com/. Archivierte Version: https://perma.cc/GE83-KW27. Kostenpflichtig.

Name Changer[15], für Linux GPRename[16] oder GNOME Commander[17], unter Unix können die Befehle *rename* oder *mmv* genutzt werden).

4.9.3 Versionierung

Die Versionierung erfasst die Änderung an Dokumenten und Daten und verhindert somit Datenverlust. Die Historie der Daten bleibt auf diese Weise transparent, lückenlos und nachvollziehbar. Bei jeder Änderung an den Daten wird eine „neuere“ Version abgelegt. Dies ermöglicht es auch, problemlos einen Schritt zurückzugehen. Eine Version ist dabei ein eindeutiger Stand dieser Dateien.

Üblicherweise werden neue Versionen mit der Vergabe von ganzen Zahlen gekennzeichnet. Für kleinere Änderungen werden mit Unterstrich verbundene Zahlen gewählt (z. B. v3_1 oder v5_3). Benennungen wie *final*, *revision* oder *letzte* sollten vermieden werden. Für die Softwareentwicklung existieren Vorschläge für einfache Regelwerke, welche definieren, wie Versionsnummern gewählt und erhöht werden sollten. Diese Regeln basieren auf bereits existierenden und weit verbreiteten Verfahren, aber beschränken sich nicht zwingend auf diese (Preston-Werner, 2013).

Bei der Arbeit mit mehreren Personen ist es empfehlenswert, eine Versionskontrolltabelle zu führen. In dieser wird bei der Erstellung einer neuen Version von Daten ein neuer Eintrag hinzugefügt, der beschreibt, wie die neue Version benannt wurde, wer die Veränderung wann durchgeführt hat, und welche Veränderung vorgenommen wurde. Dies hilft insbesondere bei allen Daten, die nicht mit einer Änderungsverfolgung (wie bei Microsoft Word oder Google Docs) nachvollziehbar gemacht werden können.

Bei textbasierten Daten (z. B. Softwarecode, Texte, LaTeX) ist die Nutzung einer automatisierten Verionskontrollsoftware zu empfehlen. Diese legt ihre Daten in einem zentralen Verzeichnis oder einer Datenbank ab, wobei die Versionen mit einem Zeitstempel und einer Nutzerkennung gesichert sind. Alle Versionen bilden einen „Stapel“, dabei bilden neuen Versionen eines Dokumentes den Anfang. Die Versionskontrollsysteme können auch die Zugriffsrechte auf die Daten koordinieren. Durch die Optionen des *Check-In*s und des *Check-Out*s wird die Zusammenarbeit unkompliziert geregelt.

Git

Die mit Abstand gängigste Versionskontrollsoftware ist Git[a] – ein aktiv gepflegtes Open-Source-Projekt. Es registriert Unterschiede zwischen den Dateien und kann daher automatisch und selbständig Veränderungen in verschiedenen Bereichen einarbeiten.

Da es sich bei Git um ein verteiltes System (so genanntes Distributed Version Control System – DVCS) handelt, befindet sich der volle Versionsverlauf nicht an einem Ort (wie z. B. bei Subversion). Jede Arbeitskopie des Bearbeitenden befindet sich gleich-

[15] https://mrrsoftware.com/namechanger/. Archivierte Version: https://perma.cc/5ZSP-WW68. Kostenfrei.

[16] http://gprename.sourceforge.net/. Archivierte Version: https://perma.cc/A43G-GJW3. Kostenfrei.

[17] https://gcmd.github.io/. Archivierte Version: https://perma.cc/R7CT-J6HB. Kostenfrei.

zeitig auf einem Repository, das den vollständigen Verlauf aller Änderungen enthält. Durch die Einbeziehung von zentralen Servern können mehrere Personen gemeinsam an denselben Dateien arbeiten.
Git ermöglicht multiple lokale Branches (Zweige), die völlig unabhängig voneinander sein können (vgl. Abbildung 4.8). Auf diese Weise können neue Ideen ausprobiert werden, ohne das Original zu verändern. Sie können auch unterschiedlichen Funktionen dienen (z. B. Produktiv- und Testlauf) oder der Entwicklung neuer Funktionen. Die Branches können im weiteren Schritt zusammengeführt werden (*merge*). Es können jedoch auch einzelne Zweige auf das Repository übertragen werden.

[a] https://git-scm.com/. Archivierte Version: https://perma.cc/TV4T-QYN7.

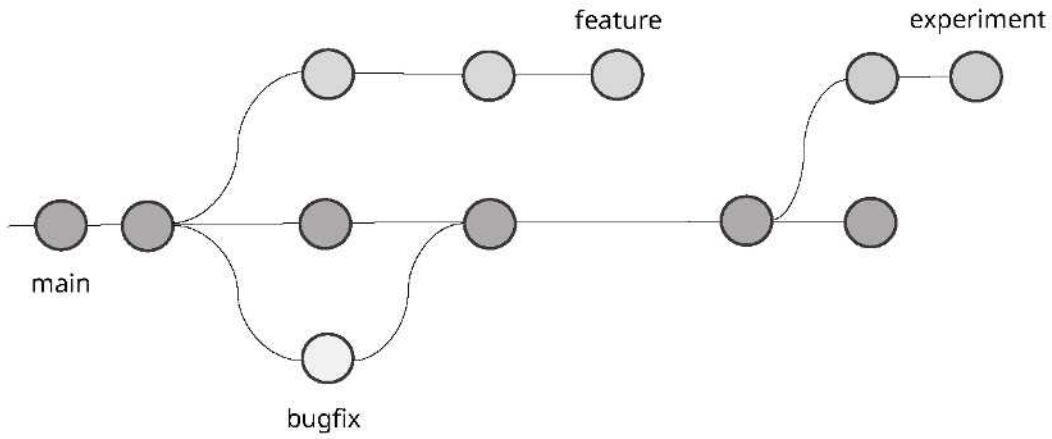


Abbildung 4.8: Beispiel für Branching

Klump et al. (2021) identifizieren sechs Prinzipien der Datenversionierung:

1. Versionskontrolle und Revisionen (Revision)
2. Identifizierung von Versionen eines Datenprodukts (Release)
3. Identifizierung von Datensammlungen (Granularity)
4. Identifizierung von Erscheinungsformen von Datensätzen (Manifestation)
5. Anforderungen an die Provenienz von Datensätzen (Provenance)
6. Anforderungen an die Zitierung von Daten (Citation)

Diese Prinzipien bilden eine Richtlinie zur Datenversionierung.

4.9.4 Literaturempfehlungen

- Haenel, V. & Plenz, J. (2016). Git: verteilte Versionsverwaltung für Code und Dokumente. [Archivierte Version: https://perma.cc/9SQS-EN9Q]. https://github.com/gitbuch
- Pilato, C. M., Collins-Sussman B. & Fitzpatrick, B. W. (2009). *Versionskontrolle mit Subversion.* O'Reilly.

4.9.5 Anwendung auf die Szenarien

Ordnung und Struktur bei Szenario 1

In diesem Szenario sollte insbesondere darauf geachtet werden, dass die Dateibenennung einheitlich und anhand von Regeln vorgenommen wird. Da die Interviewtranskripte auch für andere Forschende interessant sein können, sollte direkt die Möglichkeit der Langzeitarchivierung (vgl. Kapitel 5.14) berücksichtigt werden. Dafür ist eine allgemein verständliche Benennung besonders wichtig, um die Nachvollziehbarkeit zu gewährleisten. Eine Namenskonvention für Interviewdaten könnte sein: `[ID oder Studiennummer]_[Kürzel für Interview]_[laufende Nummer]_[Version]`, wobei

- `ID oder Studiennummer` die Zuordnung zu einer konkreten Studie bezeichnet,
- `Kürzel für Interview` angibt, ob es sich um ein Interview (int) handelt,
- `laufende Nummer` die Durchnummerierung der Daten umfasst,
- `Version` die Versionsnummer angibt, die sich beispielsweise durch Bearbeitungsschritte wie die Anonymisierung verändern kann.

Ein konkretes Beispiel nach dieser Konvention wäre: collab2022_int_01_v1.

Ordnung und Struktur bei Szenario 2

Da Timo in einem Team arbeitet und somit auch weitere Personen Zugriff auf die Forschungsdaten benötigen, ist es wichtig die genutzten Namenskonventionen zu dokumentieren und bei einheitlichen Benennungen zu bleiben. In vielen Git-Systemen lassen sich die Berechtigungen so anpassen, dass die Konventionen automatisch eingehalten werden müssen. Eine Namenskonvention bei der Nutzung von Multi-Repo-Variante könnte sein: `<projektname>-<app>-<component/module>`, wobei

- `projektname` der Prefix für alle Repository, die zu diesem Projekt gehören, ist,
- `app` der Kurzname der Applikation ist,
- `component/module` der eigentliche Name der Komponente bzw. des Moduls darstellt.

Die Versionierung der Daten wird innerhalb von Git durchgeführt. Falls es verschiedene Versionen geben sollte, die in weiteren Tools abgelegt werden, wird dies entspre-

chend gekennzeichnet. Ein konkretes Beispiel dafür könnte sein: DigPubHealth4all-Bills-DataTransfer.

Ordnung und Struktur bei Szenario 3

In diesem Szenario wird eine große Anzahl an Daten gesammelt, wobei darauf zu achten ist, dass die gesammelten Datensätze nach einheitlichen Konventionen benannt werden. Eine Grundstruktur für die Benennung quantitativer Daten kann wie folgt aussehen: `[ID oder Studiennummer]_[Kennzeichner für Datensatz]_[Version]`, wobei

- `ID oder Studiennummer` die Zuordnung zu einer konkreten Studie bezeichnet,
- `Kürzel für Datensatz` angibt ob es sich um ein Datensatz (ds) handelt,
- `Version` die Versionsnummer angibt.

Für die Verarbeitung und Analyse der Daten, wird die Benennungskonvention einen Datum- und Zeitstempel des jeweiligen Tweets enthalten. Zusätzlich wird ein Datei-Präfix eingefügt mit dem Pseudonym der tweetenden Person. Dementsprechend kann eine Namenskonvention für Alex wie folgt zusammengesetzt sein: `[ID oder Studiennummer]_[Kennzeichner für Datensatz]_[Datum]_[Zeit]_[Name der tweetenden Person]_[Version]`. Ein konkretes Beispiel in Alex' Datenbenennung kann wie folgt aussehen: LA2ICST_ds0001_2022-04-01_12-01_Paula125_v1.

Bei der Publikation der Forschungsdaten werden die Dateinamen anonymisiert sowie mit dem Namen des Projekts und einer zufallsgenerierten Nummer ersetzt.

4.10 Speicherung und Back-up

Die Auswahl von geeigneten Speichermedien für Forschungsdaten ist ein wichtiger Faktor für die Aufbewahrung und Auffindbarkeit der Daten. Eine bewusste und strategisch sinnvolle Wahl der Medien kann auch dabei helfen, Datenverlust oder Doppelarbeit zu vermeiden.

4.10.1 Speichermedien

Die unterschiedlichen Medien weisen verschiedene Stärken und Schwächen auf, die je nach Anwendungsfall zur Geltung kommen. Komplexe Daten aus Experimenten, in denen mehrere 100 Megabytes an Daten pro Sekunde entstehen, müssen anders abgelegt werden, als die Fragebögen aus sozialwissenschaftlichen Umfragen. Tabelle 4.2 gibt einen Überblick über die Vor- und Nachteile der unterschiedlichen Speichermedien.

Tabelle 4.2: Vergleich von Speichermedien

Speichermedium	**Vorteile**	**Nachteile**
Eigener PC	• Kontrolle über den Zugriff auf Daten • Eigenverantwortung bzgl. der Sicherheit der Daten und des Back-ups	• Know-how zur Erstellung von Back-ups und Wartung des Systems notwendig • Einzellösungen sind kostenaufwendig
Mobile Speichermedien	• leichter Transport • einfache Aufbewahrungsmöglichkeiten	• unsicher gegen Verlust und/oder Diebstahl • stoß- und verschleißanfällig
Institutionelle Speicherorte	• Back-up der Daten ist sichergestellt • professionelle Wartung des Systems • Speicherung entsprechend der Datenschutzrichtlinien der Institution • für mobiles Arbeiten nutzbar	• Zugriff auf Back-ups evtl. verzögert durch längeren Dienstweg • für den Zugriff ist eine Netzwerkverbindung notwendig

Externe Speicherorte		
	• weit verbreitet und intuitiv in der Nutzung • professionelle Wartung • für mobiles Arbeiten nutzbar	• häufig kostenpflichtig • die Verbindung kann unsicher sein • der Datenschutz abhängig vom Land, in dem der Server steht • von manchen Institutionen geblockt

Ob auf DVDs, Festplatten, USB-Sticks oder auf Blu-rays – die Forschungsdaten werden zu verschiedenen Zwecken auf diesen Medien gespeichert. Neben den genannten Vor- und Nachteilen dieser Speichermedien, ist auch die Lebensdauer bei der Auswahl von Relevanz. Diese kann stark variieren und somit die Lesbarkeit der Daten einschränken. Anders als die in Stein gemeißelten Informationen, können Daten auf modernen Datenträgern schnell verloren gehen. Tabelle 4.3, basierend auf Schasche (2018), stellt die geschätzte Lebensdauer sowie die größte Bedrohung für die jeweiligen Speichermedien dar.

4.10.2 Back-up

Um Forschungsdaten zu schützen, ist die Auswahl des Speicherortes von großer Bedeutung. Darüber hinaus ist es jedoch auch wichtig, die Daten in regelmäßigen Abständen zu sichern. Die Erstellung einer Sicherheitskopie auf weiteren Speichermedien wird als Back-up bezeichnet. Dieses sollte geplant und systematisch durchgeführt werden, damit man im Bedarfsfall die Daten wiederherstellen kann.

Betriebssysteme sind heutzutage bereits mit Back-up-Programmen ausgestattet, was den Installationsaufwand reduziert. Ist man mit dieser Standardsoftware unzufrieden, gibt es auf dem Markt eine große Auswahl an Back-up-Software. Dabei können sie sich bei der Sicherungsart unterscheiden: vollständig, inkrementell oder differenziell. In allen Fällen wird im ersten Schritt ein Voll-Back-up von allen existierenden Dateien erstellt. Im zweiten Schritt werden bei der inkrementellen Datensicherung nur die Dateien oder Teile von Dateien gespeichert, die sich seit der letzten inkrementellen Sicherung geändert haben oder neu hinzugekommen sind. Bei der differenziellen Sicherung werden hingegen alle Daten gespeichert, die sich seit dem letzten vollständigen Back-up geändert haben oder neu hinzugekommen sind.

Tabelle 4.3: Lebensdauer von Speichermedien

Speichermedium	**Lebensdauer in Jahren**	**Bedrohung**
Optische Medien:		
DVD	≤ 30	Wärme, Licht, Feuchtigkeit und Kratzer
Blu-ray	50–100	
CD	<80	
Gepresste optische Medien:		
DVD	100	Temperaturen über 25 Grad Celsius und Luftfeuchtigkeit von über 80 Prozent
Blu-ray	80	
Externe Festplatten	10	Feuchtigkeit, Stöße, Magnetismus
Interne Festplatten	5–10	Wärme im Betrieb
SSD		
USB-Stick	≥ 5	Begrenzte Schreibzyklen
Cloud-Speicher	theoretisch unbegrenzt	Zugriff durch Dritte, Pleite des Anbieters

Eine empfehlenswerte Back-up-Strategie stellt die 3-2-1-Regel dar:

- **drei** Kopien einer Datei,
- auf mindestens **zwei** verschiedenen Speichermedien,
- wovon **eine** dezentral abgelegt sein sollte.

Die Erstellung der Sicherheitskopien sollte regelmäßig geschehen und auch in regelmäßigen Abständen auf Lesbarkeit und Vollständigkeit überprüft werden.

Viele Universitäten bieten automatisierte Back-up-Lösungen bei der Nutzung von institutionellen Speichermedien an. An der Humboldt-Universität zu Berlin und der Universität Hamburg wird beispielhaft das IBM Spectrum Protect™ genutzt (vorher Tivoli Storage Manager (TSM) genannt, was weiterhin als Bezeichnung in den Dokumentationen zur Software seitens IBM genutzt wird). Dank der professionellen Konfiguration und Wartung sind die Daten hier gut aufgehoben.

4.10.3 Literaturempfehlungen

- Hanson, K., Surkis, A. & Yacobucci, K. (2013). Data Sharing and Management Snafu in 3 Short Acts [Zuletzt geprüft 2022-03-01]. https://youtu.be/66oNv_DJuPc

4.10.4 Anwendung auf die Szenarien

Speicher und Back-up bei Szenario 1

In dem Fall von Carla reicht ein lokales Back-up und eine zusätzliche Sicherheitskopie auf externem Medium vermutlich aus. Da Carla ihre Interviews an verschiedenen Orten führen wird und dafür auf Reisen ist, sollte zusätzlich am besten ein mobiler Speicher oder institutioneller Speicherort (sofern für Studierende nutzbar) genutzt werden.

Speicher und Back-up bei Szenario 2

Sofern in dem Open-Source-Projekt weitere Tools genutzt werden, die kein automatisches Back-up erstellen, muss eine regelmäßige Speicherung bei diesem Szenario erfolgen. Welche Formen der Speicherung genutzt werden sollten, ist abhängig von der Projektgröße und Form der Zusammenarbeit. Insbesondere bei diesem kollaborativen Projekt ist die Speicherung auf institutionellen Speicherorten mit automatischem Back-up ratsam.

Speicher und Back-up bei Szenario 3

Bei Drittmittelprojekten ist ein institutioneller gesicherter Speicherort ratsam, um eine gemeinsame Arbeit unter den Projektbeteiligten zu ermöglichen und einen Datenverlust zu vermeiden. Es ist je nach Mittelgeber zu prüfen, welche individuellen Vorgaben zur Speicherung und zum Back-up existieren. Oftmals dürfen die Daten nicht auf privaten Geräten gespeichert werden (vgl. Kapitel 4.8).

4.11 Dokumentation und Metadaten

Die reine Verfügbarmachung der Daten ermöglicht keine Nachnutzung dieser Daten. Eine gute Datendokumentation ist für eine transparente und reproduzierbare Forschung unabdingbar. Sie ermöglicht den Nachnutzenden nicht nur das Verstehen der Daten, sondern auch vor allem das Finden dieser Daten.

Ein großer Vorteil einer ausführlichen Dokumentation ist, dass diese selbst in einem Data Paper (vgl. Kapitel 4.13.2) publiziert werden kann und somit direkt zur Publikationsliste der Forschenden beiträgt. Darüber hinaus ist es bereits nachgewiesen, dass eine gute Datendokumentation zur Steigerung der Zitation der dazugehörigen Artikel führt (Piwowar & Vision, 2013).

Die Vorteile beschränken sich jedoch nicht nur für die Nachnutzenden. Auch für einen selbst ist es hilfreich, eine Beschreibung der Daten und der vorgenommenen Verarbeitungsschritte zu haben, da viele Details mit der Zeit in Vergessenheit geraten. Bereits nach wenigen Jahren wird man vieles nicht mehr wiedergeben können.

4.11.1 Inhalte einer Dokumentation

Zu den grundlegenden Inhalten einer Dokumentation gehören:

- Beschreibung des Forschungsvorhabens
- Projektziele
- Hypothesen
- Informationen zur Erhebung der Daten (Methoden, Einheiten, Zeiträume, Orte, verwendete Instrumente und Software)
- Maßnahmen zur Datenbereinigung
- Struktur der Daten und deren Beziehungen zueinander
- Erläuterung von Variablen, Labels und Codes
- Unterschiede zwischen verschiedenen Versionen
- Informationen zum Zugang und Nutzungsbedingungen

Je detailreicher die Datendokumentation ist, desto einfacher können diese Daten zum späteren Zeitpunkt verstanden werden.

Ein Teil der Inhalte einer Dokumentation überschneidet sich mit Elementen eines DMP (vgl. Kapitel 4.6). Insbesondere die Beschreibung des Forschungsvorhabens, die Projektziele und Hypothesen können in beiden Dokumenten gleich sein. Der Unterschied zwischen diesen beiden Dokumenten liegt darin, dass der DMP vor Beginn eines Forschungsvorhabens geschrieben wird und dort die Daten als Ganzes sowie die Motivation und die geplanten Verarbeitungsschritte beschrieben werden. Eine Dokumentation wird kontinuierlich geschrieben – immer wenn ein Schritt vorgenommen wurde. Auch wenn man einen DMP in regelmäßigen Abständen aktualisieren kann und sollte, ist eine Dokumentation dennoch viel detaillierter und konkreter.

4.11.2 Formen einer Dokumentation

Es gibt verschiedene Möglichkeiten, Daten zu dokumentieren. Zu den häufigsten Formen gehören:

- ReadMe-Datei (alternative Schreibweise: README bzw. readme) ist eine reine Textdatei, wodurch sie auf allen Betriebssystemen gelesen werden kann. Am häufigsten wird sie im .txt-Format gespeichert, es gibt jedoch auch die Varianten .doc, read.me oder readme.1st. In der Ordnerstruktur der Forschungsdaten sollte sich die ReadMe-Datei ganz oben befinden.
- Data Dictionary (Datenkatalog) führt alle zum Datensatz dazugehörigen Dateien inklusive deren Metadaten und Beziehungen zueinander auf (vgl. Abbildung 4.9). Bei relationalen Datenbanken bezieht sich der Begriff auf eine Menge von Tabellen mit

Research Data Management - The Data Dictionary

File name	Data Type	Method	Creator	Date	Description	Rights	Long-term availability	

Abbildung 4.9: Beispiel für ein Data Dictionary nach Biernacka et al. (2020)

Ansichten, die sich bei der Abfrage im Read-only-Modus befinden. Diese Tabellen enthalten nur Metadaten.

- Codebook (Codebuch) ist eine Dokumentationsform der genutzten Variablen, Labels und Codes (vgl. Abbildung 4.10). Idealerweise stellt eine Zeile eine Variable und jede Information zu einer Variable eine Spalte dar. Für jede Variable sollten die folgenden Informationen gegeben sein: Name, Beschreibung der Variable, Maßeinheiten, Kodierung der Werte (z. B. 1 = weiblich, 2 = männlich), mögliche Wertebereiche, Bezeichnung für fehlende Werte und Beziehungen zu anderen Variablen. Darüber hinaus sollte die Quelle einer Maßnahme und Informationen zur genutzten Skala vorhanden sein.

  Je nach Anwendungsfall und teilweise je nach Disziplin, werden die beiden Dokumentationsformen – Data Dictionary und Codebook – als Synonyme genutzt.

Show rows with cells including:

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

Abbildung 4.10: Beispiel für ein Codebook nach OSF Support (2022)

- Electronic Lab Notebooks (ELN) bzw. Elektronische Laborbücher (ELB) ersetzen die analogen Laborbücher und unterstützen die Dokumentation im Laboralltag. Dabei können nicht nur die Rohdaten dokumentiert werden, sondern auch die Konzeption,

Durchführung und Auswertung von wissenschaftlichen Experimenten, Beobachtungen oder Versuchen und den in diesem Zusammenhang erstellten Forschungsdaten. Auf diese Weise können die Rohdaten unmittelbar mit Protokollen, Prozessen und Workflows verlinkt werden. Einen guten Überblick über ELNs bieten Adam und Lindstädt (2019).

- Software-Dokumentation erläutert je nach Perspektive sowohl den Entwickler:innen, Anwender:innen als auch Endnutzer:innen, wie eine Software funktioniert, was für ihren Betrieb notwendig ist, wie sie entwickelt wurde und welche Daten von ihr erzeugt und verarbeitet werden. Es gibt verschiedene Wege, um Software zu dokumentieren:
  - Programmierdokumentation (auch: Inline Source Documentation) obwohl der Quellcode selbsterklärend sein sollte, wird die notwendige Dokumentation möglichst weit direkt in den Quellcode eingearbeitet. Dies geschieht durch Kommentare und Kommentarzeilen in unmittelbarer Nähe der betroffenen Stelle.
  - Methodendokumentation behandelt weniger die Programmierung selbst als vielmehr die Erläuterung der angewendeten Methoden, Algorithmen, Flussdiagramme sowie die Hintergründe der Softwareentwicklung. Es ist eine Art detailliertes Pflichtenheft.
  - Schnittstellendokumentation beschreibt die vorhandenen und genutzten Schnittstellen wie Bussysteme, Laufzeitumgebungen, GUI, Webdienste oder API. Auf diese Weise erhält man eine gute Übersicht über die externen Möglichkeiten der Software sowie der verwendeten Technologien.
  - Technische Dokumentation bietet Hintergrundinformationen zu den verwendeten Technologien. Sie kann auch als ein ausführliches Handbuch angesehen werden. Zur Erstellung sollten die ISO-Normen in Betracht gezogen, sowie die Produkthaftung geklärt werden. Zur Zielgruppe einer technischen Dokumentation gehören in erster Linie technische Mitarbeitende.
  - Benutzerdokumentation (auch: Handbuch, Manual bzw. Benutzerhandbuch) dient der Erklärung des Programms für die Nutzer:innen. Sie wird in einer benutzerfreundlichen Sprache geschrieben und sollte auch für weniger technisch versierte Personen verständlich sein. Sie beinhaltet Informationen zur Funktionalität der Software, ihrer Eingabedaten und der erzeugten Ergebnisse, eine Bedienungsanleitung, mögliche Lösungen für Problemfälle, häufig gestellte Fragen (FAQ) und ein Glossar.

Es gibt auf dem Markt auch immer mehr kommerzielle Software, die bei der Dokumentation von Daten unterstützen soll. So ermöglicht zum Beispiel *colectica*[18] die Dokumentation von Variablen, Codelisten und Datensätzen direkt aus Microsoft Excel heraus.

[18] https://www.colectica.com/software/colecticaforexcel/. Archivierte Version: https://perma.cc/6TWF-V2ZR.

4.11.3 Metadaten

Eine Dokumentation dient vor allem der Verständlichkeit und der Reproduzierbarkeit von Daten. In Textform mit Beispielen ist sie sehr gut menschenlesbar. Um die Daten jedoch auch maschinenlesbar zu gestalten und sie somit über Suchmaschinen auffindbar zu machen, ist die Vergabe von Metadaten notwendig.

Metadaten

Metadaten sind strukturierte Daten über Daten.

Metadaten werden nach deren Funktion kategorisiert. Auf diese Weise kann man unter anderem zwischen den folgenden Kategorien unterscheiden:

- administrative Metadaten (z. B. Erstellungsdatum, Dateigröße oder -typ),
- bibliografische Metadaten (z. B. Autor:in, Titel oder Abstract),
- beschreibende Metadaten (geben zusätzliche Informationen zu Inhalt und Erstellung der Daten),
- technische Metadaten (z. B. Blende oder Geo-Location),
- legale Metadaten (z. B. Informationen zur Lizenzierung oder dem Urheberrecht),
- statistische Metadaten (z. B. Statistiken aus Berichten oder Umfragen),
- Prozessmetadaten (beschreiben die einzelnen Phasen und Maßnahmen, die zur Erstellung und Bearbeitung der Daten verwendet wurden),
- Provenienzmetadaten (Herkunftsinformationen der Daten),
- strukturelle Metadaten (weisen die Zusammenhänge zwischen verschiedenen Daten auf),
- Nutzungsmetadaten (werden bei jedem Zugriff gesammelt).

Metadaten können sowohl manuell als auch automatisiert erstellt werden. Die technischen Metadaten werden häufig direkt von den genutzten Geräten aufgezeichnet und in die Datei eingebettet, während z. B. beschreibende Metadaten von den Ersteller:innen der Daten hinzugefügt werden.

Um Interoperabilität zu gewährleisten, werden Metadaten häufig im XML-Format gespeichert (vgl. Abbildung 4.11). Dies ermöglicht auch einen besseren Informationsaustausch zwischen unterschiedlichen Systemen. Aus Sicht der Forschenden sind Metadaten Attribute, die sie ihren Daten – meist über ein Formular über eine Weboberfläche – vergeben.

Metadatenschema

Ein Metadatenschema ist eine Zusammenstellung von Attributen (Metadaten) zur Beschreibung von Daten.

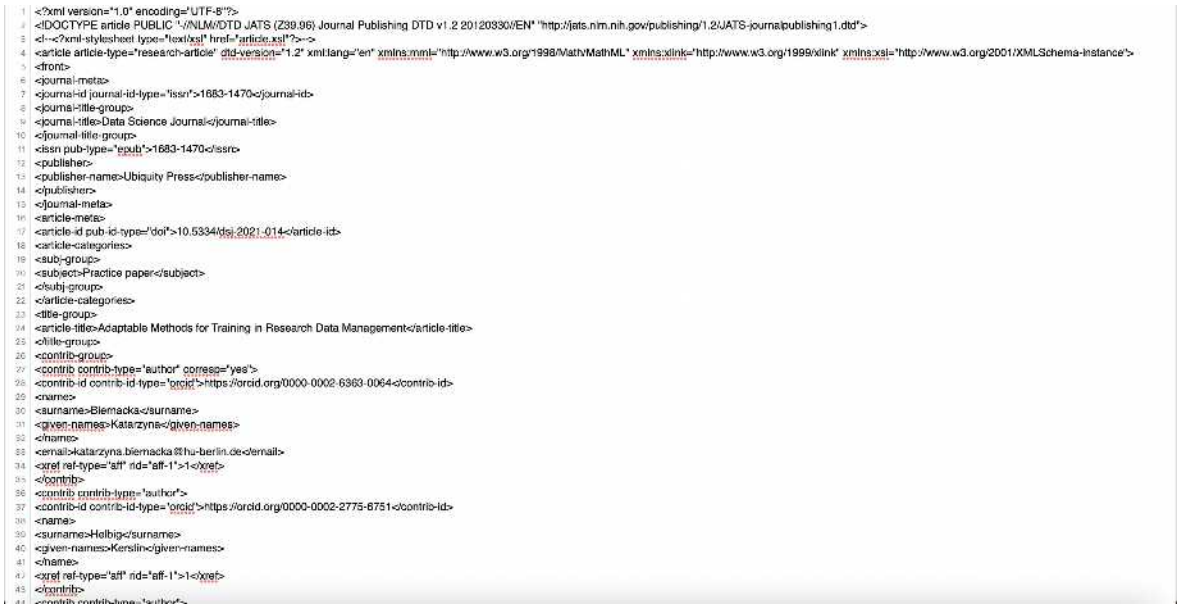


Abbildung 4.11: Ausschnitt von Metadaten im XML-Format am Beispiel von Biernacka et al. (2021b)

Auch für die Beschreibung von Software ist ein umfassendes Metadatenschema notwendig. Hierfür eignet sich CodeMeta[19] als umfassendes Beschreibungsschema für Forschungssoftware und -code im JSON-Format besonders gut (Jones et al., 2017). Der entstehende JSON-Code kann entweder in Form einer codemeta.json–Datei im Softwarerepositorium hinterlegt oder auf Dokumentations- oder Beschreibungsseiten eingebettet werden (TU9-FDM, 2019). Dabei beruht CodeMeta auf schema.org (vgl. Tabelle 4.4) und sieht wie folgt aus (Beispiel von Jones et al., 2017[20]): {

```
"@context": "http://schema.org",
"@type": "Code",
"author": [
  {
    "@id": vhttp://orcid.org/0000-0002-3957-2474",
    "@type": "Person",
    "email": "arfon@github.com",
    "name": "Arfon Smith"
  },
  {
    "@id": "http://orcid.org/0000-0002-7217-4494",
    "@type": "Person",
```

[19] https://doi.org/10.5063/schema/codemeta-2.0.

[20] https://github.com/codemeta/codemeta/blob/master/examples/example-code-jsonld.json. Archivierte Version: https://perma.cc/3VEB-QRU4.

```
            "email": "kaitlin@mozillafoundation.org",
          "name": "Kaitlin Thaney"
        }
    ],
    "citation": "http://dx.doi.org/10.6084/m9.figshare.828487",
    "codeRepository": "https://github.com/arfon/fidgit",
    "dateCreated": "2013-10-19",
    "description": "An ungodly union of GitHub and Figshare
                    http://fidgit.arfon.org",
    "keywords": "publishing, DOI, credit for code",
    "license": "http://opensource.org/licenses/MIT",
    "name": "Fidgit"
}
```

4.11.4 Standardisierung von Metadaten

Um mehr Nutzen aus Metadaten zu ziehen, die Daten besser auffindbar zu machen sowie die Interoperabilität zu gewährleisten, ist eine Standardisierung der Metadaten notwendig. Dies garantiert die Verknüpfung der Metadaten. Standards ermöglichen darüber hinaus eine inhaltlich und strukturell gleichförmige Beschreibung von ähnlichen Datensätzen. Jeder Standard basiert auf einem spezifischen Schema, das eine übergreifende Struktur für alle seine Metadaten bietet (Kranz, 2021). Beispiele von den häufigsten generischen Metadatenstandards können Tabelle 4.4 entnommen werden.

Metadatenstandards werden auch disziplinspezifisch definiert, um die Interoperabilität innerhalb einer Community zu vereinfachen. Eine Auflistung dieser Standards führt die Research Data Alliance (2021). Einige wichtige Metadatenstandards in der Informatik sind in Tabelle 4.5 aufgelistet.

Tabelle 4.4: Die häufigsten generischen Metadatenstandards

Standard	**Beschreibung**
Dublin Core[21]	Dublin Core ist ein allgemeiner Metadatenstandard und beschreibt die Attribute von 15 Kernelementen. Dublin Core ist bei webbasierten, digitalen Metadaten weit verbreitet.
schema.org[22]	schema.org bietet eine Sammlung von Metadatenschemata für strukturierte Daten im Internet, auf Webseiten, in E-Mails und darüber hinaus an.

[21] https://dublincore.org/. Archivierte Version: https://perma.cc/RQA2-3WGH.
[22] https://schema.org/. Archivierte Version: https://perma.cc/AJG2-BCAM.

Metadata Object Description Schema (MODS)[23]	Das Metadata Object Description Schema ist ein bibliografischer XML-basierter Metadatenstandard für Bibliotheken.
DataCite Metadata Schema[24]	Das DataCite Metadata Schema ist eine Liste der Kernmetadaten, die für eine genaue und konsistente Identifizierung einer Ressource zu Zitier- und Auffindungszwecken notwendig sind, zusammen mit empfohlenen Nutzungsanweisungen.
MARC 21[25]	Machine-Readable Cataloging (MARC) ist ein Standard- und Serialisierungsformat für die Darstellung bibliografischer Metadaten, das ursprünglich für den Austausch bibliografischer Datensätze zwischen Bibliothekskatalogen entwickelt wurde. MARC 21 ist die weltweit am häufigsten verwendete Version.
Metadata Encoding & Transmission Standard (METS)[26]	METS ist ein Metadatenstandard für die Kodierung von beschreibenden, administrativen und strukturellen Metadaten zu Objekten innerhalb einer digitalen Bibliothek.

4.11.5 Kontrolliertes Vokabular

Das XML-Format von Metadatenstandards und -schemata gibt die Struktur der Attribute vor. Diese sagen jedoch noch nichts über die Inhalte der Metadaten aus. Um auch diese zu standardisieren, ist kontrolliertes Vokabular in Form von Taxonomien, Glossaren, Thesauri, Ontologien oder Normdaten notwendig. Diese Sammlungen von eindeutigen Begriffen ermöglichen eine gemeinsame, einheitliche Sprache zur Beschreibung von Daten und sollten nach Summann (2015) folgende Qualitätskriterien erfüllen:

- Begriffe sind eindeutig identifiziert und definiert,
- Synonyme sind möglichst vollständig erfasst,
- Homonyme sind eindeutig geklärt.

[23] http://www.loc.gov/standards/mods/. Archivierte Version: https://perma.cc/JQ5A-G2AS.
[24] https://schema.datacite.org/. Archivierte Version: https://perma.cc/6BFZ-NCBX.
[25] https://www.dnb.de/DE/Professionell/Metadatendienste/Exportformate/MARC21/marc21_node.html#doc158972bodyText1. Archivierte Version: https://perma.cc/6G2W-Y6UM.
[26] http://www.loc.gov/standards/mets/. Archivierte Version: https://perma.cc/S27P-BDEM.

Tabelle 4.5: Beispiele für Metadatenstandards in der Informatik

Standard	**Beschreibung**
CodeMeta[27]	CodeMeta ist ein minimales Metadatenschema für wissenschaftliche Software und Code, in JSON und XML. Das Ziel von CodeMeta ist es, ein Konzeptvokabular zu erstellen, das zur Standardisierung des Austauschs von Software-Metadaten zwischen Repositorien und Organisationen verwendet werden kann.
Learning Object Metadata (LOM)(IEEE, 2020)	LOM ist ein offener Metadatenstandard zur Beschreibung von Lernobjekten.
NISO Metadata for Images in XML (MIX)[28]	MIX bietet Zusatzinformationen im XML-Format für Bilddateien und digitale Bildsammlungen.
Standard Generalized Markup Language (SGML)	SGML dient der inhaltlichen Markierung elektronisch erstellter Texte.

Ein Thesaurus ist eine strukturierte Sammlung von Begriffen, in der Wörter zusammen mit ihrem semantischen Kontext verwaltet werden. Ein bekanntest Beispiel ist der Getty Thesaurus of Geographic Names (TGN).[29]

Ontologien bilden eine Art Beziehungsnetz zwischen den Daten (vgl. Abbildung 4.12). Die bekannteste Definition stammt von Gruber (2016), der Ontologien als explizite formale Spezifikation einer Konzeptualisierung bezeichnet. In der Informatik unterscheiden Gruninger und Lee (2002) drei Anwendungsfelder von Ontologien:

1. Kommunikation,
2. automatisches Schließen,
3. Repräsentation sowie Wiederverwendung von Wissen.

Somit sind Ontologien unter anderem in den Bereichen der Künstlichen Intelligenz, Datenbanken, Informationssysteme oder Multimedia-Kommunikation von Bedeutung. Eine ausführliche Erläuterung von Ontologien in der Informatik bieten Gruninger und Lee (2002).

Klassifizierung kann als eine systematische Einteilung in Gruppen oder Kategorien nach festgelegten Kriterien verstanden werden (vgl. Abbildung 4.12). Anwendung finden Klassifikationen unter anderem in Form von Taxonomien.

Taxonomien sind strukturierte Vokabulare, in denen Begriffe miteinander in Beziehung gesetzt sind. Sie ermöglichen es, Objekte nach bestimmten Kriterien (meist Klassen oder Hierarchien) zu klassifizieren. Nach Lambe (2006)

[29] http://www.getty.edu/research/tools/vocabularies/tgn/index.html. Archivierte Version: https://perma.cc/P9ES-ZD8W

- ist eine Taxonomie eine Form eines Klassifikationsschemas,
- sind Taxonomien semantisch,
- ist eine Taxonomie eine Art Wissenslandkarte[30].

Während Ontologien Inhalte und ihre Beziehungen beschreiben, formalisiert eine Taxonomie die hierarchischen Beziehungen zwischen Konzepten und spezifiziert den Begriff, der für jedes Konzept zu verwenden ist (vgl. Abbildung 4.12).

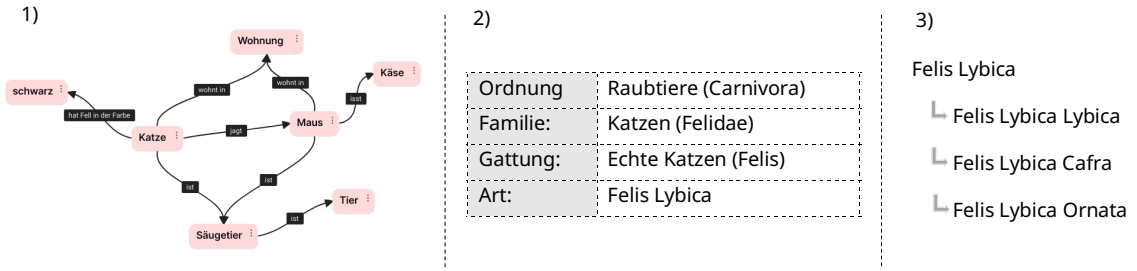


Abbildung 4.12: Vergleich von 1) Ontologien, 2) Taxonomien und 3) Klassifikationen

Normdaten unterstützen eine eindeutige Zuweisung von Personen, Orten oder Institutionen. Zu den bekanntesten gehören:

- Gemeinsame Normdatei (GND)[31]
- International Standard Name Identifier (ISNI)[32]
- Virtual International Authority File (VIAF)[33]

Insbesondere in der Informatik werden Identifikatoren zur eindeutigen Identifizierung von Objekten benutzt.

Um eine möglichst breite Nachnutzbarkeit der Metadaten zu gewährleisten, sollen Vokabeln verwendet werden, die im Sinne von Linked Open Data in maschinenlesbarer Form frei zur Verfügung stehen. Darüber hinaus wird ausdrücklich empfohlen, Vokabeln zu verwenden, die innerhalb einer Community und bestenfalls darüber hinaus anerkannt und verbreitet sind. Sie sollten standardkonform sein und von einer Institution veröffentlicht und gepflegt werden.

Das Basic Register of Thesauri, Ontologies & Classifications (BARTOC)[34] bietet eine Suchmöglichkeit und hat viele anerkannte Thesauri und Klassifikationen gelistet. Die Suche ist in 20 europäischen Sprachen verfügbar und bietet zwei Suchoptionen: die einfache Suche nach Stichwörtern und die erweiterte Suche nach Taxonomiebegriffen. Somit kann disziplinspezifisches kontrolliertes Vokabular einfach gefunden werden.

[30] auch *Knowledge Map* genannt: eine grafische Darstellung von Wissen.
[31] https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html. Archivierte Version: https://perma.cc/E6HU-URFP.
[32] https://isni.org/. Archivierte Version: https://perma.cc/J6VG-UVJ9.
[33] http://viaf.org/. Archivierte Version: https://perma.cc/652D-2J2N.
[34] http://bartoc.org/. Archivierte Version: https://perma.cc/78SL-H3XE.

4.11.6 Vorgehen bei Dokumentation

Nach CESSDA ERIC (2020)[35] gibt es die folgenden sechs Schritte der Datendokumentation:

1. Keine Panik. Viele Dokumentationen sind einfach gute Forschungspraktiken, also machen Sie wahrscheinlich schon viel davon.

2. Fangen Sie früh an! Eine sorgfältige Planung Ihrer Dokumentation zu Beginn Ihres Projektes hilft Ihnen, Zeit und Aufwand zu sparen. Warten Sie nicht mit der Dokumentation bis zum Ende des Projekts. Denken Sie daran, Schritte zur Dokumentation in Ihren DMP aufzunehmen.

3. Denken Sie über die Informationen nach, die benötigt werden, um die Daten zu verstehen. Was werden andere Forschende und Nachnutzende benötigen, um Ihre Daten zu verstehen?

4. Erstellen Sie eine separate Dokumentationsdatei für die Daten, die die grundlegenden Informationen zu den Daten enthält. Sie können auch ähnliche Dateien für jeden Datensatz erstellen. Denken Sie daran, Ihre Dateien so zu organisieren, dass eine Verbindung zwischen der Dokumentationsdatei und den Datensätzen besteht.

5. Planen Sie, wo die Daten nach Abschluss des Projekts abgelegt werden sollen. Das Repositorium folgt wahrscheinlich einem bestimmten Metadatenstandard, den Sie übernehmen sollten.

6. Dokumentieren Sie kontinuierlich während des gesamten Projekts. Die Datendokumentation liefert kontextuelle Informationen über Ihre Datensätze. Es legt die Ziele des ursprünglichen Projekts fest und enthält erläuterndes Material, einschließlich der Datenquelle, der Methodik und des Prozesses der Datenerhebung, der Datensatzstruktur und der technischen Informationen. Umfangreiche und strukturierte Informationen helfen Ihnen, einen Datensatz zu identifizieren und Entscheidungen über seinen Inhalt und seine Benutzerfreundlichkeit zu treffen.

Um eine größere Verbreitung der Daten und somit auch der Dokumentation zu gewährleisten, ist es empfehlenswert, diese (zusätzlich) auf Englisch zu verfassen.

4.11.7 Literaturempfehlungen

- Social Science Research Council. (o. J). *Principles of Documenting Data.* In: Managing Qualitative Social Science Data. An interactive online course. [Archivierte Version: https://perma.cc/WJL2-UUA4]. https://managing-qualitative-data.org/modules/2/a/

- The Carpentries. (o. J). 4 Simple recommendations for Open Source Software. Make software easy to discover by providing software metadata via a popular community registry. [Archivierte Version: https://perma.cc/N7PY-ZQXX]. https://softdev4research.github.io/4OSS-lesson/05-use-registry/index.html

[35] Übersetzung von Biernacka et al. (2021a).

4.11.8 Anwendung auf die Szenarien

Dokumentation und Metadaten bei Szenario 1

Da Carla eine Interviewstudie durchführt, wird sie mit Fragen-Codes arbeiten. Diese gilt in der Dokumentation zu erläutern. Es empfiehlt sich daher in ihrem Fall, auf mehrere Dokumentationsformen zuzugreifen: ReadMe.txt, Codebook und ein Data Dictionary. Zur Vergabe von Metadaten kann sich Carla an den allgemeinen Metadatenstandard Dublin Core oder an den Standard der Sozialwissenschaften DDI halten.

Dokumentation und Metadaten bei Szenario 2

Timo schreibt in seinem Projekt eine Software. Hier sollte darauf geachtet werden, die Inline Source Documentation durchzuführen. Darüber hinaus sollte eine ReadMe-Datei erstellt werden, sowie eine Benutzerdokumentation zur Erläuterung des Programms für die Nutzer:innen. Zur umfangreichen Beschreibung der Software mit Metadaten sollte Timo CodeMeta verwenden.

Dokumentation und Metadaten bei Szenario 3

Eine ReadMe-Datei ist auch bei Alex die Grundlage einer Dokumentation. Alex arbeitet mit Daten aus den sozialen Medien. Diese gilt es mindestens zu pseudonymisieren, woraus sich die Notwendigkeit für ein Codebook ergibt. Darüber hinaus sollte Alex die *search queries* dokumentieren. Da es für Learning Analytics noch keinen etablierten Metadatenstandard gibt, würde man an dieser Stelle auf eine Kombination der Standards LOM und Dublin Core zurückgreifen.

4.12 Zugriffssicherheit

Forschungsdaten sind die Grundlage der Forschung und somit besonders wertvoll. Sie sollten sowohl vor Verlust geschützt werden (vgl. Kapitel 4.10.2) als auch vor unerwünschtem Zugriff Dritter. Dies kann vor allem bei sensiblen und personenbezogenen Daten notwendig sein, als auch bei Daten, deren Schutz vertraglich zugesichert wurde (z. B. Firmengeheimnisse, Auftragsforschung oder Daten mit Patentpotenzial).

4.12.1 Verschlüsselung

Bei der Datenverschlüsselung handelt es sich um ein Verfahren zur Datensicherung. Sie verhindert unbefugten Zugriff auf die Daten, indem sie offene in verschlüsselte Daten anhand eines geheimen Schlüssels umwandelt. Es wird zwischen symmetrischen, asymmetrischen und hybriden Verschlüsselungen unterschieden.

Bei der symmetrischen Verschlüsselung wird der gleiche Schlüssel sowohl für die Ver- als auch Entschlüsselung genutzt. Dieses Verfahren eignet sich vor allem bei Einzelarbeiten, da sonst der Schlüssel an die Empfänger:innen weitergegeben werden müsste, was ein zusätzliches Sicherheitsrisiko darstellen kann.

Die asymmetrische Verschlüsselung nutzt zwei unterschiedliche Schlüssel: einen privaten (*Private Key*) und einen öffentlichen (*Public Key*). Beide Schlüssel sind mathematisch miteinander verknüpft. Während der öffentliche Schlüssel mit beliebigen Personen geteilt werden kann, sollte der private Schlüssel immer geheim bleiben. Die Forschungsdaten werden mit dem öffentlichen Schlüssel kodiert, können jedoch nur mit dem privaten Schlüssel entschlüsselt werden. Diese Funktion erinnert an die Funktionalität eines Tresors. Jeder kann, solange die Tresortür offen ist, etwas reinlegen und zuschließen, jedoch nur die Person, die den Code kennt, kann den Tresor wieder öffnen.

Die hybride Verschlüsselung verbindet das symmetrische Verfahren mit dem asymmetrischen. Hierbei wird der Datensatz mit einem schnell arbeitenden symmetrischen Verschlüsselungsverfahren kodiert. Im weiteren Schritt wird der dafür benutzte Schlüssel mit dem asymmetrischen Verfahren verschlüsselt und dann so an die Empfänger:innen der Nachricht versandt.

Damit die Sicherheit der Daten über Verschlüsselung tatsächlich gewährleistest werden kann, müssen nicht nur die Forschungsdaten selbst verschlüsselt werden, sondern auch alle ihre Kopien sowie das Back-up.

4.12.2 Passwortschutz

Die Daten sind nur so sicher verschlüsselt bzw. geschützt, wie das genutzte Passwort sicher ist. Da dies schnell zur Schwachstelle werden kann, ist ein starkes Passwort besonders wichtig. Nach dem Bundesamt für Sicherheit in der Informationstechnik (2021) gehören folgende Merkmale zu einem sicheren Passwort:

- gut zu merken,
- je länger, desto besser, jedoch mindestens acht Zeichen lang; ein starkes Passwort kann auch „kürzer und komplex“ oder „lang und weniger komplex“ sein,
- Verwendung aller verfügbaren Zeichen, z. B. Groß- und Kleinbuchstaben, Ziffern und Sonderzeichen,
- keine einheitlichen Passwörter,
- das Passwort sollte nicht in Wörterbüchern vorkommen.

Vermieden werden sollten Familiennamen, Geburtsdaten, Namen der Kinder, Haustiere, Freund:innen oder der Lieblingsstars. Das Hinzufügen von Sonderzeichen oder Zahlen an ein einfaches Passwort gewährt keine zusätzliche Sicherheit. Es sollte zudem nicht aus gängigen Varianten und Wiederholungs- oder Tastaturmustern wie *asdfgh* oder *1234abcd* bestehen.

Passwortmanager können bei der Erstellung und Verwaltung der Passwörter behilflich sein. Auf diese Weise muss man sich nur noch ein starkes Passwort merken (das für den Zugang zum Passwortmanager selbst) und kann dennoch unterschiedliche starke Passwörter für andere Zugänge vergeben.

4.12.3 Rechtevergabe

Zugriffsrechte bezeichnen die Regeln der Zugangskontrolle. Über die Vergabe von Berechtigungen wird festgelegt, welche Personen bzw. Personenkreise mit welchen Rechten auf bestimmte Verzeichnisse und Dateien zugreifen dürfen (vgl. Abbildung 4.13). Dabei kann zwischen Lese-, Schreib- (bzw. Änderungs-) und Ausführungsrechten gewählt werden, die in verschiedenen Abstufungen gewährt werden können: *geschlossen* (bzw. *vertraulich*), *offen* oder *eingeschränkt*. *Geschlossene* Daten (im Englischen *closed access*) sind für die Welt unzugänglich. Die Daten können zwar auffindbar und die Metadaten einsehbar sein (u. U. besteht auch eine Vorschau der Daten), die Forschungsdaten selbst bleiben jedoch unter Verschluss. Forschungsdaten, die *eingeschränkt* (im Englischen *restricted access*) zugänglich sind, sind nur für einen bestimmten Personenkreis oder für bestimmte Zwecke zugänglich. Für alle anderen Personen werden sie wie *geschlossene* Daten angezeigt. Die Daten können jedoch auch *offen* (im Englischen *open access*) sein und somit für alle uneingeschränkt zugänglich. Das ist im Sinne von Open Science und Open Data die beste Wahl, aber nicht immer für alle Daten umsetzbar.

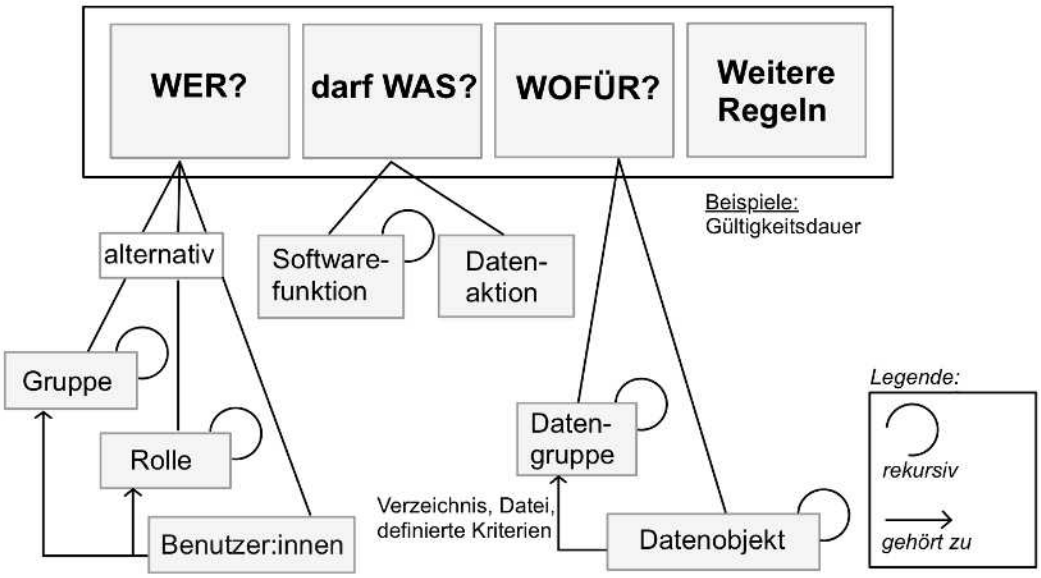


Abbildung 4.13: Vergabe von Zugriffsrechten nach VÖRBY (2012)

4.12.4 Literaturempfehlungen

- Bundesamt für Sicherheit in der Informationstechnik. (o.J). [Archivierte Version: https://perma.cc/D649-RFP4]. https://www.bsi.bund.de

4.12.5 Anwendung auf die Szenarien

Zugriffssicherheit bei Szenario 1

In Carlas Fall ist die Verschlüsselung der Daten auf dem eigenen Gerät notwendig, da mit sensiblen und personenbezogenen Daten gearbeitet wird und die Personen geschützt werden müssen. Ein Passwortschutz ist, unter Berücksichtigung der Merkmale für sichere Passwörter, notwendig.

Zugriffssicherheit bei Szenario 2

In Timos Open-Source-Projekt sind Verschlüsselung und Passwortschutz notwendig. Auch die Zugriffsrechte bei der kollaborativen Arbeit müssen geregelt werden, das bedeutet konkret: Für welche Daten in Git benötigt Kayla Lese-, Schreib- (bzw. Änderungs-) und Ausführungsrechte, damit eine Kollaboration möglich ist? Die gemeinsame Nutzung der Daten muss nach den Angaben erfolgen, wie es in der informierten Einwilligung beschrieben wurde.

Zugriffssicherheit bei Szenario 3

In Alex' BMBF-Projekt sind Verschlüsselung und Passwortschutz aufgrund der Arbeit mit sensiblen und personenbezogenen Daten notwendig. Darüber hinaus müssen Lese-, Schreib- (bzw. Änderungs-) und Ausführungsrechte für die kollaborative Arbeit geregelt werden. Insbesondere in Projekten muss abgesichert sein, ob externe Institutionen Zugriff auf die Daten bekommen dürfen (da das Projekt an mehreren deutschen Universitäten durchgeführt wird).

4.13 Publikation von Forschungsdaten

Die freie Verfügbarkeit von Daten ermöglicht eine weltweite Kollaboration. Forschende können existierende Daten suchen und darüber entscheiden, ob sie für die eigene Forschung nützlich sein können. Sie können in die eigenen Analysen integriert werden, oder ganz neue gemeinsame Forschungsfragen beantworten.

Häufig werden Datenpublikation bzw. -veröffentlichung, -austausch bzw. -verfügbarmachung, -freigabe, FAIRe Daten und offene Daten als Synonyme verwendet (im Englischen: data publication, data sharing, data release, FAIR data und open data), jedoch gibt es Unterschiede.

Das Zugänglichmachen von Forschungsdaten ist Voraussetzung für deren Verbreitung, Nachnutzung und Reproduzierbarkeit. Dies kann ein direkter Austausch zwischen Forschenden sein oder die Freigabe auf einer privaten Webseite. Um diese Daten jedoch auch nachhaltig erreichbar und somit zitierbar zu machen, ist es nicht nur notwendig, sie zur Verfügung zu stellen, sondern auch zu publizieren. Publizierte Daten werden in einem Repositorium abgelegt, wodurch sie folgende Merkmale aufweisen sollten:

- Sie sind durch Metadaten beschrieben (vgl. Kapitel 4.11.3),
- die Metadaten ermöglichen die Auffindbarkeit der Daten (vgl. FAIR-Prinzipien, Kapitel 4.5),
- den Forschungsdaten wurde ein persistenter Identifikator zugewiesen (vgl. Kapitel 4.13.4),
- der Zugang zu den Forschungsdaten wird über Zugriffsrechte geregelt (vgl. Kapitel 4.12.3).

FAIRe Daten (vgl. Kapitel 4.5) sind eine Teilmenge der publizierten Daten. Um die Daten im Sinne von Open Science zu öffnen – also offene Daten zu erhalten –, müssen die Nutzungseinschränkungen minimiert werden. Die Mindestanforderung an Open Data ist, dass ihnen offene Lizenzen vergeben wurden (vgl. Kapitel 4.14.5).

Nicht alle Forschungsdaten werden publiziert. Von denen, die publiziert wurden, sind nicht alle automatisch FAIR (z. B. ist bei der Publikation von Forschungsdaten die Interoperabilität keine Voraussetzung). Offene Daten (im Sinne von Open Data) können sich sowohl auf publizierte Forschungsdaten (mit einer DOI versehen) oder auf Forschungsdaten beziehen, die z. B. einfach auf einer Webseite zur Verfügung gestellt wurden (ohne persistente Identifikatoren). Da der offene Zugang zu Forschungsdaten nicht Voraussetzung für FAIRe Daten ist, sind nicht alle FAIRen Daten zugleich offen im Sinne von Open Data. Da nicht alle offenen Daten einen persistenten Identifikator besitzen oder in einem Repositorium abgelegt werden müssen, sind auch nicht alle offenen Daten gleichzeitig FAIR. Abbildung 4.14 zeigt die Abhängigkeiten und Teilmengen zwischen diesen Begriffen. Die grau markierte Fläche zeigt die Teilmenge der publizierten Forschungsdaten, die sowohl FAIR als auch offen sind.

4.13.1 Datenauswahl

Bei allen Forschungsvorhaben werden viele digitale Forschungsdaten produziert. Nicht alle eignen sich für die Publikation. Man muss sich daher die Frage stellen, welche Daten dafür ausgewählt werden sollten.

Alle Daten, die einer wissenschaftlichen Publikation zugrunde liegen (z. B. einem Zeitschriftenartikel oder einer Dissertation), sollten nach Möglichkeit immer publiziert werden, um die Reproduzierbarkeit der Forschung zu gewährleisten und Transparenz zu schaffen. Dabei kann es auch relevant sein, nicht nur finale Versionen, sondern auch weitere Meilenstein-Versionen zu publizieren. Bei der Auswahl sollte ein kritischer Blick auf die Qualität der Daten geworfen werden. Damit ist sowohl die inhaltliche als auch die technische Qualität gemeint. Während die Forschenden den ersten Aspekt selbst beurteilen können, ist es beim zweiten Aspekt empfehlenswert, sich im Vorfeld über nachhaltige Formate (vgl. Kapitel 4.15.2) und Interoperabilität zu informieren. Darüber hinaus sollten die Daten einzigartig sein. Es ist z. B. nicht hilfreich, eine weitere Temperaturaufzeichnung von exakt dem gleichen Ort und der gleichen Zeit zu publizieren. Ein weiterer wichtiger Punkt ist die Klärung der Rechte. Dabei gilt zu prüfen, wer die Urheber:innen sind, wer die Nutzungsrechte besitzt und ob es institutionelle Vorgaben zur Publikation von Forschungsdaten gibt.

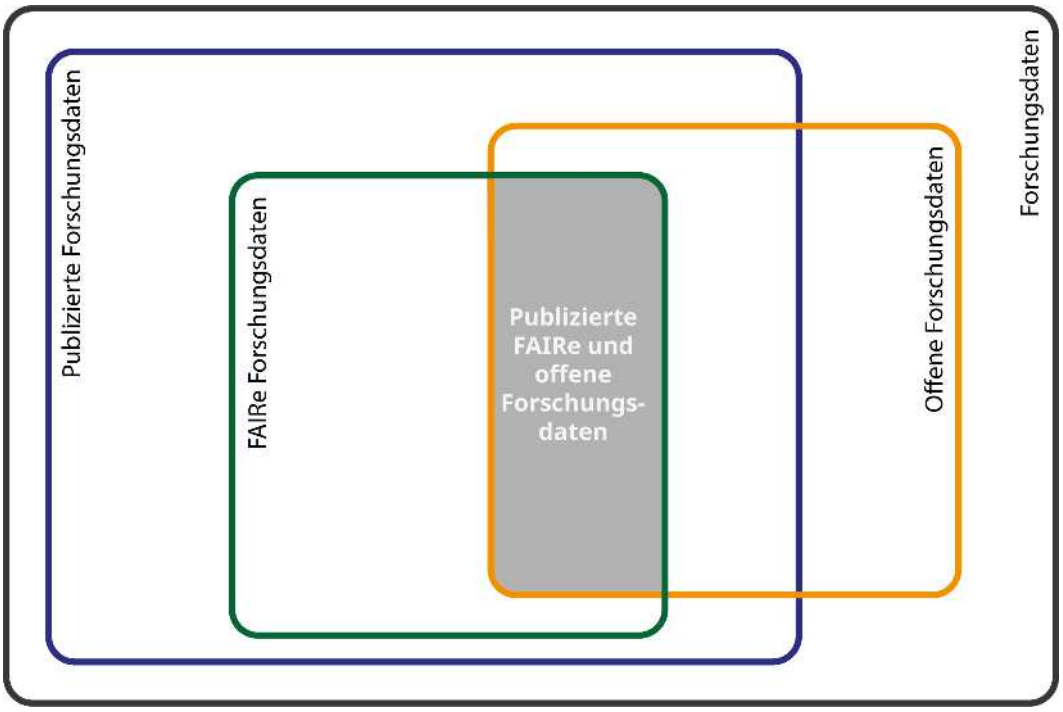


Abbildung 4.14: Abgrenzung der Begriffe im Zusammenhang mit der Publikation von Forschungsdaten

Der Prozess der Publikation von Forschungsdaten erfordert bekanntermaßen zeitlichen Aufwand. Innerhalb eines Forschungsvorhabens ist Zeit jedoch häufig knapp bemessen. Die Frage, die man sich daher stellen muss, ist: Wie hoch ist der Zeit-Kosten-Nutzen-Faktor?

4.13.2 Publikationswege

Es gibt verschiedene Wege für die Publikation von Forschungsdaten. Diese variieren in Abhängigkeit vom Inhalt der Daten und von der Art der Forschung:

- als Enhanced Publication,
- begleitend zu einer wissenschaftlichen Publikation,
- in Form eines Data Paper,
- als eigenständiges Informationsobjekt in einem Forschungsdaten- oder Software-Repositorium oder als
- Publikation von Code in einem öffentlichen Code-Repository.

Bei dem ersten Weg können Daten als Supplemente zu Veröffentlichungen von elektronischen wissenschaftlichen Fachartikeln über Verlage publiziert und mit dem Artikel direkt verknüpft werden. Diese Daten unterstützen und verdeutlichen die im Artikel präsentierten Forschungsergebnisse. Die in dieser Form veröffentlichten Daten sind in der Regel nur in Kombination mit dem zugehörigen Artikel auffindbar und zugänglich.

Der zweite Weg ermöglicht die autarke Publikation der Forschungsdaten, die eine wissenschaftliche Arbeit begleiten. Dabei werden die Forschungsdaten nicht vom Verlag mit dem Artikel verknüpft, sondern werden separat auf einem beliebigen Repositorium veröffentlicht. Auf diese Daten wird in dem Forschungsartikel verwiesen.

Data Paper widmen sich der Veröffentlichung von Informationen über publizierte Daten, die in Repositorien veröffentlicht werden. Bei diesen Informationen handelt es sich um eine ausführliche Dokumentation dieser Daten. Die Daten im Repositorium und das Data Paper in der Zeitschrift (Data Journal) werden mittels persistenten Identifikatoren (vgl. Kapitel 4.13.4) miteinander verknüpft und sind somit eindeutig auffindbar. Für die Forschungsdaten aus der Informatik gibt es eine Reihe von möglichen Data Journals z. B. F1000Research[36], Journal of Open Research Software[37], Scientific Data[38], Data[39], Journal of Open Source Software[40], Data in Brief[41] oder SoftwareX[42]. Eine Liste mit weiteren Data Journals aus unterschiedlichen Fachrichtungen führt forschungsdaten.org (2021). Wie auch bei anderen wissenschaftlichen Zeitschriften, können hier sowohl die Texte als auch die Daten einem Peer-Review unterliegen.

Der vorletzte Weg ermöglicht es, die Forschungsdaten auch ohne dazugehörigen Artikel zu publizieren – als eigenständiges Informationsobjekt in einem Repositorium.

Der letzte Weg bezieht sich direkt auf die Publikation von Software-Code in einem öffentlichen Code-Repository. Dort wird der aktuelle Code gebündelt und sicher aufbewahrt. Dieser Weg wird genauer in Kapitel 4.13.5 beleuchtet.

4.13.3 Repositorien

In den ersten vier vorgestellten Wegen der Publikation von Forschungsdaten, werden die Daten selbst in einem Repositorium abgelegt.

Repositorium

Repositorien sind Speicherorte für digitale Forschungsdaten und dienen der Veröffentlichung dieser Daten. Die Nutzer:innen können einen offenen oder eingeschränkten Zugang zu den Forschungsdaten erhalten. Viele Repositorien können auch die Funktion eines Archivs ausführen.

Je nach Repositorium können Daten, Datensätze, Experiment- und Auswertungsbeschreibungen, audiovisuelle Objekte wie Bild- und Videodateien, Modelle von Simulationen oder auch Software veröffentlicht werden. Es wird zwischen den folgenden Repositorien unterschieden:

[36] https://f1000research.com/. Archivierte Version: https://perma.cc/P8DJ-CN9G.
[37] https://openresearchsoftware.metajnl.com/. Archivierte Version: https://perma.cc/6ASU-VUJ8.
[38] https://www.nature.com/sdata/. Archivierte Version: https://perma.cc/KN4L-EL3X.
[39] https://www.mdpi.com/journal/data. Archivierte Version: https://perma.cc/GKZ4-CBNN.
[40] https://joss.theoj.org/. Archivierte Version: https://perma.cc/SE9C-2EEB.
[41] https://www.journals.elsevier.com/data-in-brief. Archivierte Version: https://perma.cc/G38K-9L4V.
[42] https://www.journals.elsevier.com/softwarex. Archivierte Version: https://perma.cc/6GYV-58JZ.

- disziplinspezifische,
- institutionelle,
- disziplinübergreifende (auch: generische).

Um die höchste Sichtbarkeit der eigenen Forschungsdaten zu erreichen, empfiehlt es sich, ein in der Community anerkanntes disziplinspezifisches Repositorium für die Publikation der Daten zu wählen. Neben dem offensichtlichen Vorteil der Sichtbarkeit in der Forschungsgemeinschaft hat es den Vorteil, dass auch das Metadatenschema disziplinspezifisch gewählt worden ist und sich somit die Suche nach den Daten einfacher gestaltet. Darüber hinaus verfügen disziplinspezifische Repositorien häufig über Kurator:innen, die die Daten vor der Aufnahme auf inhaltliche und technische Qualität sowie auf rechtliche Aspekte prüfen.

Institutionelle Repositorien sind meist kleiner und nicht unbedingt in der Community bekannt. Dafür bieten sie sich für alle Forschungsdaten an, die im Rahmen einer Abschlussarbeit oder einer anderen wissenschaftlichen Publikation im Rahmen des Anstellungsverhältnisses erstellt worden sind. Das institutionelle Repositorium kann auch neben einem disziplinspezifischen als Ort für die Zweitveröffentlichung der Daten genutzt werden. Zu den institutionellen Repositorien gehört beispielsweise der edoc-Server[43] der Humboldt-Universität zu Berlin, das Forschungsdatenrepositorium (FDR)[44] der Universität Hamburg oder das Open Access Repository and Archive (OpARA) der Technischen Universität Dresden[45].

Falls weder ein anerkanntes disziplinspezifisches noch ein institutionelles Repositorium zur Verfügung stehen, würde man ein disziplinübegreifendes wählen, bei dem eine ganze Bandbreite an Fächern abgedeckt wird. Auch hier gibt es einige, die in bestimmten Disziplinen, für die es noch keine gesonderten Repositorien gibt, in der Community verbreitet sind. Die Daten werden hier vor der Aufnahme selten kuratiert und das Metadatenschema ist wie beim institutionellen Repositorium eher generisch, was die Suche weniger spezifisch gestaltet. Zu den bekanntesten generischen Repositorien gehören Zenodo[46], Figshare[47] oder Data Dryad[48].

Kriterien für die Auswahl von Repositorien

Bei der Auswahl eines Repositoriums kann die Registry of Research Data Repositories (re3data)[49] behilflich sein. Es handelt sich dabei um einen Katalog von Repositorien aus der ganzen Welt, bei dem man nach vielerlei Merkmalen filtern kann (z. B. Disziplin, Land, Zugriffsrechte oder Datenarten).

Neben der Suche nach bestimmten Stichworten, sollten bei der Auswahl von einem geeignetem Repositorium folgende Kriterien beachtet werden (nach Gerlach et al., 2020):

[43] https://edoc.hu-berlin.de/. Archivierte Version: https://perma.cc/QH78-V4LK.
[44] https://www.fdr.uni-hamburg.de/. Archivierte Version: https://perma.cc/HK93-VXQQ.
[45] https://opara.zih.tu-dresden.de/xmlui/. Archivierte Version: https://perma.cc/8FJV-X23A.
[46] https://zenodo.org/. Archivierte Version: https://perma.cc/XYF5-ZYBD.
[47] https://figshare.com/. Archivierte Version: https://perma.cc/36LW-H7FR.
[48] https://www.datadryad.org/. Archivierte Version: https://perma.cc/C4Q7-A959.
[49] https://www.re3data.org/. Archivierte Version: https://perma.cc/6JU3-24QW.

- Zertifizierung des Repositoriums (vgl. Kapitel 4.13.3)
- Vergabe von persistenten Identifikatoren (vgl. Kapitel 4.13.4)
- Zugriffsmöglichkeiten (vgl. Kapitel 4.12.3)
- Lizenzierung (vgl. Kapitel 4.14.5)
- Metadatenschema (vgl. Kapitel 4.11.3)
- Download- und Exportmöglichkeiten
- Versionierung (vgl. Kapitel 4.9.3)
- Aufbewahrungsdauer (vgl. Kapitel 4.15)

Auch mögliche Kosten, Embargozeiten[50] oder geforderte Formate (vgl. Kapitel 4.15.2) sollten betrachtet werden.

CoreTrustSeal

Standardisierte Qualitätskriterien helfen den Forschenden bei der Auswahl eines Repositoriums. Zertifikate bzw. Gütesiegel vermitteln die Sicherheit einer langfristigen Speicherung der Daten, wobei die technische Umsetzung von der zertifizierten Initiative überprüft wird.

Eins der gängigsten Zertifizierungsorgane ist das CoreTrustSeal. Es entstand aus dem ICSU World Data System (ICSU-WDS) und dem Data Seal of Approval (DSA). Die CoreTrustSeal-Data-Repository-Zertifizierung löst die DSA-Zertifizierung und die WDS Regular Members Zertifizierung ab. Die Richtlinien zur Vergabe des Zertifikats beinhaltet 16 Kriterien (CoreTrustSeal Standards and Certification Board, 2019):

1. Auftrag/Umfang
2. Lizenzen
3. Kontinuität des Zugangs
4. Vertraulichkeit/Ethik
5. Organisatorische Infrastruktur
6. Fachliche Anleitung
7. Datenintegrität und -authentizität
8. Begutachtung
9. Dokumentierte Speicherverfahren

[50] Embargozeit ist die Zeitspanne, in der die Forschungsdaten (oder auch der wissenschaftliche Artikel) zwar auf einem Repositorium abgelegt, jedoch noch nicht der Öffentlichkeit zur Verfügung gestellt werden. Die Daten werden erst nach dieser Zeit freigeschaltet.

10. Plan für die Aufbewahrung
11. Qualität der Daten
12. Arbeitsabläufe
13. Datenermittlung und -identifizierung
14. Wiederverwendung von Daten
15. Technische Infrastruktur
16. Sicherheit

Die CoreTrustSeal-Zertifizierung ist als erster Schritt in einem globalen Rahmen für die Zertifizierung von Repositorien vorgesehen.

4.13.4 Persistente Identifikatoren

Persistenter Identifikator

Ein Persistenter Identifikator (PID) ist eine eindeutige Benennung (Referenzierung) einer digitalen Ressource (z. B. Zeitschriftenartikel oder Forschungsdaten) durch Vergabe eines Codes, der im Internet dauerhaft eindeutig referenziert werden kann. Dadurch wird verhindert, dass tote Links entstehen, wenn beispielsweise Verlage die Internetadresse eines Servers ändern (forschungsdaten.org, 2017).

PID ermöglicht die dauerhafte Auffindbarkeit und somit Zitierfähigkeit der Forschungsdaten. Neben der eindeutigen Identifizierung einer Person, eines Ortes oder eines Objektes, bietet ein PID auch einen großen Vorteil aufgrund der mit ihm verbundenen Metadaten.

Bei dem Thema FDM haben sich vor allem zwei PID etabliert: der Digital Object Identifier und die Open Researcher and Contributor ID. Im Folgenden wird auf beide eingegangen.

Digital Object Identifier

Der DOI ist ein persistenter Identifikator, der der eindeutigen Identifikation von digitalen Objekten dient. Die Referenz ist verfolgbar, dauerhaft und interoperabel. Mit dem DOI werden vor allem digitale Artikel, Datensätze, Dokumente und Medien referenziert. Der Identifikator ist eine einzigartige Folge von alphanumerischen Zeichen und besteht aus zwei Teilen: einem Präfix, der die vergebende Organisation kennzeichnet, und einem Suffix, der das Objekt identifiziert.

Die International DOI Foundation bietet die notwendige Infrastruktur z. B. für die Persistenz, das Back-up, Verhalten bei einem Ausfall u. s. w. Aus technischer Sicht wird das Handle System für die persistente Identifikation genutzt und das Vocabulary Mapping Framework für die assoziierten Metadaten.

Die Forschenden selbst können ihren Daten keine DOI vergeben, das geschieht immer über eine Einrichtung (z. B. die Bibliothek, das Repositorium o. Ä.). Die Kosten für den Erhalt eines DOI tragen die vergebenden Einrichtungen mit einer jährlichen Gebühr.

Open Researcher and Contributor ID

Die Open Researcher and Contributor ID (ORCID iD) ist eine globale, gemeinnützige Organisation, die sich aus den Beiträgen der Mitgliedsorganisationen finanziert. Das Ziel der ORCID iD ist es, transparente und vertrauenswürdige Verbindungen zwischen Forschenden, ihren Beiträgen und ihren Zugehörigkeiten zu ermöglichen, indem eine eindeutige, dauerhafte Kennung für Einzelpersonen bereitgestellt wird, die bei ihren Aktivitäten in den Bereichen Forschung, Wissenschaft und Innovation verwendet werden kann (ORCID, 2022).

Die ORCID iD hat sich mittlerweile als ein Standard, eine Art Ausweis für Wissenschaftler:innen, etabliert. Die ID kann innerhalb weniger Minuten auf der ORCID-Webseite[51] erstellt werden. Die Forschenden können das Onlineportfolio mit einer kurzen Biografie, den bisherigen Arbeitsorten, Stipendien und Publikationen anreichern. Darüber hinaus ist ORCID iD zu Web of Science[52], Scopus[53], Zenodo[54], DataCite[55], u. a. verbunden, wodurch Einträge auch automatisiert zum Profil hinzugefügt werden können.

Zu den wichtigsten Vorteilen einer ORCID iD gehören:

- eindeutige Identität als Forschende,
- korrekte Zuordnung von Forschungsergebnissen und -aktivitäten,
- zuverlässige und einfache Verbindung der wissenschaftlichen Beiträge und Zugehörigkeiten,
- Übernahme von Einträgen aus dem ORCID-Profil in andere Systeme,
- bessere Auffindbarkeit und Wiedererkennung,
- der ORCID-Eintrag ist für die Forschenden kostenlos und dauerhaft.

4.13.5 Informatikspezifische Forschungsdaten veröffentlichen

Zu den Repositorien, die nach Lucke (2021) für die Informatik-Forschungsdaten geeignet sind, gehören unter anderem:

- Archive of Formal Proofs[56]
- Camunda Modeler
- CDSTAR[57]
- Dataverse[58]
- de.NBI[59]
- ELIXIR[60]

[51] https://orcid.org/. Archivierte Version: https://perma.cc/B4TY-37S8.
[52] https://www.webofscience.com. Archivierte Version: https://perma.cc/HQ4X-WZDX.
[53] https://www.scopus.com. Archivierte Version: https://perma.cc/4TUK-77JQ.
[54] https://zenodo.org/. Archivierte Version: https://perma.cc/XYF5-ZYBD.
[55] https://datacite.org/. Archivierte Version: https://perma.cc/D6MZ-URGY.
[56] https://www.isa-afp.org/. Archivierte Version: https://perma.cc/JK9D-ZEGT.
[57] https://camunda.com/de/download/modeler/. Archivierte Version: https://perma.cc/73E7-X26Z.
[58] https://dataverse.org/. Archivierte Version: https://perma.cc/G5L6-N542.
[59] https://www.denbi.de/. Archivierte Version: https://perma.cc/VY7G-GB7U.
[60] https://elixir-europe.org/. Archivierte Version: https://perma.cc/FSJ9-EEG8.
[61] https://www.ebi.ac.uk/. Archivierte Version: https://perma.cc/ABL2-RT7S.
[62] https://fair-dom.org/. Archivierte Version: https://perma.cc/5XLC-LM8J.

- EMBL-EBI[61]
- FAIRdom[62]
- GitHub[63]
- GitLab[64]
- IACR Cryptology ePrint[65]
- iRODS[66]
- koala long-term archiving[67]
- Learning Online Network
- Linked Open Vocabularies[68]
- EMBL-EBI[69]
- Mizar Mathematical Library[70]
- OER repositories
- Open Research Knowledge Graph[71]
- PIKA[72]
- PROMISE[73]
- ReMoDD[74]
- Repository of solutions to programming exercises
- xAPI definitions[75]
- Score-P[76]
- SSELab[77]
- SNIAIOTTA[78]
- TREC Data Archive
- Virus total[79]

Software kann in speziellen Code-Repositories[80] wie GitHub, GitLab, BitBucket[81] oder SourgeForge[82] publiziert werden. Bei der Auswahl eines geeigneten Code-Repositoriums für die Software muss entschieden werden, welche Funktionalitäten benötigt werden, z. B.:

- Mailinglisten, Listenverwaltung und Archive,
- Bug-/Issue-Tracker,
- einfacher Webserver für Projekt-/Software-Seiten,
- Nachrichten,

63 https://github.com/. Archivierte Version: https://perma.cc/N3VK-P2H6.
64 https://gitlab.com/gitlab-org/gitlab. Archivierte Version: https://perma.cc/C7VK-9Y7R.
65 https://eprint.iacr.org/. Archivierte Version: https://perma.cc/B6YJ-M3NS.
66 https://irods.org/. Archivierte Version: https://perma.cc/W745-B8UW.
67 https://koala-docs.gwdg.de/. Archivierte Version: https://perma.cc/B7PK-54Q3.
68 https://lov.linkeddata.es/dataset/lov/. Archivierte Version: https://perma.cc/YAG2-W2LJ.
69 https://www.ebi.ac.uk/. Archivierte Version: https://perma.cc/V9U8-MDGZ.
70 http://mizar.org/library/. Archivierte Version: https://perma.cc/7SJB-EGBM.
71 https://www.orkg.org/orkg/. Archivierte Version: https://perma.cc/9FPC-NXUS.
72 https://pika.readthedocs.io/en/latest/index.html. Archivierte Version: https://perma.cc/N2Q5-3YPG.
73 http://promise.site.uottawa.ca/SERepository/. Archivierte Version: https://perma.cc/8CXL-SBX9.
74 https://doi.org/10.1109/ICSE.2012.6227059.
75 https://xapi.com/overview/. Archivierte Version: https://perma.cc/YL3Z-U39H.
76 https://www.vi-hps.org/projects/score-p. Archivierte Version: https://perma.cc/Z6FS-MLEJ.
77 https://sselab.de/lab1/. Archivierte Version: https://perma.cc/ZZX2-5K88.
78 http://iotta.snia.org/. Archivierte Version: https://perma.cc/8WR8-JVRP.
79 https://www.virustotal.com/gui/hunting-overview. Archivierte Version: https://perma.cc/7WPR-FBFN.
80 Ein Vergleich von Code-Repositories ist auf Wikipedia unter https://en.wikipedia.org/wiki/Comparison_of_source-code-hosting_facilities zu finden. Archivierte Version: https://perma.cc/FQ6U-SMQF.
81 http://bitbucket.org/. Archivierte Version: https://perma.cc/Y43N-GQRK.
82 http://sourceforge.net/. Archivierte Version: https://perma.cc/WCJ5-4PAP.

- Hosting/Veröffentlichung von Softwarepaketen,
- Statistikberichte (z. B. Anzahl der Übertragungen, Anzahl der Downloads),
- Foren,
- Wikis,
- Projekt-/Release-Management,
- Zugriffskontrolle.

Darüber hinaus sollten bei der Auswahl folgende Aspekte beachtet werden (N. Chue Hong, 2021):

- Wie einfach ist es, in Zukunft zusätzliche Funktionen hinzuzufügen?
- Welches Versionskontrollsystem wird verwendet, z. B. CVS, SVN, Git, Mercurial?
- Ist der Code öffentlich zugänglich?
- Wie einfach ist es, andere Elemente (z. B. eine Webseite), in das Repositorium zu integrieren?
- Wie gut ist die Unterstützung für die unterschiedlichen IDEs?
- Gibt es Unterstützung für Authentifizierungssysteme wie OpenID oder SSH-Schlüssel?
- Wo werden ähnliche Projekte wie das eigene gehostet?
- Wie schnell ist der Upload/Download?
- Wie einfach ist es, das gesamte Code-Repositorium zu sichern (Code, Mailinglisten, Tickets, etc.)
- Wie etabliert und stabil ist das Projektarchiv?
- Wie gut ist die Benutzer:innenunterstützung?
- Wie hoch ist der Aufwand für die Pflege des Projektarchivs?
- Wäre es besser, mehr als ein Code-Repositorium zu verwenden, z. B. Code in GitHub und einen Link zu Assembla für die zusätzlichen Tools?
- Wie lauten die Service Level Agreements für Betriebszeit, Ausfallzeit, Zeit zur Behebung von Ausfällen und Bandbreite?

Unabhängig von der Publikation in einem Code-Repositorium, kann eine Kopie des Codes bei Zenodo publiziert werden, um einen DOI zu erhalten und somit die Software zitierfähig zu machen. Dabei gilt zu beachten, dass Zenodo nur auf öffentliche Code-Repositorien zugreifen kann. Daher ist es wichtig, dass das Code-Repositorium, das publiziert werden soll, öffentlich ist. Falls es zu einer Organisation gehört, müssen die Eigentümer:innen der Organisation den Zugriff für die Zenodo-Anwendung möglicherweise genehmigen. Die Vergabe von Lizenzen stellt sicher, dass die Lesenden wissen, wie man den publizierten Code weiterverwenden kann (vgl. Kapitel 4.14.5).

4.13.6 Literaturempfehlungen

- Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS One, 2*(3), 5. https://doi.org/10.1371/journal.pone.0000308
- Vierkant, P., Beucke, D., Deinzer, G., Hartmann, S., Herwig, S., Höhner, K., Müller, U., Schirrwagen, J. & Summann, F. (2018). Autorenidentifikation anhand der Open Researcher and ContributorID (ORCID) – Positionspapier. Humboldt-Universität zu Berlin – edocServer. https://doi.org/10.18452/1952810.18452/19528

4.13.7 Anwendung auf die Szenarien

Publikation von Forschungsdaten bei Szenario 1

In diesem Szenario scheint eine Publikation besonders sinnvoll zu sein. Insbesondere bei der Erhebung von Interviewdaten in internationalen Kontexten wird der Kreis der Interessent:innen potenziell hoch sein. Die Daten könnten Lehrende und Forschende der Informatik interessieren, wenn sie kollaborative Prozesse untersuchen, aber auch Forschende, die einen internationalen Vergleich vornehmen wollen. Austauschprogramme für die Studierendenqualifizierung haben ebenfalls ein potenzielles Interesse an den Daten.
Obwohl bei Bachelorarbeiten nicht zwangsläufig eine Veröffentlichung der gesamten Arbeit angestrebt wird, kann eine Publikation von qualitativ hochwertigen Forschungsdaten in einem Repositorium trotzdem im Sinne der Nachnutzung von Vorteil sein. Die Wahl eines geeigneten Repositoriums ist in diesem Fall nicht trivial. Aufgrund der Arbeit mit Interviews kann es von Vorteil sein, die anonymisierten Daten nur mit einem eingeschränkten Zugriff zu veröffentlichen. Insbesondere für die Bachelorarbeit kann Carla das FDR der Universität Hamburg nutzen. Da in dem Projekt sowohl Daten über technische Tools, der Zusammenarbeit von Studierenden im Allgemeinen als auch demografische Daten erhoben werden, kann ein Forschungsinteresse in verschiedenen Bereichen liegen. So könnte sich z. B. Qualiservice für qualitative sozialwissenschaftliche Forschungsdaten anbieten. Die Nutzung von einem interdisziplinären Repositorium (z. B. Zenodo) ist jedoch auch sinnvoll und sollte anhand der konkreten Daten gegenüber dem institutionellen und fachspezifischen Repositorium abgewogen werden.

Publikation von Forschungsdaten bei Szenario 2

Ein Forschungsdatum im Fall von Timo ist die entwickelte Software, die als Open Source Software zur Verfügung steht und entsprechend auch weiterhin frei verfügbar sein muss. Die Veröffentlichung dieses Datums ist besonders wichtig, um Anwender:innen transparent zu zeigen, welche Daten erhoben und wie sie verarbeitet werden. Dies soll die Nachvollziehbarkeit gewährleisten. Zur Publikation können GitHub oder GitLab als Code-Repository verwendet werden. Darüber hinaus kann ein DOI über ein anderes

Repositorium (z. B. Zenodo) erzeugt werden, bei der die zugehörige Dokumentation für die Software hinterlegt ist.

Publikation von Forschungsdaten bei Szenario 3

Die Publikation der Forschungsdaten wird nach den BMBF-Standards erwartet, da sie zur Transparenz von Forschung beiträgt. Darüber hinaus wird eine nachhaltige Forschung angestrebt, da insbesondere die Algorithmen zur Auswertung von Twitter-Daten auch für andere Forschende hilfreich sein können. Leider ist es in Alex' Fall schwierig, da die AGB von Twitter die Publikation der Sammlung der Twitterdaten erschweren, wenn nicht sogar unmöglich machen (vgl. Kapitel 4.17.6). Somit könnte nur der Algorithmus selbst publiziert werden. Dafür können GitHub oder GitLab als Code-Repositorium verwendet werden. Zusätzlich kann ein DOI über ein anderes Repositorium (z. B. Zenodo) für die Software mit zugehöriger Dokumentation erzeugt werden.

4.14 Urheberrecht und Lizenzierung

4.14.1 Einführung in das Urheberrechtsgesetz

Das Gesetz über Urheberrecht und verwandte Schutzrechte (kurz Urheberrechtsgesetz (UrhG)) dient dem Schutz des geistigen Eigentums von im weitesten Sinne Kreativschaffenden und der Kreativwirtschaft (Kreutzer & Lahmann, 2021), wobei Urheber:innen nur natürliche Personen sein können und keine Unternehmen. Der Schutz wird aufgrund des Grundrechts des Eigentums gemäß der Verfassung gerechtfertigt und als Belohnung für die Urheber dafür angesehen, dass sie die Kultur und Gesellschaft mit ihren Werken bereichern (Amt der Europäischen Union für geistiges Eigentum, 2021). Das Urheberrecht erteilt somit den Urheber:innen das alleinige Recht an der Veröffentlichung, Vervielfältigung, Aufführung und dem Verleih ihres Werkes und bestimmt die Folgen der Verletzung der Rechte.

4.14.2 Verwandte Schutzrechte

Im Gegensatz zum Urheberrecht, dienen die verwandten Schutzrechte (auch: Leistungsschutzrechte) dem Schutz bestimmter Leistungen im kulturellen Bereich, die selbst aber keine Werke sind, sondern als künstlerische Leistungen oder (technische, finanzielle oder organisatorische) Investitionen betrachtet werden, die für die Kultur als ausreichend wichtig und daher als schutzwürdig angesehen werden (Amt der Europäischen Union für geistiges Eigentum, 2021). Dabei handelt es sich vor allem um Schutzrechte für Fotograf:innen, Interpreter:innen oder Vermittler:innen von Inhalten (z. B. Tonträger- oder Datenbankhersteller:innen).

4.14.3 Schutzwürdige Werke und Leistungen

Damit ein Werk urheberrechtlich geschützt ist, muss es Schöpfungshöhe erreichen, d. h., es muss einer schöpferischen Leistung und mit Originalität entstanden sein. In der Regel bedeu-

tet das, dass eine geistige Leistung erbracht werden muss, die in einer konkreten Schöpfung zum Ausdruck gebracht wurde. Laut § 2 Abs. 1 UrhG werden „Werke der Literatur, Wissenschaft und Kunst“ geschützt, jedoch können auch Studienarbeiten oder Werbegrafiken urheberrechtlich geschützt sein, wenn sie gewissen individuelle Züge aufweisen.

Nicht geschützt sind dagegen Ideen, Informationen, Fakten, unstrukturierte Messdaten oder andere abstrakte Ressourcen, die lediglich die Basis eines Werkes bilden und notwendige Voraussetzung für kulturelles Schaffen, Wissenschaft, Kommunikation, Meinungsfreiheit und vieles mehr sind (Kreutzer & Lahmann, 2021). Auch sehr einfache Texte, Musik oder Computerprogramm genießen keinen Urheberrechtsschutz. Daraus wird schnell ersichtlich, dass auch die meisten (quantitativen) Forschungsdaten, die in den MINT-Fächern zur Anwendung kommen, keinen Urheberrechtsschutz genießen. Eine grundsätzliche Pauschalisierung ist dennoch schwierig, da z. B. auch in MINT-Fächern qualitative Forschung betrieben wird (Brettschneider et al., 2021). Es entsteht also regelmäßig ein Konglomerat aus geschützten und ungeschützten Daten, was zu einer erheblichen Komplexität der rechtlichen Beurteilung führen kann. In Zweifelsfällen sollte daher von einer grundsätzlichen Schutzfähigkeit der Forschungsdaten ausgegangen (Lauber-Rönsberg, 2021) und eine Einzelfallüberprüfung durchgeführt werden.

Grundsätzlich sind Computerprogramme urheberrechtlich geschützt, wenn es sich um „individuelle Werke“ handelt, die das Ergebnis der eigenen geistigen Schöpfung ihrer Urheber:innen sind (§ 69a Abs. 3 Satz 1 UrhG). Wird hingegen ein Computerprogramm im Rahmen eines Arbeits- oder Dienstverhältnisses erstellt, liegen die Nutzungsrechte an diesem dienstlich geschaffenen Computerprogramm in der Regel beim Arbeitgeber, z. B. der Hochschule (§ 69b UrhG).

Anders als beim Urheberrecht, werden bei der Entscheidung, ob bestimmte Leistungen schutzfähig sind, keine qualitativen Anforderungen gestellt. Die Leistungserbringer:innen erhalten die Rechte unabhängig vom Umfang oder von der Qualität des Werkes selbst (Kreutzer & Lahmann, 2021).

Gemeinfreiheit

Nach 70 Jahren, beginnend mit dem Ablauf des Todesjahres des Urhebers (*post mortem auctoris*, abgekürzt p. m. a.), beginnt die Gemeinfreiheit geschützter Werke. Dies bedeutet, dass keine Rechte mehr bestehen und dass der jeweilige Inhalt frei von jeglichen Einschränkungen oder Rechtspflichten genutzt werden darf.

4.14.4 Autorenschaft

Wenn mehrere Personen an der Erstellung eines Werkes beteiligt waren, sind sie alle Urheber:innen und es steht ihnen das Urheberrecht gemeinsam zu (Miturheberschaft, § 8 UrhG). Nicht selten führt das zu Streitigkeiten, da alle Co-Autor:innen nach dieser gesetzlichen Regelung gemeinsam über die Veröffentlichung entscheiden müssen, jedoch unterschiedlich viel zum Werk beigetragen haben. Der Grad des Beitrags lässt sich manchmal an der Reihenfolge der genannten Autor:innen erkennen, in der die Namen in der Kopfzeile der Veröffentlichung

erscheinen, aber es ist oft schwierig, die relativen Beiträge auf diese Weise genau zu bestimmen.

Die Contributor Roles Taxonomy (CASRAI, 2021) soll dabei helfen, dieses Problem zu lösen, indem der Grad der Beiträge der Autor:innen anhand 14 vordefinierter Rollen bestimmt wird. Zu diesen Rollen gehören:

- Konzeptualisierung
- Datenkuratierung
- Formale Analyse
- Akquisition von Fördermitteln
- Forschung
- Methodik
- Projektverwaltung
- Ressourcen
- Software
- Beaufsichtigung
- Validierung
- Visualisierung
- Schreiben – ursprünglicher Entwurf
- Schreiben – Überprüfung und Bearbeitung

Dieser Ansatz bringt nicht nur mehr Transparenz in den Grad des Beitrages der Autor:innen, sondern verbessert auch die Konsistenz, Nützlichkeit und Vergleichbarkeit der verschiedenen Daten, die für analytische Erkenntnisse und den täglichen Forschungsbetrieb benötigt werden.

4.14.5 Übertragung und Einräumung von Nutzungsrechten

Urheberrechte als solche können nicht übertragen werden. Urheber:innen können Dritten lediglich die Nutzung ihrer Werke erlauben. Dies wird als Einräumung bzw. Übertragung der Nutzungs- oder Verwertungsrechte bezeichnet (Kreutzer & Lahmann, 2021).

Die Übertragung von Rechten kann eine unterschiedliche Reichweite haben. Eine sehr weitreichende Einräumung der Rechte ist die Übertragung der Exlusivnutzungsrechte mit Einmalvergütung (*total buy-out*). Dieser Weg führt dazu, dass die Urheber:innen selbst weitgehend von der Nutzung des eigenen Werkes ausgeschlossen werden. Wenn z. B. Wissenschaftlerin:innen bei der Publikation ihrer Artikel dem Verlag das ausschließliche Recht übertragen, den Beitrag abzudrucken, online zu stellen und zu verbreiten, dürfen sie selbst diese Handlungen dann nicht mehr ohne Erlaubnis des Verlags vornehmen (Kreutzer & Lahmann, 2021).

Die Übertragung von Rechten erfolgt meist über Verträge. Es handelt sich dabei meist um Lizenzverträge, kann aber auch in Form von Arbeits-, Dienst- oder Werkverträgen geschehen. Auch Open-Content-Lizenzen sind Formulare für (Lizenz-)Verträge. Die Verträge

müssen nicht zwingend schriftlich, sondern können auch mündlich oder implizit („konkludent“) geschlossen werden (Kreutzer & Lahmann, 2021).

Lizenzierung

Um urheberrechtlich geschützte Forschungsdaten (oder Werke im Allgemeinen) dem breiten Publikum zu öffnen, muss eine freie und offene Lizenz vergeben werden. Diese Lizenzen zeichnen sich dadurch aus, dass sie die Werknutzung weitestgehend gestatten und den Nutzer:innen nur wenige Pflichten und Restriktionen auferlegen (Kreutzer & Lahmann, 2021). Es gibt unterschiedliche Lizenzmodelle, die für Forschungsdaten gut geeignet sind. Am gängigsten ist dabei die Nutzung von Creative-Commons(CC)-Lizenzen[83].

Zusätzlich zu dem eigentlichen Lizenzvertrag werden bei Creative Commons die Lizenzbedingungen meist anhand standardisierter Symbole beschrieben:


Public Domain


Namensenennung


Weitergabe unter gleichen Bedingungen


Keine Bearbeitung


Nicht-kommerziell

Diese Symbole ermöglichen eine eindeutige und unmissverständliche Zuordnung der Lizenz.

Creative-Commons-Lizenzen

Aus den genannten fünf Lizenzbedingungen ergeben sich sechs mögliche CC-Lizenzen:

- CC BY (Namensnennung)
- CC BY-SA (Namensnennung – Weitergabe unter gleichen Bedingungen)
- CC BY-ND (Namensnennung – Keine Bearbeitung)
- CC BY-NC (Namensnennung – Nicht-kommerziell)
- CC BY-NC-SA (Namensnennung – Nicht-kommerziell – Weitergabe unter gleichen Bedingungen)
- CC BY-NC-ND (Namensnennung – Nicht-kommerziell – Keine Bearbeitung)

Eine besondere Form der Lizenzierung stellt CC0 (Public Domain Dedication) dar (vgl. Abbildung 4.15).
Mit dieser Lizenz können die Urheber:innen ihr Werk in die Gemeinfreiheit entlassen. Das bedeutet, dass eine freie Nachnutzung ohne jegliche Einschränkungen möglich ist. In Deutschland ist der vollständige Verzicht auf das Urheberrecht nicht möglich, die Vergabe von einer

[83] https://creativecommons.org/. Archivierte Version: https://perma.cc/825H-QFL9.

Abbildung 4.15: Public Domain Dedication

CC0-Lizenz ermöglicht es jedoch, auf Rechte am eigenen Werk im vollen gesetzlich zulässigen Umfang zu verzichten und klärt eindeutig den Status des eigenen Werks weltweit.

Die CC0-Lizenz sollte allerdings nicht für gemeinfreie Werke verwendet werden. Zur Kennzeichnung von Werken, die bereits frei von bekannten Urheberrechtsbeschränkungen und weltweit gemeinfrei sind, sollte die Public Domain Mark genutzt werden (vgl. Abbildung 4.16).

Abbildung 4.16: Public Domain Mark

Zudem sind die Lizenzbedingungen der Creative-Commons-Lizenzen maschinenlesbar, indem sie in Form von Metadaten im RDF-Format auf den Seiten der Verwender:innen hinterlegt werden. Auf diese Weise kann gezielt nach Ressourcen, die unter einer CC-Lizenz veröffentlicht wurden, gesucht werden.

Neben den CC-Lizenzen sind auch die ODC[84] hervorzuheben. Diese wurden speziell für die Lizenzierung von Daten und Datenbanken entwickelt. Seit der Version 4.0 International von Creative Commons gibt es jedoch kaum noch Unterschiede zwischen den Lizenzen. Die ODC-Lizenzen eignen sich vor allem für Datenbankrechte, Patentrechte und Warenzeichen. Bei den Datenbanklizenzen können auch für die in entsprechend lizenzierten Datenbanken enthaltenen Inhalte unabhängige Bedingungen festgelegt werden.

Open Data Commons

Es gibt drei ODC-Lizenzen:

- Open Data Commons Open Database License (ODbL) – entspricht CC BY-SA
- Open Data Commons Attribution License (ODC-By) – entspricht CC BY
- Open Data Commons Public Domain Dedication and License (PDDL) – entspricht CC0

Zu den Lizenzen, die mit der Open Definition konform sind und somit die Idee von Open Science fördern, gehören: CC0, CC BY, CC BY-SA, PDDL, ODC-By und ODbL (Open Knowledge Foundation, 2021). Abbildung 4.17 stellt die Konformität aller CC- und ODC-Lizenzen in Bezug auf Open Access dar.

[84] https://opendatacommons.org/. Archivierte Version: https://perma.cc/9EL8-CGXJ.

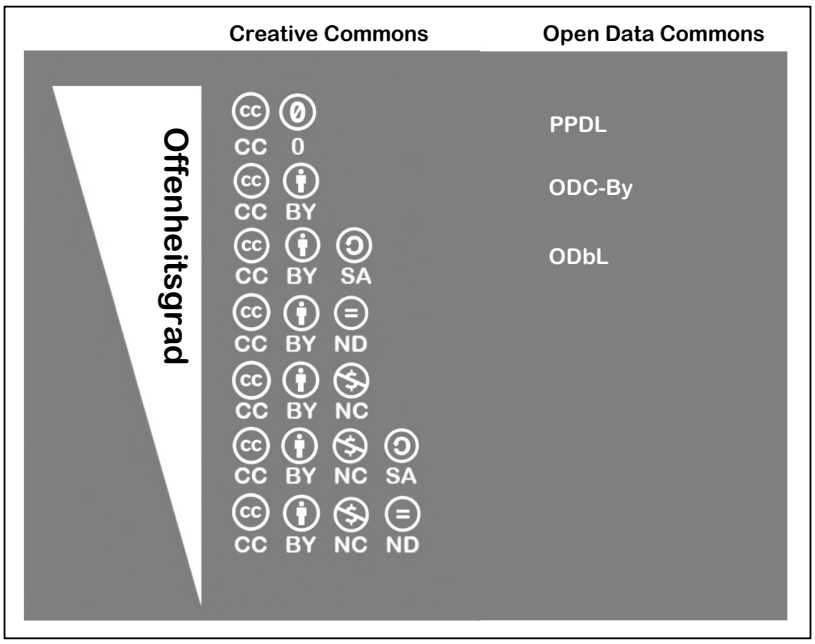


Abbildung 4.17: Open-Access Konformität der CC- und ODC-Lizenzen

Softwarelizenzierung

Beide Lizenzmodelle – sowohl Creative Commons als auch Open Data Commons – eignen sich nicht für die Lizenzierung von Software. Um den besonderen Anforderungen dieses Formats gerecht zu werden (Einverständnis zur Installation, Veränderung, Ausführung, Zweck oder Ort der Nutzung, Anzahl der Benutzenden etc.), empfiehlt es sich, eine der gängigen Softwarelizenzen zu benutzen. Es gibt unterschiedliche Softwarelizenzen, die für die Lizenzierung von Quellcode genutzt werden können. Besonders empfehlenswert sind Open Source-Lizenzen, die die freie Nutzung, Modifizierung und Weitergabe ermöglichen. Dabei wird zwischen den *permissiven* und *non-permissiven* Lizenzen unterschieden. Während es bei den permissiven (auch: freigebigen) Lizenzen wenige Beschränkungen bei der Nutzung des Quellcodes gibt, müssen Nutzer:innen bei den non-permissiven strengere Restriktionen beachten, wenn sie den Code weiterentwickeln oder mit anderem Quellcode verbinden wollen.

Bei der Nutzung von Softwarelizenzen ist es wichtig, den Copyleft-Effekt zu berücksichtigen. Copyleft bedeutet nicht den Verzicht auf das Copyright. Es verpflichtet die Lizenznehmer:innen, jegliche Bearbeitung der Software (z. B. Erweiterung, Veränderung) unter die Lizenz der ursprünglichen Software zu stellen. Somit wird sicher gestellt, dass Weiterentwicklungen von freien Programmen frei sind und auch frei bleiben. Die einzige Ausnahme davon stellen die Lizenzgeber:innen selbst dar, da sie nicht an das eigene Copyleft gebunden sind und neue Versionen auch unter proprietärer Lizenz veröffentlichen können.

Open Source-Lizenzen

- MIT-Lizenz (auch: X-Lizenz oder X11-Lizenz; non-Copyleft-Lizenz)
- Apache-Lizenz[a] (non-Copyleft-Lizenz)
- BSD (non-Copyleft-Lizenz; die Software kann für kommerzielle sowie proprietäre Produkte genutzt werden)
- GNU Lesser General Public License (LGPL)[b] (verfolgt ein beschränktes Copyleft, d. h. Verlinkungen in anderen Programmen sind möglich, ohne neue Produkte unter der LGPL lizenzieren zu müssen)
- GNU General Public License (GPL)[c] (non-permissiv; verfolgt ein striktes Copyleft und garantiert somit, dass freie Software frei bleibt)

[a] https://www.apache.org/licenses/LICENSE-2.0. Archivierte Version: https://perma.cc/3URE-GUSK.
[b] https://www.gnu.org/licenses/lgpl-3.0.de.html. Archivierte Version: https://perma.cc/2NPJ-MWAM.
[c] http://www.gnu.org/licenses/gpl-3.0.de.html. Archivierte Version: https://perma.cc/CD8D-SBSY.

Schutzrechtsberühmung

Lizenzen dienen dazu, die Nutzung eines geschützten Werkes oder einer geschützten Leistung zu ermöglichen. Das Einräumen von Lizenzen für ungeschützte Werke ist hingegen unrechtmäßig und wird als Copyfraud bezeichnet.

Copyfraud (Schutzrechtsberühmung)

Der Versuch, urheberrechtliche Ansprüche für gemeinfreie Werke (aus der „Public Domain Mark“ von Creative Commons oder gemeinfrei nach UrhG) zu erheben.

Creative-Commons-Lizenzen antizipieren das Problem des Copyfrauds und stellen ausdrücklich klar, dass die Lizenzvergabe auf nicht urheberrechtlich geschütztes Material keine Wirkung entfaltet. Dennoch bleibt die faktische Wirkung, die andere von der Nachnutzung der Daten abhalten könnte (Brettschneider et al., 2021).

4.14.6 Literaturempfehlungen

- Baumann, P., Krahn, P. & Lauber-Rönsberg, A. (2021). *Forschungsdatenmanagement und Recht. Datenschutz-, Urheber- und Vertragsrecht.* W. Neugebauer.
- Brettschneider, P., Biernacka, K., Böker, E., Danker, S. A., Jacob, J., Perry, A., Wiljes, C. & Wuttke, U. (2021). Urheberrecht und Lizenzierung bei Forschungsdaten. https://doi.org/10.5281/zenodo.5243232

- Kreutzer, T. & Lahmann, H. (2021). *Rechtsfragen bei Open Science. Ein Leitfaden.* Hamburg University Press. https://doi.org/10.15460/HUP.211

4.14.7 Anwendung auf die Szenarien

Urheberrecht bei Szenario 1

In diesem Szenario fallen qualitative Daten als Forschungsdaten an, die der Schöpfungshöhe im Urheberrecht entsprechen. Dementsprechend ist Carla die Urheberin der Daten und sollte die Daten für die Überprüfbarkeit und Nachnutzung zur Verfügung stellen. Sie könnte dafür eine CC-BY-Lizenz vergeben, und somit die Nachnutzung weniger restriktiv halten. Das hätte den Vorteil, dass auch Projekte/Firmen, die forschungsnah sind, trotzdem aber auch kommerzielle Interessen haben, die Daten nutzen dürfen.

Urheberrecht bei Szenario 2

Timo ist der Urheber der Software, wobei er Nutzungsrechte an das Projekt übertragen muss, damit die Software genutzt werden kann. Damit die Software urheberrechtlich geschützt ist, muss sie eigene Lösungsansätze enthalten und darf nicht nur aus Code bestehen, der bereits in anderer freier Software in genau dieser Form bereits offen zur Verfügung gestellt wurde. Um die Software wieder frei verfügbar zu machen im Open Source-Projekt, muss darauf geachtet werden, dass keine Code-Fragmente genutzt werden, die unter einer restriktiveren Lizenz stehen. Beispielsweise muss im Falle der Wiederverwendung von Code, ebenfalls auf die Nutzungslizenz geachtet werden, das heißt, ob der Code überhaupt verändert und verwendet werden darf. Es ist zu prüfen, welche Restriktionen sich auf dieser Grundlage für die Lizenz von Timos Software ergeben.

Urheberrecht bei Szenario 3

In Alex' Fall muss zunächst überprüft werden, wie Twitters AGB eine Nutzung der Daten regeln. Da Alex einen Arbeitsvertrag mit der Humboldt-Universität zu Berlin im Rahmen des BMBF-Projekts abgeschlossen hat, ist zu klären, inwieweit notwendige Verwertungsrechte an die HU übertragen wurden. Das kann zum Beispiel der Fall sein, wenn eine Software zur Analyse von Twitterdaten von Alex geschrieben wird, die auch später im Projekt genutzt werden muss und nicht von Alex' Anstellungsverhältnis abhängen darf.

4.15 Langzeitarchivierung

Sowohl die Gute Wissenschaftliche Praxis der DFG als auch viele andere internationale Förderer verlangen eine Aufbewahrungsfrist von Forschungsergebnissen sowie der zugrundeliegenden Forschungsdaten nach Möglichkeit für einen Zeitraum von zehn Jahren (Deutsche Forschungsgemeinschaft, 2019, S. 22). Dabei sollten die Daten entweder in der Einrichtung, in

der sie entstanden sind, oder auf einem Repositorium (vgl. Kapitel 4.13) bzw. in einem Langzeitarchiv zugänglich gemacht werden. Die Aufbewahrungsfrist beginnt mit der öffentlichen Zugänglichmachung der Forschungsdaten.

Langzeitarchivierung

„Unter der LZA von Daten versteht man ein Verfahren, das Daten (z. B. Forschungsdaten) für einen unbestimmten Zeitraum, der über nicht vorhersehbare technologische und soziokulturelle Veränderungen hinausreicht, verfügbar und für Menschen interpretierbar hält.“ (forschungsdaten.org, 2019)
Die LZA strebt den Erhalt der Integrität, Authentizität, Zugänglichkeit und Verständlichkeit der archivierten Daten an.

Die LZA soll demnach die langfristige Nutzbarkeit der Forschungsdaten über einen undefinierten Zeitraum hinweg sicherstellen. Um dies zu gewährleisten, muss die Standardisierung von Arbeitsschritten und Prozessen beachtet werden. Die Integrität und Authentizität dieser Daten können mit unter durch die Anwendung von Standards bei ihrer Erzeugung und Dokumentation erreicht werden. Dabei sollten insbesondere technische, rechtliche, inhaltliche und Provenienzmetadaten vergeben werden (vgl. Kapitel 4.11.3). Darüber hinaus bedarf es einer regelmäßigen Überprüfung der Daten im Hinblick auf den technischen und soziokulturellen Wandel, um die Lesbarkeit dieser Forschungsdaten dauerhaft zu gewährleisten. Dies kann mithilfe von Emulation oder Migration geschehen. Dafür dürfen die Daten nicht untrennbar mit einem Datenträger oder Auslesegerät verbunden sein.

Bei der Migration werden die Daten immer wieder in neuen Versionen abgespeichert, um sie so dem neuen Umfeld anzupassen. Dabei kann sich nicht nur der Datenträger verändern, sondern auch das Dateiformat, die -darstellung oder -navigation, somit geht die Authentizität der Daten verloren. Um dem entgegenzuwirken, ist es also notwendig, akribisch zu dokumentieren, welche Software in welcher Version, welches Betriebssystem und Hardware benötigt werden, um den ursprünglichen Zustand der Daten wiederherstellen zu können. Eine verlustfreie Migration ist nur dann möglich, wenn:

- das Originalformat eindeutig spezifiziert ist,
- das Zielformat eindeutig spezifiziert ist,
- diese Spezifikationen bekannt sind und
- eine Übersetzung von dem einen in das andere Format ohne Probleme möglich ist (Funk, 2010b).

Die Emulation hingegen basiert auf der Nachstellung der originären Umgebung der Daten (Betriebssystem, Software und/oder Hardware), also der Anpassung der neuen Umgebung an die Daten. So werden die originalen Daten nicht verändert, stattdessen muss man für jede neue Hardwarearchitektur die Emulationssoftware anpassen, die im schlimmsten Fall jedes Mal neu entwickelt werden muss (Funk, 2010a).

NFDIxCS[85] plant einen Research Data Management Container, der die Emulation von Forschungsdaten sowie Forschungssoftware erleichtern soll (vgl. Abbildung 4.18). Neben den Daten, den Verweisen auf verwendete Software und der Ausführungsumgebung, werden weitere Mechanismen diesem Container hinzugefügt, die die Nutzung der Daten in der Zukunft ermöglichen sollen. Dazu gehören Mechanismen zur Zugriffskontrolle, Workflows sowie Filter und Transformationen (z. B. zur Verschlüsselung oder Pseudonymisierung der Daten; Lucke, 2021).

Abbildung 4.18: Der Research Data Management Container nach Lucke (2021)

4.15.1 Abgrenzung der Begrifflichkeiten

Während die reine physische Speicherung der Daten darauf abzielt, die Daten im Zustand zum Zeitpunkt ihrer Ablieferung (*Ingest*) zu erhalten, ist es bei der LZA auch relevant, die Lesbarkeit dieser Daten in der undefinierten Zukunft trotz technischer und soziokultureller Veränderungen zu gewährleisten.

Auch unterscheidet sich LZA vom Back-up. Während es eine regelmäßige automatische Sicherung aller Daten und aller Versionen ist, um den Datenverlust vorzubeugen (z. B. durch technischen Defekt oder menschlichen Fehler), werden nur endgültige Daten und Versionen langfristig in einem LZA archiviert.

Sowohl Repositorien als auch Langzeitarchive sind dafür gedacht, Daten zu halten. Der Unterschied jedoch besteht in der primären Zielsetzung der Infrastruktur: Repositorien sollen die Daten für die Öffentlichkeit verfügbar machen (mit möglichen Zugangsbeschränkungen),

[85] https://nfdixcs.org/. Archivierte Version: https://perma.cc/5LUY-WA5H.

während die Langzeitarchive die Daten für unbestimmte (bzw. eine bestimmte lange) Zeit aufbewahren und lesbar erhalten sollen. Repositorien können gleichzeitig Langzeitarchive sein, in dem sie die langfristige Aufbewahrung der Daten durch notwendige konzeptionelle und technische Infrastruktur gewährleisten. Das bietet jedoch nicht jedes Repositorium an. Andererseits ist auch nicht jedes Langzeitarchiv gleichzeitig ein Repositorium. Die Daten im Archiv können verschlossen sein und der Öffentlichkeit unzugänglich gemacht werden. Auch die Auswahl der Daten zur Publikation in einem Repositorium kann sich von der Wahl der zu archivierenden Daten unterscheiden. Im ersteren Fall werden Daten mit möglichem Nachnutzungspotenzial bzw. Daten, die eine Publikation zugrunde liegen gewählt. Im zweiten Fall können auch zusätzliche Daten z. B. Projektdokumente und -berichte für Dokumentationszwecke mit abgelegt werden.

4.15.2 Nachhaltige Dateiformate

Im digitalen Forschungsalltag ist die Handhabung unterschiedlicher Dateiformate nicht wegzudenken. Viele Forschende kennen auch die damit verbundenen Probleme.

Im Kontext der LZA werden verschiedene Formatkategorien unterschieden. Zu den geläufigsten gehören: Text, Bild, Audio und Video. Darüber hinaus können in Abhängigkeit vom Archiv noch weitere Kategorien gelistet werden, z. B. Tabellenkalkulation, Datenbanken, medizinische Bildformate, Statistikdaten u. v. m. Innerhalb dieser Formatkategorien gibt es bestimmte Formatempfehlungen (vgl. Tabelle 4.6), die als Empfehlungsgrundlage für Langzeitarchive dienen können.

Diese Formate zeichnen sich dadurch aus, dass sie anhand bestimmter Kategorien als langzeittauglich gelten. Zu diesen Kategorien zählen unter anderem die Offenheit, Verbreitung, Funktionalität, Verifizierbarkeit, sowie Best Practices und die zukunftsfähige Perspektive der Formate. Zu vermeiden sind insbesondere proprietäre Formate, da sie an bestimmte Software gebunden sind. Wenn möglich sollte hier immer eine offene Alternative gewählt werden oder beide Formate – das proprietäre und der offene Export – archiviert werden.

4.15.3 Anforderungen an Langzeitarchive

Ähnlich wie bei Repositorien, gibt es bestimmte Merkmale zur Erkennung von vertrauenswürdigen Langzeitarchiven. Dazu gehören Verfahren, die die Umsetzung der grundlegenden Funktionalitäten und damit die Vertrauenswürdigkeit von Langzeitarchiven prüfen, wie z. B.:

- das CoreTrustSeal (CTS) (vgl. Kapitel 4.13.3),
- der nestor-Siegel/DIN 31644 oder
- die ISO 16363.

Neben den Siegeln für vertrauenswürdige Langzeitarchive, sind auch die technischen Anforderungen, die Kosten der Services, die Zugänglichmachung der Daten, sowie die Langlebigkeit des Dienstleister:innen bei der Auswahl des Langzeitarchivs von Bedeutung.

Tabelle 4.6: Vergleich der Empfehlungen für archivische Dateiformate (Überschneidungen sind anhand einer Hervorhebung kenntlich gemacht)

Formatkategorie	**Formatempfehlungen nach Koordinatiosstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST) (2021)**	**Formatempfehlungen nach IANUS (2016)**
Text	**TXT**, PDF, **PDF/A**, ODF, OOXML	DOCX, ODT, **PDF/A**, **TXT**
Bild	**TIFF**, JPEG, JPEG2000, **PNG**, **DNG**, PDF/A-2 für Bilddaten, Vektorgrafiken	**TIFF**, **DNG**, **PNG**, SVG
Audio	**WAV**, **FLAC**, ALAC, **MP3**, Ogg Vorbis	**FLAC**, **WAV**, BWF, Matroska, MBWF, AAC, **MP3**
Video	Uncompressed Video, Digital Video, FFV1, **MPEG-2**, **MPEG-4**, **MJPEG2000**, ProRes, Containerformate	MKV, **MJPEG2000**, **MPEG-4**, MXF, **MPEG-2**
Tabellenkalkulation	XLS, **ODS**, **XLSX**, PDF/A-2 für Tabellenkalkulation	CSV, TSV, **ODS**, **XLSX**, XML, HTML
Datenbanken	CSV, **SIARD**, SQLX, **SQL**	**SIARD**, **SQL**, XML
Hypertext	**HTML**, HTML5, **MHTML**, ARC, **WARC**, **PDF/A-2**	**HTML**, **MHTML**, **WARC**, **PDF/A-2**
CAD/CAM	**DWG**, **DXF**, STEPS, IFC, Vektorgrafik in PDF	**DWG**, **DXF**, PDF/A

4.15.4 Literaturempfehlungen

- Whyte, A. (2014). Five steps to decide what to keep: a checklist for appraising research data. [Archivierte Version: https://perma.cc/HG6N-W4UZ]. Digital Curation Centre. https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep

4.15.5 Anwendung auf die Szenarien

Langzeitarchivierung bei Szenario 1

Es sollte darauf geachtet werden, dass nachhaltige Formate genutzt werden. Beispielsweise können .txt oder .docx für Interviewtranskripte und .mp3 für die Audio- bzw. .mpeg-4 für Videodaten genutzt werden, um eine LZA zu ermöglichen. Falls eine direkte Bearbeitung der Transkripte unterbunden werden soll, ist auch die Speicherung unter .pdf (PDF/A) möglich.

Langzeitarchivierung bei Szenario 2

Da Timo im Projekt Software entwickelt, die an verschiedene Systeme angebunden werden soll, ist insbesondere bei der Schnittstellenentwicklung auf die Verwendung nachhaltiger Formate zu achten. Ist die Software bereits bei Git verfügbar, so kann sie z. B. über Zenodo publiziert und somit auch archiviert werden.

Langzeitarchivierung bei Szenario 3

Software als Forschungsdatum sollte langzeitarchiviert werden. Ist die Software bereits bei Git verfügbar, so kann sie z. B. über Zenodo publiziert und somit auch archiviert werden.

4.16 Nachnutzung

Die Nachnutzung von Forschungsdaten bringt viele Vorteile, sowohl für die Forschenden, die die Daten publizieren, als auch für die Forschenden, die diese nachnutzen. Publizierte Daten steigern die wissenschaftliche Reputation und erhöhen die Ziatationsrate. Die Nachnutzung publizierter Forschungsdaten ermöglicht den Vergleich verschiedener Stichproben oder Vergleiche über Zeit. Da die eigene Erhebung entfällt, werden auf diese Weise sowohl Zeit- als auch Kostenressourcen gespart.

4.16.1 Forschungsdaten finden

Es gibt unterschiedliche Möglichkeiten, nach geeigneten Forschungsdaten zu suchen. Passende Daten für die Nachnutzung zu finden, erfordert meist eine Suche in verschiedenen Quellen (Biernacka et al., 2021a). Dazu gehören:

- Suche direkt in Forschungsdatenrepositorien,
- mittels einer Metasuchmaschine[86], z. B. B2FIND Datensuche[87], BASE[88], DataCite Metadata Search[89] oder Google Dataset Search[90] sowie die
- Suche in Data Journals.

4.16.2 Nutzungsbedingungen und Zugriffsrechte

Bereits bei der Publikation von Forschungsdaten können Forschende entscheiden, was die Nachnutzenden mit ihren Daten machen sollen dürfen. Am einfachsten geht dies über die Vergaben von Lizenzen (vgl. Kapitel 4.14.5). So weiß man ganz genau, in welchem Umfang die Nachnutzung zugelassen ist. Neben der Lizenzvergabe kann die Nachnutzung von Forschungsdaten jedoch auch durch Zugriffsrechte reguliert werden (vgl. Kapitel 4.12.3).

4.16.3 Kompatibilität von Lizenzen

Eine Nachnutzung von Daten in Form eines Verweises ist immer möglich, wenn jedoch die Datensätze vermischt und/oder verändert werden sollen, müssen die Lizenzen, unter denen die Daten verfügbar sind, beachtet werden. Nicht alle Lizenzen sind miteinander kompatibel. Die Tabelle 4.7 zeigt die Kompatibilität von CC-Lizenzen auf, für den Fall, dass zwei (oder mehr) Datensätze mit einander kombiniert (vermischt) werden sollen.

Abbildung 4.19 nach Wheeler (2012) zeigt die Kompatibilitäten der gängigen Softwarelizenzen. Die Pfeilrichtung zeigt an, welche Lizenzen miteinander kombiniert werden dürfen. Die Pfeilspitze weist dabei auf die Lizenz hin, die nach der Kombination der Quellcodes, vergeben werden muss.

[86] Eine Metasuchmaschine ist eine Suchmaschine, die mit nur einer Abfrage in mehreren Suchmaschinen suchen kann.
[87] http://b2find.eudat.eu/. Archivierte Version: https://perma.cc/F52D-BEWF.
[88] https://www.base-search.net/about/de/. Archivierte Version: https://perma.cc/UVG5-WZLQ.
[89] https://search.datacite.org/. Archivierte Version: https://perma.cc/L7QZ-D9CJ.
[90] https://datasetsearch.research.google.com/. Archivierte Version: https://perma.cc/JG7X-3SWF.

Tabelle 4.7: Kompatibilität von Creative-Commons-Lizenzen

	PUBLIC DOMAIN	PUBLIC DOMAIN	BY	BY SA	BY NC	BY ND	BY NC SA	BY NC ND
PUBLIC DOMAIN	✓	✓	✓	✓	✓	×	✓	×
PUBLIC DOMAIN	✓	✓	✓	✓	✓	×	✓	×
BY	✓	✓	✓	✓	✓	×	✓	×
BY SA	✓	✓	✓	✓	×	×	×	×
BY NC	✓	✓	✓	×	✓	×	✓	×
BY ND	×	×	×	×	×	×	×	×
BY NC SA	✓	✓	✓	×	✓	×	✓	×
BY NC ND	×	×	×	×	×	×	×	×

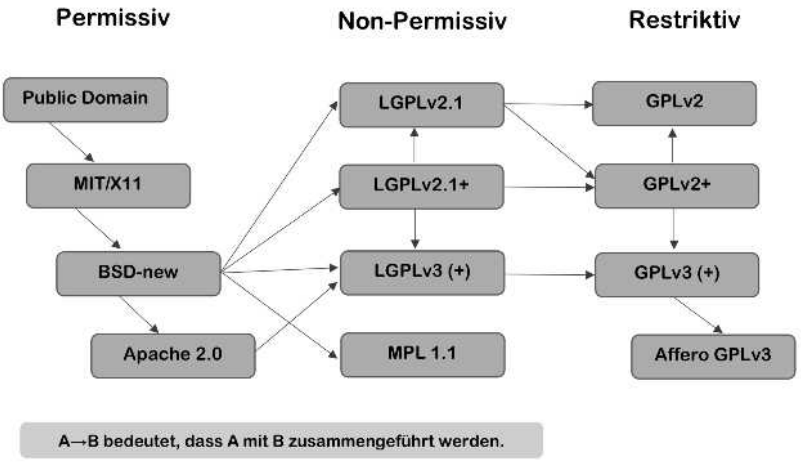


Abbildung 4.19: Kompatibilität von Softwarelizenzen nach Wheeler (2012)

4.16.4 Zitation von Forschungsdaten

Ähnlich wie bei der Nachnutzung von Passagen eines wissenschaftlichen Textes, wird auch bei der Nachnutzung von Forschungsdaten die Quelle genannt und die Urheber:innen/ Ersteller:innen gewürdigt. Lediglich bei der Nachnutzung von Daten unter der Lizenz CC0 oder gemeinfreien Daten, müssen die Ersteller:innen nicht namentlich genannt werden. Die Gute Wissenschaftliche Praxis (vgl. Deutsche Forschungsgemeinschaft, 2019) besagt jedoch, dass man auch in diesem Fall die Arbeit der Ersteller:innen korrekt zitieren sollte.

Grundsätzlich gelten bei der Zitation von Forschungsdaten die gleichen Empfehlungen, wie bei der Zitation von Forschungsartikeln. Zu den wichtigsten Angaben gehören:

- Urheber:innen,
- Veröffentlichungsjahr,
- Titel,
- Publikationsagent (Repositorium oder Archiv),
- Ressourcentyp (Datendomäne),
- Persistenter Identifikator (vorzugszweise als Link),
- ggf. Versionsnummer.

Bei der Nachnutzung von Werken, die mit einer Creative-Commons-Lizenz versehen wurden, sollte darüber hinaus auf die korrekte Attribution der Lizenz geachtet werden. Dies bedeutet, dass die Quellenangabe des Werks die folgenden Angaben beinhalten sollte:

- Quelle,
- Benennung der Lizenz inklusive der Version und Link zur Beschreibung der Lizenz,
- ggf. Bearbeitungsinformationen (ab Version 4.0),
- ggf. Titel des Werkes (ab Version 4.0).

Da Software nicht nur ein wichtiger Bestandteil der Forschung, sondern auch deren Forschungsdatum sein kann, sollte auch diese konsistent und genauso wie andere Forschungsdaten zitiert werden (mit dem Ressourcentyp Software oder Computer Software). Darüber hinaus formulierten Smith et al. (2016) anhand der Datenzitierungsprinzpien der Data Citation Synthesis Group (2014) die Softwarezitierungsprizipien:

- Bedeutung,
- Anerkennung und Namensnennung,
- Eindeutige Identifizierung,
- Persistenz,
- Zugänglichkeit,
- Spezifität.

Beide Zitierungsprinzipien (Daten und Softwarezitierungsprinzipien) ermöglichen sowohl die Menschen- als auch Maschinenlesbarkeit.

4.16.5 Literaturempfehlungen

- Data Citation Synthesis Group. (2017). Joint Declaration of Data Citation Principles. FORCE11. https://doi.org/10.25490/a97f-egyk

- Steward, G. (2017). Licence Compatibility and Transition. [Archivierte Version: https://perma.cc/5NS3-FSDY]. https://docs.google.com/spreadsheets/d/1wUtpMs-FNOQ8Ez2l6ePdEmTgk5kHnr34KinR0YV7hII/edit#gid=888728748

4.16.6 Anwendung auf die Szenarien

Nachnutzung bei Szenario 1

In diesem Szenario werden keine Daten nachgenutzt.

Nachnutzung bei Szenario 2

In dem Open-Source-Projekt wird eine App für eine bereits existierende Software entwickelt. Diese Software wird dabei genutzt und um entsprechende Schnittstellen erweitert. Für die Nachnutzung muss darauf geachtet werden, mit welcher Lizenz die Software lizenziert ist und ob diese eine Nutzung und Veränderung zulässt (vgl. Abbildung 4.19). Eine Nachnutzung der Daten kann zusätzlich in einem separaten Datennutzungsvertrag unter den Projektpartner:innen vereinbart werden.

Nachnutzung bei Szenario 3

In diesem Szenario werden die Daten von Hussein, Sherif (2021), „Twitter Sentiments Dataset“, Mendeley Data, V1, doi: 10.17632/z9zw7nt5h2.1 nachgenutzt.

4.17 Weitere rechtliche Aspekte

Das Forschungsdatenmangement birgt einige rechtliche Tücken, insbesondere wenn die Daten der Öffentlichkeit zur Verfügung gestellt werden sollen. Die häufigsten Rechtsgebiete, die davon betroffen sind, sind Datenschutz und Urheberrecht, die bereits in vorherigen Kapiteln besprochen wurden (vgl. Kapitel 4.8 und 4.14). Darüber hinaus können jedoch viele andere Gebiete eine Rolle spielen (vgl. Abbildung 4.20). In der Informatik stechen vor allem das Patentrecht, Text- und Data-Mining sowie vertragliche Vereinbarungen hervor, auf die nun eingegangen wird.

Patentrecht	Urheberrecht	Wettbewerbsrecht	Datenschutz
Was ist zu beachten, wenn FD Patentreife erlangen (können)?	Unterliegen FD überhaupt dem Urheberrechtsgesetz?	Werden Daten im unternehmerischen Geschäftsverkehr unfair genutzt?	Welche FD sind schützenswert?
Wissenschaftsrecht	**Grundrechte**	**Internationales Recht**	**EU-Recht**
Können Lizenz- und Veröffentlichungsvorgaben für FD per Mandatierung erfolgen?	Welche verfassungsrechtlichen Grenzen sind zu beachten?	Welche Rechtsbestimmungen bestehen außerhalb Deutschlands?	Was bringt z. B. die European Data Economy für FD?
Verträge	**Arbeits-/Dienstrecht**	**Förderbedingungen**	**Policys**
Bestehen Absprachen zum geistigen Eigentum an FD?	Wem gehören die an Hochschulen erhobenen FD?	Welche Bedingungen geben Förderer (DFG; Industrie) vor?	Welche rechtliche Bindung können Policys entfalten?

Abbildung 4.20: Landkarte „Terra incognita – digitale Forschungsdaten auf der Suche nach einer rechtlichen Heimat“ nach Hartmann, 2018

4.17.1 Patentrecht

Patent

Das Patent dient dem Schutz technischer Innovationen. Es handelt sich dabei um ein von einer staatlichen Behörde gewährtes Schutzrecht, welches den Inhaber:innen des Patents das Recht verleiht, die Erfindung zu nutzen und andere von der Nutzung für gewerbliche Zwecke auszuschließen (Götting et al., 2014).

Nach § 1 Abs. 1 Patentgesetz (PatG) können alle Erfindungen auf dem Gebiet „der Technik patentiert werden, sofern sie neu sind, auf einer erfinderischen Tätigkeit beruhen und gewerblich anwendbar sind“. Darüber hinaus können auch Erzeugnisse, die aus biologischem Material sind oder dieses enthalten, patentiert werden. „Biologisches Material, das mithilfe eines technischen Verfahrens aus seiner natürlichen Umgebung isoliert oder hergestellt wird, kann auch dann Gegenstand einer Erfindung sein, wenn es in der Natur schon vorhanden war“. Als Erfindungen gelten jedoch nicht:

- „Entdeckungen sowie wissenschaftliche Theorien und mathematische Methoden;
- ästhetische Formschöpfungen;
- Pläne, Regeln und Verfahren für gedankliche Tätigkeiten, für Spiele oder für geschäftliche Tätigkeiten sowie Programme für Datenverarbeitungsanlagen;
- die Wiedergabe von Informationen“.

Computerprogramme als solche sind im Gegensatz zu sog. „computerimplementierten Erfindungen“ oder Softwarepatenten (vgl. § 1 Abs. 3 Nr. 3 PatG) nicht patentfähig (Baumann et al., 2021).

Eine Erfindung gilt als neu, wenn sie nicht zum Stand der Technik gehört. Dieser umfasst alle Kenntnisse, die vor dem für den Zeitpunkt der Anmeldung maßgeblichen Tag

durch schriftliche oder mündliche Beschreibung, durch Benutzung oder in sonstiger Weise der Öffentlichkeit zugänglich gemacht worden sind. Demnach muss dem Stand der Technik etwas bisher Unbekanntes hinzugefügt werden.

Laut § 6 PatG haben die Erfinder:innen oder die Rechtsnachfolger:innen das Recht auf das Patent. „Haben mehrere gemeinsam eine Erfindung gemacht, so steht ihnen das Recht darauf gemeinschaftlich zu. Haben mehrere die Erfindung unabhängig voneinander gemacht, so steht das Recht dem zu, der die Erfindung zuerst beim Deutschen Patent- und Markenamt angemeldet hat.“.

Diensterfindung

Ein besonderer Fall liegt jedoch vor, wenn die Erfindung im Rahmen der Dienstpflichten gemacht wurde. Die Arbeitnehmer:innen bleiben zwar Erfinder:innen, jedoch können unter bestimmten Voraussetzungen die Rechte an der Erfindung der Arbeitgeber:innen zugewiesen werden (Arbeitnehmererfindergesetz (ArbnErfG)). Es muss sich dabei um eine sog. *Diensterfindung* handeln (§ 4 Abs. 2 ArbnErfG), d. h., dass sie während der Dauer des Arbeitsverhältnisses aus der Tätigkeit im Betrieb entstanden sein muss oder maßgeblich auf Erfahrungen oder Arbeiten des Betriebs beruhen (Baumann et al., 2021). Es ist jedoch irrelevant ob die Diensterfindung im Büro gemacht wurde, sie könne auch von zu Hause oder während des Urlaubs gemacht worden sein. Wurde eine Diensterfindung gemacht, ist sie nach § 5 Abs. 1 ArbnErfG unverzüglich den Arbeitgeber:innen zu melden. Die Arbeitgeber:innen können diese durch Erklärung gegenüber den Arbeitnehmer:innen in Anspruch nehmen, müssen es jedoch nicht. Nehmen sie sie in Anspruch, gehen alle vermögenswerten Rechte an der Diensterfindung auf die Arbeitgeber:innen über (§ 7 Abs. 1 ArbnErfG). Daraufhin müssen die Arbeitgeber:innen die Erfindung zum Patent anmelden. Diese Verpflichtung entfällt nur, wenn berechtigte Belange des Betriebs es erforderlich machen.

Haben die Arbeitnehmer:innen während des Dienstverhältnisses eine freie Erfindung gemacht, müssen sie dies nach § 18 Abs. 1 ArbnErfG den Arbeitgeber:innen unverzüglich durch Erklärung in Textform mitteilen. Dabei muss über die Erfindung und, wenn dies erforderlich ist, auch über ihre Entstehung so viel mitgeteilt werden, dass die Arbeitgeber:innen beurteilen können, ob die Erfindung frei ist.

Wissenschaftsfreiheit und Patentrecht

Für Wissenschaftler:innen an Hochschulen (Hochschullehrerende, wissenschaftliche Mitarbeitende, Assistent:innen sowie studentische Hilfskräfte; jedoch nicht Studierende, externe Promovierende und Honararprofessor:innen) ergeben sich aus § 42 ArbnErfG einige Besonderheiten, die dem Schutz der Wissenschaftsfreiheit aus Art. 5 Abs. 3 des Grundgesetzes gelten (Baumann et al., 2021). Diese beziehen sich insbesondere auf die Veröffentlichung eigener Forschungsergebnisse. Die Erfinder:innen sind berechtigt, die Diensterfindung im Rahmen der eigenen Lehr- und Forschungstätigkeit zu offenbaren, wenn dies dem Dienstherrn rechtzeitig, in der Regel zwei Monate zuvor, angezeigt wurde (§ 42 Nr.1 ArbnErfG).

Darüber hinaus bleibt nach § 42 Nr. 3 ArbnErfG den Erfinder:innen „im Fall der Inanspruchnahme der Diensterfindung ein nichtausschließliches Recht zur Benutzung der Diensterfindung im Rahmen der eigenen Lehr- und Forschungstätigkeit". § 42 Nr. 4 ArbnErfG hingegen legt für Hochschulerfinder:innen eine gesetzliche Vergütung von 30 % der durch die Verwertung der Erfindung erzielten Einnahmen fest.

4.17.2 Vertragliche Vereinbarungen

Wurden Forschungsdaten auf Grundlage vertraglicher Vereinbarungen zur Verfügung gestellt, sind diese Vorgaben einzuhalten. Eine vertragliche Vereinbarung liegt auch dann vor, wenn kein schriftliches Dokument unterzeichnet wurde, sondern auch insbesondere dann, wenn die Daten von einem Repositorium oder andere Einrichtungen bzw. Institutionen zum freien Download im Internet zur Verfügung gestellt werden (Baumann et al., 2021). Die Bedingungen der Nutzung können dabei für Einzelfälle frei festgelegt werden. Dazu gehören Beschränkungen auf nicht-kommerzielle Zwecke, Geheimhaltungsabreden, Quellenangaben oder Änderungsversuche.

4.17.3 Text- und Data-Mining

Text- und Data-Mining (TDM)[91] – also die automatisierte Auswertung von großen Datenmengen – spielen in der Forschung eine immer größere Rolle. Das Urheberrecht bildet mit § 44b UrhG-E und § 60d UrhG-E hierfür die rechtliche Grundlage.

Abbildung 4.21 zeigt auf, welche technischen Vorgänge laut Gesetz möglich sind. Demnach darf das Ursprungsmaterial kopiert und vervielfältigt werden, um daraus durch Normalisierung, Strukturierung, Kategorisierung oder andere Aufbereitungsmethoden ein *Korpus* zu erzeugen. Innerhalb einer Forschungsgruppe – also einem abgrenzbaren Personenkreis – kann dieses *Korpus* frei zugänglich gemacht werden. Die Forschenden müssen dabei nicht an der gleichen Einrichtung tätig sein. Auch im Rahmen von Begutachtungs- bzw. Review-Prozessen darf das *Korpus* verfügbar gemacht werden. Das *Korpus* selbst darf desweiteren auch automatisch ausgewertet werden, da es urheberrechtlich irrelevant ist.

Alle Vervielfältigungen der Ursprungsdaten müssen gelöscht werden, sobald eine Speicherung zum Zweck des TDM nicht länger erforderlich ist, da die dauerhafte Aufbewahrung dieser Daten nicht erlaubt ist. Um die Referenzier- und spätere Überprüfbarkeit der Ergebnisse der Forschung trotzdem zu gewährleisten, dürfen nach § 60d UrhG-E das *Korpus* und die Vervielfältigungen des Ursprungsmaterials an eine Bibliothek, Archiv, Museum oder andere öffentlich zugängliche Bildungseinrichtungen zur Archivierung übermittelt werden. Bei allen anderen (nicht-wissenschaftlichen) Zwecken ist die Archivierung erlaubnispflichtig.
Bevor Daten automatisiert analysiert werden bzw. Datensammlungen zu diesem Zweck angelegt werden, sollte geklärt werden, ob eine solche Nutzung zulässig ist. Eine Hilfestellung bietet dafür die Abbildung 4.22. Dabei wird im ersten Schritt geprüft, ob die Daten dem Urheberrecht unterliegen. Ist dies nicht der Fall, kann TDM problemlos durchgeführt werden. Sind die Daten jedoch urheberrechtlich geschützt (dabei kann auch die Datenbank, aus der

[91] Dieses Kapitel basiert auf den Informationen von forschungsdaten.info (2021).

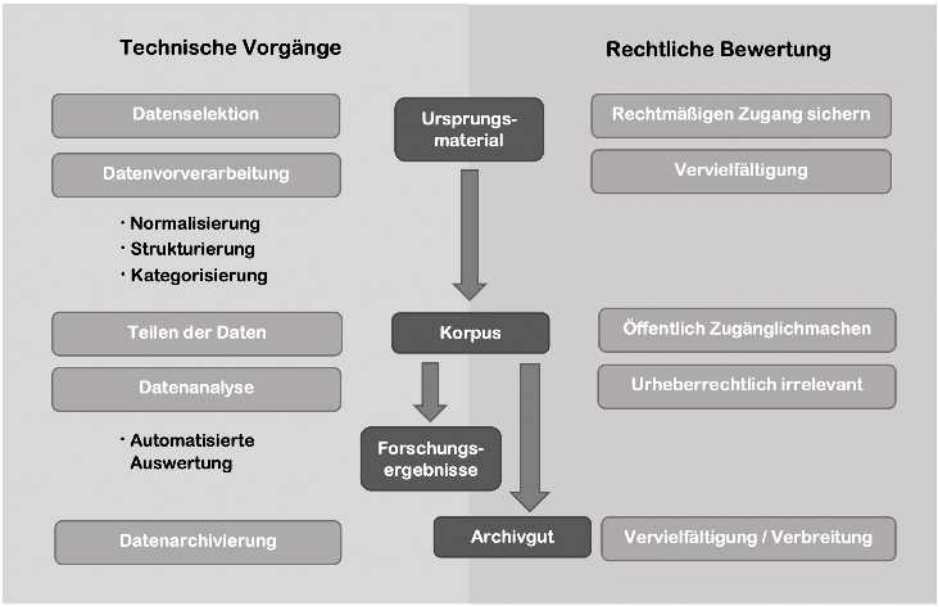


Abbildung 4.21: „Text- und Data-Mining nach dem Verständnis des Gesetzgebers“ nach Brettschneider (2021b)

diese entnommen werden sollen, urheberrechtlich geschützt sein), muss der Zugang zu den Daten überprüft werden. Der § 60d UrhG gewährt diesen nicht automatisch, sondern setzt ihn voraus.

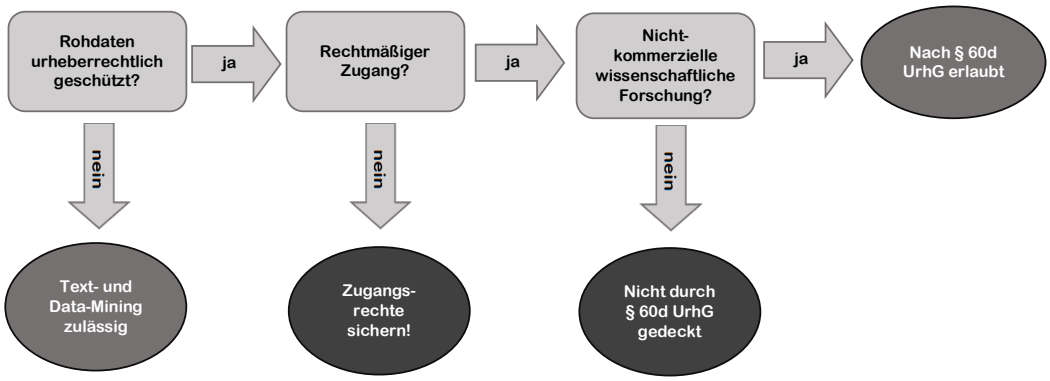


Abbildung 4.22: Checkliste: „Ist Text- und Data-Mining erlaubt?“ nach Brettschneider (2021a)

Mit der Umsetzung der Richtlinie (EU) 2019/790 in 2021 ist TDM nur zulässig, wenn die Rechtsinhaber:innen sich diese nicht vorbehalten haben. Ein Nutzungsvorbehalt bei online zugänglichen Werken ist nur dann wirksam, wenn er in maschinenlesbarer Form erfolgt. Anders als bei § 44b UrhG-E kann die gesetzliche Privilegierung für wissenschaftliches TDM nicht durch vertragliche Abrede beschnitten werden. Weiterhin besteht bei der Vervielfältigung zum Zweck des TDM nicht länger ein Vergütungsanspruch.

4.17.4 Zusammenarbeit mit Schule/Schulbehörde

Bei der Datenerhebung an Schulen sind besondere Rahmenbedingungen hinsichtlich der rechtlichen Lage zu beachten. Die Vorgaben hierzu sind jedoch stark vom jeweiligen Bundesland abhängig und unterscheiden sich hinsichtlich ihrer Vorgaben zum Genehmigungsverfahren, Datenschutz, Einwilligungserklärungen, etc. Einen Überblick über bundeslandsspezifische Bestimmungen können beim Verbund Forschungsdaten Bildung[92] nachgelesen werden. Ein exemplarischer Abriss wird für das Land Hamburg dargestellt. Hier ist zunächst die Schulbehörde zu kontaktiert, wofür entsprechende Formulare auf der Webseite der Behörde für Schule und Berufsbildung[93] zu finden sind. Neben der inhaltlichen Darstellung der Untersuchung und des Forschungsdesigns müssen zu diesem Zeitpunkt auch exemplarische Dokumente eingereicht werden, wie die Einwilligungserklärung, Testinstrumente, etc. Dieses Verfahren soll mindestens drei Monate vor der geplanten Untersuchung begonnen werden. Erst nachdem die Genehmigung der Behörde vorliegt, dürfen Schulen kontaktiert werden. Es muss eine Zustimmung der Schulleitung eingeholt und die Schulkonferenz sollte informiert werden. Eine Ausnahme bilden hier jedoch Studien, die im Rahmen des Lehramtsausbildung von Studierenden durchgeführt werden. Sofern die Studien nur an bis zu zwei Schulen durchgeführt werden, muss keine Genehmigung der Behörde für Schule und Berufsausbildung eingeholt, sondern es kann direkt an die Schulleitung herangetreten werden.

Bei der Arbeit mit Minderjährigen sind weitere Aspekte in besonderer Form zu berücksichtigen:

1. Bei der informierten Einwilligungserklärung muss darauf geachtet werden, welche Personen bei welchem Alter der Schüler:innen schriftlich zustimmen müssen.
2. Die informierte Einwilligungserklärung muss sprachlich für die Schüler:innen verständlich und nachvollziehbar sein.
3. Forschungsethisch muss berücksichtigt werden, dass die Schüler:innen gegebenenfalls die langfristigen Folgen bei der Publikation von Daten nicht abschätzen können. Der Schutz der Schüler:innen muss deshalb an erster Stelle stehen.

4.17.5 Literaturempfehlungen

- Baumann, P., Krahn, P. & Lauber-Rönsberg, A. (2021). *Forschungsdatenmanagement und Recht. Datenschutz-, Urheber- und Vertragsrecht.* W. Neugebauer.
- Brettschneider, P. (2021). Text und Data-Mining – juristische Fallstricke und bibliothekarische Handlungsfelder. *Bibliotheksdienst, 55*(2): 104–126. https://doi.org/10.1515/bd-2021-0020

[92] https://www.forschungsdaten-bildung.de/genehmigungen. Archivierte Version: https://perma.cc/R6JX-ULRK.

[93] https://www.hamburg.de/bsb/bq-f/4361582/genehmigungsverfahren/. Archivierte Version: https://perma.cc/9LBX-3TH6.

4.17.6 Anwendung auf die Szenarien

Weitere rechtliche Aspekte bei Szenario 1

Carla ist als private Person die Eigentümerin der Daten, da sie die Daten im Rahmen ihrer Bachelorarbeit erhoben hat.

Weitere rechtliche Aspekte bei Szenario 2

Timo erhebt die Daten im Rahmen seiner Masterarbeit und hat kein arbeitsrechtliches Verhältnis im Rahmen des Open-Source-Projekts. Somit ist er Eigentümer der Daten.

Weitere rechtliche Aspekte bei Szenario 3

Neben der DSGVO müssen die rechtlichen Rahmenbedingungen von Twitter beachtet werden. Diese umfassen drei Dokumente: Terms of Service[a], Developer Policy[b] und das Developer Agreement[c]. Aus diesen Dokumenten lässt sich ableiten, dass die Datennutzung den wirtschaftlichen Interessen Twitters nicht schaden darf. Eine Erstellung, Anreicherung und Verbreitung von großen Datenbanken mit Tweets ist grundsätzlich untersagt (Beurskens, 2013). Dies führt zu Verwirrung, denn einerseits werden die Tweets von den Nutzer:innen selbst der breiten Öffentlichkeit zu Verfügung gestellt, sie können nicht nur über die interne Suche, sondern auch über Google gefunden und somit von jeder Person gelesen werden. Andererseits jedoch behält sich Twitter die Rechte an den Tweets vor. Beurskens (2013) empfiehlt daher, selbsterstellte Sammlungen nicht zu publizieren. Darüber hinaus ist Alex an der Universität angestellt und agiert im Auftrag des Arbeitgebers, somit ist die Universität Eigentümerin der Daten.

[a] https://twitter.com/en/tos. Archivierte Version: https://perma.cc/A5MB-QCWP.

[b] https://developer.twitter.com/en/developer-terms/policy. Archivierte Version: https://perma.cc/Y648-684Y.

[c] https://developer.twitter.com/en/developer-terms/agreement. Archivierte Version: https://perma.cc/C665-BKUT.

Teil III

Lehrmaterial

Kapitel 5

Lehrmaterial

Nachdem in den vorherigen Kapiteln die Notwendigkeit von FDM im Studium veranschaulicht wurde und wichtige inhaltliche Aspekte detailliert beschrieben wurden, wird nun konkretes Lehrmaterial zur Verfügung gestellt. Ähnlich wie in Kapitel 4.2 werden die Lehrmaterialien an FDM-Themen entlang strukturiert. Anhand der Überschriften von Arbeitsblättern ist zu erkennen, zu welchem Unterthema von FDM sie zuzuordnen sind. Direkt im Anschluss ist die Musterlösung für das jeweilige Arbeitsblatt zu finden. Zu einigen Themen werden auch zusätzliche Handreichungen in Form von Checklisten bereitgestellt. Diese Checklisten können im Rahmen von Lehrveranstaltungen an Studierende als Lehrmaterial verteilt werden.

Die gesammelten Materialien stehen ebenfalls als Download unter der DOI 10.5281/zenodo.6512432 zur Verfügung. In dieser Version ist entsprechend keine Kopfzeile enthalten.

5.1 Forschungsdaten(-management) in der Informatik

So wie bereits in Kapitel 4.1 und 4.2 zunächst Forschungsdaten und FDM im Allgemeinen beschrieben und dann auf die Informatik angewandt wurden, ist auch bei den Lehrmaterialien ein allgemeiner Teil enthalten, der dann hinsichtlich der Informatik konkretisiert wird. Tabelle 5.1 gibt einen Überblick über das Material.

Tabelle 5.1: Lehrmaterial Forschungsdaten und Forschungsdatenmanagement in der Informatik

Arbeitsblatt: Forschungsdaten und Forschungsdatenmanagement

Carla hat sich bisher noch nie mit Forschungsdatenmanagement beschäftigt, bekommt jedoch von ihrer Betreuerin der Bachelorarbeit die Aufgabe, ihre Forschungsdaten klar zu benennen und die Daten nach der Bachelorarbeit zu publizieren.

1. Definieren Sie den Begriff Forschungsdaten.

   .....

2. Mit welchen Forschungsdaten arbeiten Sie?

   .....

3. Nennen Sie Vorteile des Forschungsdatenmanagements für die Wissenschaft.

   .....

4. Warum ist es wichtig, bereits bei der Bachelorarbeit gutes Forschungsdatenmanagement zu betreiben?

   .....

   .....

   .....

   .....

Musterlösung: Forschungsdaten und Forschungsdatenmanagement

1. Definieren Sie den Begriff Forschungsdaten.

   Es gibt nicht *eine* feste Definition von Forschungsdaten. In diesem Buch wird auf der Definition von Kindling und Schirmbacher (2013)[1] gearbeitet, die Folgendes besagt: „Unter digitalen Forschungsdaten verstehen wir [...] alle digital vorliegenden Daten, die während des Forschungsprozesses entstehen oder ihr Ergebnis sind".

2. Mit welchen Forschungsdaten arbeiten Sie?

   Für diese Frage gibt es keine Musterlösung, da die Antwort individuell in Abhängigkeit zur eigenen Forschung steht.

3. Nennen Sie Vorteile des Forschungsdatenmanagements für die Wissenschaft.

   - schnellere Auffindbarkeit von Daten, z. B. durch aussagekräftige Benennung
   - Übersichtlichkeit, z. B. keine verstreute Ablage von Daten in unterschiedlichen Versionen auf verschiedenen Rechnern
   - Wissenserhalt -– Daten sind unabhängig von einzelnen Menschen, Projekten oder Institutionen zugänglich
   - Transfer der Daten in künftige Projekte
   - Erleichterung der Zusammenarbeit
   - langfristige Nachvollziehbarkeit von Ergebnissen statt neue Erzeugung (Erhalt von Primär- und Sekundärdaten)
   - beugt Datenverlust vor, z. B. wegen defekter Hard- oder Software oder von Ursprungsversionen der Dateien
   - (halb-)automatische Verarbeitung wird durch Metadaten ermöglicht
   - Weitergabe und Nachnutzung von Daten durch Verwendung von entsprechend formulierten Einwilligungserklärungen, z. B. kein Passus, dass Daten nach Ablauf des Projektes gelöscht werden
   - optimierter Mitteleinsatz, z. B. Kostenersparnis durch Nachnutzung statt neuer Erhebung
   - Erfüllung von Auflagen der Drittmittelgeber
   - Forschungsdatenzitation
   - Referenzierbarkeit
   - Erhöhung der Relevanz der eigenen Arbeit durch bessere Sichtbarkeit

[1] Kindling, M. & Schirmbacher, P. (2013). Die digitale Forschungswelt als Gegenstand der Forschung / Research on Digital Research / Recherche dans la domaine de la recherche numerique. *Information - Wissenschaft & Praxis, 64*(2-3), 127–136. https://doi.org/10.1515/iwp-2013-0017

4. Warum ist es wichtig, bereits bei der Bachelorarbeit gutes Forschungsdatenmanagement zu betreiben?

   Die Bachelorarbeit stellt schon eine erste wissenschaftliche Arbeit dar, die auch nach den Prinzipien Guter Wissenschaftlicher Praxis ausgeführt werden soll. Insbesondere bei der Arbeit mit sensiblen und personenbezogenen Daten ist es wichtig, dass im Forschungsdatenmanagement enthaltene Aspekte (wie informierte Einwilligungserklärungen, Sicherung der Daten etc.) akkurat umgesetzt werden. Auch die Nachnutzung der erhobenen Daten kann für andere Wissenschaftler:innen relevant sein und somit Ressourcen sparen.

Arbeitsblatt: Softwarelebenszyklus

Timo möchte sich an einem Open-Source-Projekt beteiligen. Vorher beschäftigt er sich mit dem Softwarelebenszyklus, um den Prozess im Vorfeld zu verstehen. Er hat ebenfalls erfahren, dass Software auch als Forschungsdatum kategorisiert wird.

1. Begründen Sie, in welchen Fällen Software ein Forschungsdatum sein kann.

   .....

   .....

   .....

   .....

2. Geben Sie zu dem abgebildeten Forschungsdatenlebenszyklus an, wo Sie Überschneidungen zum Softwarelebenszyklus finden können.

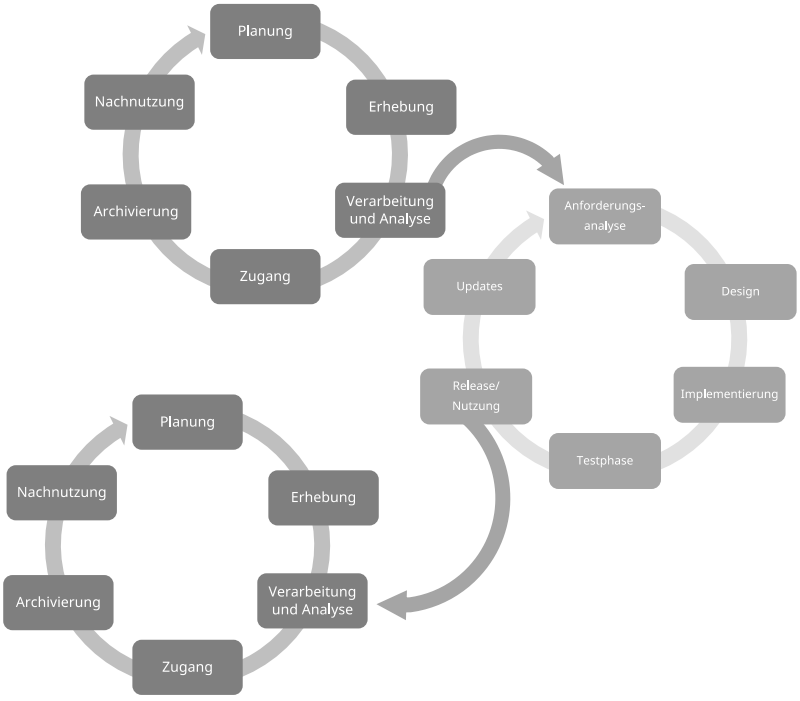
Musterlösung: Softwarelebenszyklus

Timo möchte sich an einem Open-Source-Projekt beteiligen. Vorher beschäftigt er sich mit dem Softwarelebenszyklus, um den Prozess im Vorfeld zu verstehen. Er hat ebenfalls erfahren, dass Software auch als Forschungsdatum kategorisiert wird.

1. Begründen Sie, in welchen Fällen Software ein Forschungsdatum sein kann.

   Bei Software kann es sich um ein Forschungsdatum handeln, wenn es ein eigenständiges Forschungsergebnis ist, das als effektive algorithmische Lösung eines Problems dient oder wenn es selbst ein Forschungsgegenstand ist.

2. Geben Sie zu dem abgebildeten Forschungsdatenlebenszyklus an, wo Sie Überschneidungen zum Softwarelebenszyklus finden können.



5.2 Forschungsdaten-Policys

In Kapitel 4.3 werden verschiedene Facetten von Forschungsdaten-Policys vorgestellt. Im Rahmen des Lehrmaterials wird die Anwendung von Forschungsdaten-Policys eines konkreten Beispiels vertieft. Tabelle 5.2 listet das Lehrmaterial im Überblick auf.

Tabelle 5.2: Lehrmaterial Forschungsdaten-Policys

Thema	**Seite**
Arbeitsblatt: Forschungsdaten-Policy	171
Musterlösung: Forschungsdaten-Policy	172

Arbeitsblatt: Forschungsdaten-Policy

Alex arbeitet in einem BMBF-Projekt und wird im Rahmen des Projekts einen Algorithmus entwickeln. Bei der Beschäftigung mit dem Softwarelebenszyklus fällt auf, dass Software auch als Forschungsdatum kategorisiert wird und Alex sich mit Forschungsdaten-Policys beschäftigen sollte.

1. Lesen Sie den angeführten Ausschnitt der Forschungsdaten-Policy[1]

> 1. Forschende HU-Angehörige sind verpflichtet, die Forschungsdaten sicher zu speichern, angemessen aufzubereiten und zu dokumentieren sowie langfristig aufzubewahren. Die Verantwortung für die Gewährleistung dieser Prozesse liegt bei den HU-Angehörigen, die das Forschungsvorhaben leiten.
> 2. Alle forschenden HU-Angehörigen sind aufgefordert, die in ihrer wissenschaftlichen Tätigkeit entstehenden Forschungsdaten gemäß den im jeweiligen Fachgebiet etablierten Regelungen bzw. Standards aufzubereiten. Sie dokumentieren den gesamten Forschungszyklus sowie die verwendeten Werkzeuge und Verfahren.
>
> 3. Es liegt in eigener Verantwortung der forschenden HU-Angehörigen, zu welchem Zeitpunkt und zu welchen rechtlichen Bedingungen Forschungsdaten zugänglich gemacht werden. Die Humboldt-Universität empfiehlt, Forschungsdaten ebenso wie die wissenschaftliche Publikation gemäß der Open-Access-Erklärung der HU frühestmöglich öffentlich zugänglich zu machen. Der Schutz personenbezogener Daten, des Urheberrechts und der berechtigten Interessen Dritter muss gewährleistet sein.

   (a) Wie sollen Ihre Forschungsdaten gespeichert werden? Wie können Sie dies erreichen?

   .....

   (b) Wie lange sollen Ihre Forschungsdaten aufbewahrt werden?

   .....

   (c) Welche Standards für die Dokumentation Ihrer Forschungsdaten sollen angewandt werden? Seien Sie bei der Nennung möglichst konkret.

   .....

   (d) Wann sollen Ihre Forschungsdaten publiziert werden? Was genau bedeutet dies für Ihre Forschungsdaten?

   .....

[1] Humboldt-Universität zu Berlin. (2014). Grundsätze zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin. [Archivierte Version: https://perma.cc/388W-9SMK]. https://www.cms.hu-berlin.de/de/dl/dataman/infos/policy

Musterlösung: Forschungsdaten-Policy

Alex arbeitet in einem BMBF-Projekt und wird im Rahmen des Projekts einen Algorithmus entwickeln. Bei der Beschäftigung mit dem Softwarelebenszyklus fällt auf, dass Software auch als Forschungsdatum kategorisiert wird und Alex sich mit Forschungsdaten-Policys beschäftigen sollte.

1. Lesen Sie den angeführten Ausschnitt der Forschungsdaten-Policy[1]

> 1. Forschende HU-Angehörige sind verpflichtet, die Forschungsdaten sicher zu speichern, angemessen aufzubereiten und zu dokumentieren sowie langfristig aufzubewahren. Die Verantwortung für die Gewährleistung dieser Prozesse liegt bei den HU-Angehörigen, die das Forschungsvorhaben leiten.
> 2. Alle forschenden HU-Angehörigen sind aufgefordert, die in ihrer wissenschaftlichen Tätigkeit entstehenden Forschungsdaten gemäß den im jeweiligen Fachgebiet etablierten Regelungen bzw. Standards aufzubereiten. Sie dokumentieren den gesamten Forschungszyklus sowie die verwendeten Werkzeuge und Verfahren.
>
> 3. Es liegt in eigener Verantwortung der forschenden HU-Angehörigen, zu welchem Zeitpunkt und zu welchen rechtlichen Bedingungen Forschungsdaten zugänglich gemacht werden. Die Humboldt-Universität empfiehlt, Forschungsdaten ebenso wie die wissenschaftliche Publikation gemäß der Open-Access-Erklärung der HU frühestmöglich öffentlich zugänglich zu machen. Der Schutz personenbezogener Daten, des Urheberrechts und der berechtigten Interessen Dritter muss gewährleistet sein.

   (a) Wie sollen Ihre Forschungsdaten gespeichert werden? Wie können Sie dies erreichen?

   Die Forschungsdaten sollen sicher und langfristig aufbewahrt werden. Um die Sicherheit der Daten zu gewährleisten, sollte ein regelmäßiges Back-up der Daten vorgenommen werden. Dieses sowie die langfristige Aufbewahrung wird von dem Back-up-Service des Computer- und Medienservices der Humboldt-Universität zu Berlin durchgeführt.

   (b) Wie lange sollen Ihre Forschungsdaten aufbewahrt werden?

   Die Daten sollen langfristig aufbewahrt werden. Laut Guter Wissenschaftlicher Praxis der DFG beträgt diese Zeit mindestens zehn Jahre nach der Publikation der Ergebnisse. Das Archiv des Computer- und Medienservice bietet hierbei drei Optionen: 15 Jahre (lang), 8 Jahre (medium) und 5 Jahre (Standard).

[1] Humboldt-Universität zu Berlin. (2014). Grundsätze zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin. [Archivierte Version: https://perma.cc/388W-9SMK]. https://www.cms.hu-berlin.de/de/dl/dataman/infos/policy

(c) Welche Standards für die Dokumentation Ihrer Forschungsdaten sollen angewandt werden? Seien Sie bei der Nennung möglichst konkret.

Es sollen die im jeweiligen Fachgebiet etablierten Regelungen bzw. Standards verwendet werden. In Learning Analytics gibt es noch keine etablierten Standards. Eine Lösung wäre, den Metadatenstandard LOM und Dublic Core zu kombinieren.

(d) Wann sollen Ihre Forschungsdaten publiziert werden? Was genau bedeutet dies für Ihre Forschugnsdaten?

Die Forschungsdaten sollen möglichst früh publiziert werden. In Alex' Fall können nicht alle Daten zugänglich gemacht werden, da dies den Nutzungsrechten Twitters widersprechen würde. Es wird daher nur die anonymisirte Twitter-Datensammlung direkt nach der vollendeten Analyse publiziert.

5.3 Institutionelle Infrastruktur

In Kapitel 4.4 wurde das Thema institutionelle Infrastruktur behandelt, welches im engen Zusammenhang mit FDM steht. Im Folgenden ist ein Arbeitsblatt zu finden, dass ein Selbststudium zur institutionellen Infrastruktur an einer beliebigen Universität vorstrukturiert. Tabelle 5.3 listet das entsprechende Material auf.

Tabelle 5.3: Lehrmaterial Institutionelle Infrastruktur

Arbeitsblatt: Institutionelle Infrastruktur

Alex beschäftigt sich im Rahmen eines BMBF-Projekts mit dem Thema Forschungsdatenmanagement. Alex hat bereits davon gehört, dass es an der Einrichtung eine spezielle Anlaufstelle für Forschungsdatenmanagement gibt. Um sich mit der institutionellen Infrastruktur auseinanderzusetzen, hat Alex die folgenden Leitfragen bekommen.

1. Nennen Sie Anlaufstellen Ihrer Einrichtung, an die Sie sich bei allgemeinen Fragen zum Forschungsdatenmanagement wenden können.

   .....

   .....

   .....

2. Geben Sie Ihne Ansprechpartner:innen für rechtliche Fragen zum Thema Forschungsdatenmanagement an.

   .....

   .....

   .....

3. Nennen Sie Weiterbildungsmöglichkeiten an Ihrer Einrichtung zum Thema Forschungsdatenmanagement.

   .....

   .....

   .....

4. Beschreiben Sie die Funktionsweise des Back-ups an Ihrer Einrichtung.

   .....

   .....

   .....

5. Nennen Sie von Ihrer Einrichtung angebotene Werkzeuge für das kollaborative Arbeiten.

   .....

   .....

   .....

6. Gibt es an Ihrer Einrichtung ein Forschungsdatenrepositorium? Welches?

   .....

   .....

   .....

7. Nennen Sie Speichermöglichkeiten Ihrer Einrichtung für das Forschungsdatenmanagement.

   .....

   .....

   .....

Musterlösung: Institutionelle Infrastruktur

Alex beschäftigt sich im Rahmen eines BMBF-Projekts mit dem Thema Forschungsdatenmanagement. Alex hat bereits davon gehört, dass es an der Einrichtung eine spezielle Anlaufstelle für Forschungsdatenmanagement gibt. Um sich mit der institutionellen Infrastruktur auseinanderzusetzen, hat Alex die folgenden Leitfragen bekommen.

1. Nennen Sie Anlaufstellen Ihrer Einrichtung, an die Sie sich bei allgemeinen Fragen zum Forschungsdatenmanagement wenden können.

   An die Forschungsdatenkoordinator:innen der Humboldt-Universität zu Berlin bzw. an die Ansprechperson der Universitätsbibliothek.[1]

2. Geben Sie Ihne Ansprechpartner:innen für rechtliche Fragen zum Thema Forschungsdatenmanagement an.

   Für die Beantwortung der rechtlichen Fragen zum Thema Forschungsdatenmanagement stehen als erste Anlaufstelle die Forschungsdatenkoordinator:innen der Humboldt-Universität zu Berlin zur Verfügung. Bei schwierigeren Fällen und Einzelfallbetrachtungen sollten die Datenschutzbeauftragten der Einrichtung kontaktiert werden[2].

3. Nennen Sie Weiterbildungsmöglichkeiten an Ihrer Einrichtung zum Thema Forschungsdatenmanagement.

   Die Koordinationsstelle FDM der Humboldt-Universität zu Berlin bietet regelmäßige Schulungen und Informationsveranstlatungen[3] zu diesem Thema an. Auch an der beruflichen Weiterbildung[4] der Humboldt-Universität zu Berlin werden Workshops zum Thema Forschungsdatenmanagement angeboten.

4. Beschreiben Sie die Funktionsweise des Back-ups an Ihrer Einrichtung.

   Das Back-up wird vom Zentralen Back-up-Service des Computer- und Medienservice der Humboldt-Universität zu Berlin durchgeführt. Dieses basiert auf dem Softwareprodukt Spectrum Protect™ der Firma IBM und führt alle 24 Stunden eine inkrementelle Sicherung der Daten durch. Die Back-up-Versionen werden bis zu 60 Tage aufbewahrt (in Abhängigkeit von den Einstellungen).

5. Nennen Sie von Ihrer Einrichtung angebotene Werkzeuge für das kollaborative Arbeiten.

[1] https://www.cms.hu-berlin.de/de/dl/dataman/support/kontakt. Archivierte Version: https://perma.cc/A8ER-TP6M.

[2] https://www.hu-berlin.de/de/datenschutz/kontakt. Archivierte Version: https://perma.cc/WZ8Z-MNLL.

[3] https://www.cms.hu-berlin.de/de/dl/dataman/support/schulungen. Archivierte Version: https://perma.cc/6AW5-UEGP.

[4] https://bwb.hu-berlin.de/. Archivierte Version: https://perma.cc/PYP7-9RKR.

- HU-Box
- Zoom
- OpenProject
- GitLab
- Only Office innerhalb der HU Box
- Moodle
- Big Blue Button

6. Gibt es an Ihrer Einrichtung ein Forschungsdatenrepositorium? Welches?

   Die Humboldt-Universität zu Berlin besitzt ein Forschungsdsatenrepositorium. Dies ist in den Publikationsserver integriert und wird als edoc-Server bezeichnet[5]. Darüber hinaus existiert das sogenannte Medien-Repositorium[6] für (Forschungs-)Daten, die nicht zwingenderweise einer Publikation zugrunde liegen müssen.

7. Nennen Sie Speichermöglichkeiten Ihrer Einrichtung für das Forschungsdatenmanagement.

   Der Computer- und Medienservice der Humboldt-Universität zu Berlin bietet verschiedene Speichermöglichkeiten für Forschungsdaten. Dazu gehören die HU-Box, der Windows Fileservice, das Medien-Repositorium, der Datenbankservice oder auch GitLab.

[5] https://edoc.hu-berlin.de/. Archivierte Version: https://perma.cc/QH78-V4LK.
[6] https://medien.hu-berlin.de. Archivierte Version: https://perma.cc/G5C6-9KFK.

5.4 FAIR-Prinzipien

Die FAIR-Prinzipien für die Publikation und Nachnutzung von Daten wurden in Kapitel 4.5 beschrieben und es wird auch auf eine spezielle Ausprägung von FAIR-Prinzipien für Software eingegangen. Spezielles Lehrmaterial findet sich in Tabelle 5.4.

Tabelle 5.4: Lehrmaterial FAIR-Prinzipien

Arbeitsblatt: FAIR-Prinzipien

Carla möchte ihre Interviewdaten publizieren und weiß bereits, dass die FAIR-Prinzipien für eine Datenpublikation beachtet werden sollen. Jedoch ist sie sich unsicher, mit welcher ihrer geplanten Maßnahmen welches FAIR-Prinzip adressiert wird.

1. Ordnen Sie die Aussagen einer der vier FAIR-Kategorien zu[1]:

   (a) Die Forschungsdaten haben eine klare und zugängliche Nutzungslizenz.

   - ☐ Findable
   - ☐ Accessible
   - ☐ Interoperable
   - ☐ Reusable

   (b) Den (Meta-)Daten wurde ein global eindeutiger und dauerhaft persistenter Identifikator zugewiesen.

   - ☐ Findable
   - ☐ Accessible
   - ☐ Interoperable
   - ☐ Reusable

   (c) Die Forschungs- und Metadaten verwenden allgemein übliche, zugängliche und vorzugsweise offene Standards und Formate.

   - ☐ Findable
   - ☐ Accessible
   - ☐ Interoperable
   - ☐ Reusable

   (d) Die Metadaten sind und bleiben verfügbar, auch für den Fall, dass die zugehörigen Forschungsdaten nicht mehr vorhanden sind.

   - ☐ Findable
   - ☐ Accessible
   - ☐ Interoperable
   - ☐ Reusable

[1] Quelle der Aufgabe: Asef, E., Biernacka, K., Böker, E., Danker, S. A., Juliane, J., Neumann, J., Petersen, B., Rex, J. & Trautwein-Bruns, U. (2021). Data Sharing interaktiv vermitteln. https://doi.org/10.5281/zenodo.4585487

Musterlösung: FAIR-Prinzipien

Carla möchte ihre Interviewdaten publizieren und weiß bereits, dass die FAIR-Prinzipien für eine Datenpublikation beachtet werden sollen. Jedoch ist sie sich unsicher, mit welcher ihrer geplanten Maßnahmen welches FAIR-Prinzip adressiert wird.

1. Ordnen Sie die Aussagen einer der vier FAIR-Kategorien zu[1]:

   (a) Die Forschungsdaten haben eine klare und zugängliche Nutzungslizenz.

   - ☐ Findable
   - ☐ Accessible
   - ☐ Interoperable
   - ☒ Reusable

   (b) Den (Meta-)Daten wurde ein global eindeutiger und dauerhaft persistenter Identifikator zugewiesen.

   - ☒ Findable
   - ☐ Accessible
   - ☐ Interoperable
   - ☐ Reusable

   (c) Die Forschungs- und Metadaten verwenden allgemein übliche, zugängliche und vorzugsweise offene Standards und Formate.

   - ☐ Findable
   - ☐ Accessible
   - ☒ Interoperable
   - ☐ Reusable

   (d) Die Metadaten sind und bleiben verfügbar, auch für den Fall, dass die zugehörigen Forschungsdaten nicht mehr vorhanden sind.

   - ☐ Findable
   - ☒ Accessible
   - ☐ Interoperable
   - ☐ Reusable

[1] Quelle der Aufgabe: Asef, E., Biernacka, K., Böker, E., Danker, S. A., Juliane, J., Neumann, J., Petersen, B., Rex, J. & Trautwein-Bruns, U. (2021). Data Sharing interaktiv vermitteln. https://doi.org/10.5281/zenodo.4585487

Checkliste: Wie FAIR sind deine Forschungsdaten?

WIE FAIR SIND DEINE FORSCHUNGSDATEN?

FINDABLE

Deine Forschungsdaten und deren Metadaten sollten sowohl von anderen Wissenschaftlern und Wissenschaftlerinnen als auch von Maschinen auffindbar sein. Grundlegende maschinenlesbare beschreibende Metadaten erleichtern das Finden von relevanten Datensätzen.

- ☐ Deinen (Meta)Daten wurde ein global eindeutiger und dauerhaft persistenter Identifier zugewiesen.
- ☐ Deine Forschungsdaten sind mit umfangreichen Metadaten beschrieben.
- ☐ Die Metadaten beinhalten eindeutig und explizit den Identifier der Daten, die sie beschreiben.
- ☐ Die Metadaten sind in einem durchsuchbaren Verzeichnis registriert oder indiziert.

ACCESSIBLE

Es sollte für Menschen und Maschinen möglich sein, auf Deine Forschungsdaten zuzugreifen (gegebenenfalls unter bestimmten Bedingungen oder Einschränkungen). FAIR bedeutet nicht, dass die Daten offen sein müssen! Es sollten aber Metadaten vorhanden sein, auch wenn die Daten nicht zugänglich sind.

- ☐ Deine (Meta)Daten sind über ihren Identifier mithilfe eines standardisierten, offenen und freien Kommunikationsprotokolls auffindbar.
- ☐ Die Metadaten sind und bleiben verfügbar, auch für den Fall, dass die zugehörigen Forschungsdaten nicht mehr vorhanden sind.

INTEROPERABLE

Deine Forschungsdaten und Metadaten sollten anerkannten Formaten und Standards entsprechen, damit sie in einer (teil-)automatisierter Weise kombiniert, ausgetauscht und interpretiert werden können.

- ☐ Deine Forschungsdaten und deren Metadaten verwenden allgemein übliche, zugängliche und vorzugsweise offene Standards und Formate
- ☐ Kontrollierte Vokabulare, Schlüsselwörter, Thesauri oder Ontologien wurden nach Möglichkeit verwendet
- ☐ Verweise auf verwandte (Meta)Daten sind enthalten

REUSABLE

Eine gute Beschreibung Deiner Forschungsdaten und deren Metadaten ermöglicht die Wiederverwendung der Daten für zukünftige Forschung und den Vergleich mit anderen, kompatiblen Datenquellen.

- ☐ Deine Forschungsdaten sind mit einer Vielzahl von genauen und relevanten Attributen beschrieben
- ☐ Deine Forschungsdaten haben eine klare und zugängliche Nutzungslizenz
- ☐ Deine Forschungsdaten enthalten detaillierte Provenienz-Informationen
- ☐ Deine Forschungsdaten und Metadaten entsprechen den relevanten fachspezifischen Standards

Weitere Informationen zu den FAIR-Prinzipien unter: www.go-fair.org/fair-principles

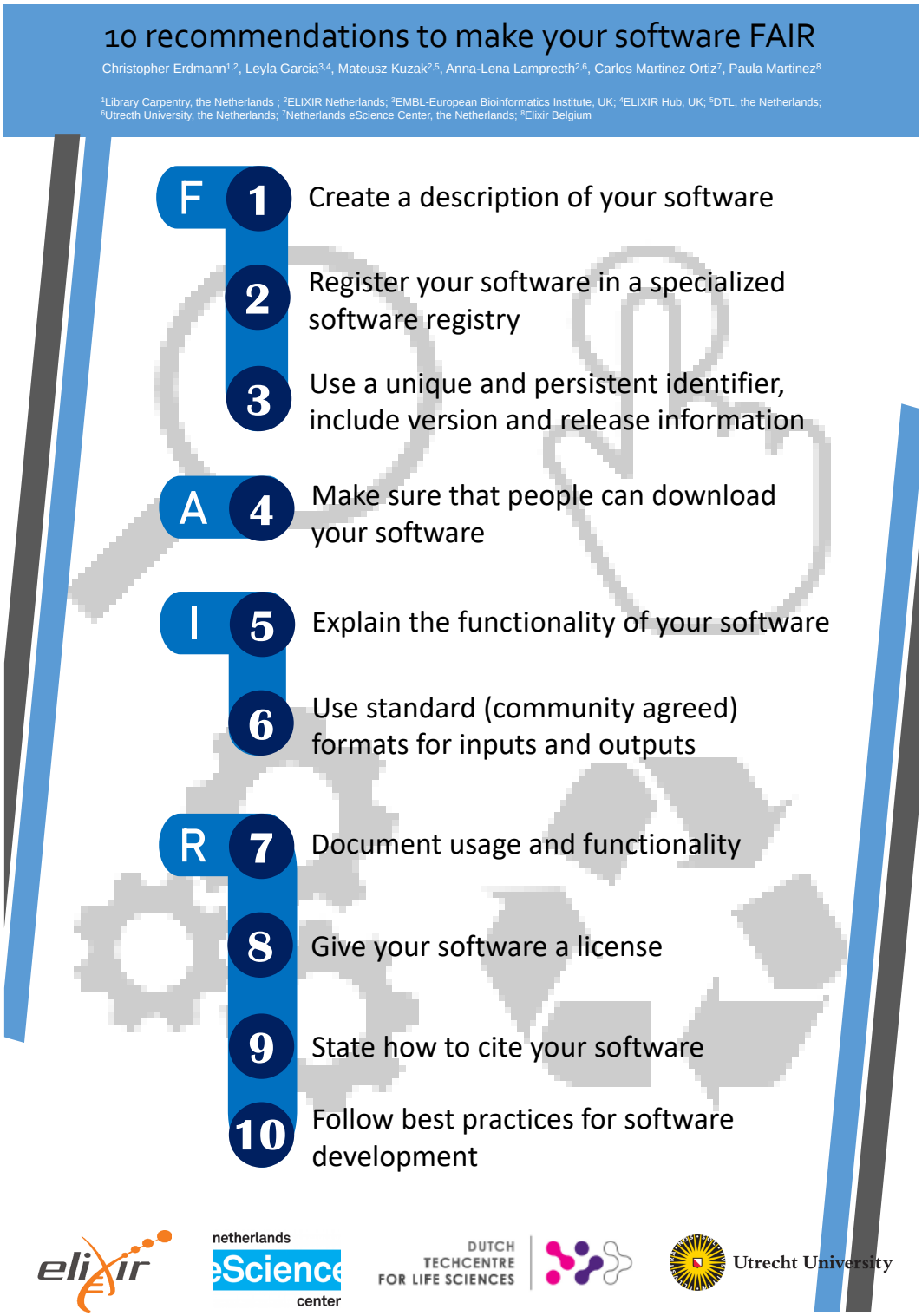
Basierend auf: Jones, Sarah und Marjan Grootveld. How FAIR are your data? Checklist. Vers. 1. 2017. DOI 10.5281/zenodo.1065991. Das Werk ist lizenziert unter der CC-BY-Lizenz.

FDMentor **Erstellt im Rahmen des FDMentor-Projektes**
Projektlaufzeit: 1. Mai 2017 bis 30. April 2019
Idee und Gestaltung: Katarzyna Biernacka, Dr. Dominika Dolzycka, Petra Buchholz, Kerstin Helbig

Kontakt: fdmentor@hu-berlin.de
Twitter: @fd_mentor
https://hu.berlin/fdmentor

gefördert vom: Bundesministerium für Bildung und Forschung

Handreichung: 10 recommendations to make your software FAIR[1]



[1] Quelle: Erdmann, C., Garcia, L., Kuzak, M., Lamprecht, A.-L., Martinez-Ortiz, C. & Martinez, P. (2019). 10 recommendations to make your software FAIR. F1000Research. https://doi.org/10.7490/f1000research.1117402.1

5.5 Datenmanagementplan

Die Wichtigkeit sowie die Inhalte eines Datenmanagementplans wurden in Kapitel 4.6 erläutert. Wie ein entsprechender Plan gestaltet wird, wenn das Forschungsdatum eine Software ist, wird anschließend im Rahmen des Softwaremanagementplans beschrieben. Auch bei dem folgenden Material wird die Differenzierung zwischen Datenmanagementplan und Softwaremanagementplan vorgenommen (siehe Tabelle 5.5).

Tabelle 5.5: Lehrmaterial Datenmanagementplan

Thema	**Seite**
Muster: Datenmanagementplan	185
Muster: Softwaremanagementplan	187

Muster: Datenmanagementplan (Seite 1)

Arbeitsblatt: Datenmanagementplan

Projektname: .....

Forschungsförderer: .....

Förderprogramm: .....

Primärfoscher*in/Wissenschaftler*in/Projektleiter*in: .....

ID Primärforscher*in/Wissenschaftler*in/Projektleiter*in: .....

Projektbeschreibung:

.....

.....

.....

.....

.....

Erstellungsdatum: .....

Änderungsdatum: .....

Datenerhebung:
.....

.....

.....

.....

.....

.....

Datenspeicherung:

.....

.....

.....

.....

Muster: Datenmanagementplan (Seite 2)

Auffindbarkeit der Daten (Findable):

.....

.....

.....

.....

.....

Datenzugang (Accessible):

.....

.....

.....

.....

.....

Interoperabilität der Daten (Interoperable):

.....

.....

.....

.....

Nachnutzung der Daten (Re-Use):

.....

.....

.....

.....

Verantwortlichkeiten:

.....

.....

.....

Muster: Softwaremanagementplan[1]

1. Über die Software
   - Abstract
   - Ziel der Software
2. Umfang der Software
   - Welche Auswirkungen hat die Software?
   - Wer sind die wichtigsten Interessen- bzw. Nutzer:innengruppen?
   - Wie trägt die Software zur Forschung bei?
3. Verwendete Infrastruktur
   - Welche Infrastruktur wird benötigt?
   - Wo wird die Infrastruktur gehostet?
4. Definitionen der Software
   - Welche programm-spezifische Definitionen werden bei der Software verwendet?
   - Wird ein bestimmtes Format genutzt?
5. Qualitätssicherung
   - Wie wird der Code verständlich zur Verfügung gestellt?
   - Wie findet die Qualitätssicherung statt?
   - Wie wird sicher gestellt, dass die Software regelmäßig getestet wird?
   - Wie werden die Nutzer:innen darüber informiert, dass Tests durchgeführt werden?
   - Wird die Software unter verschiedenen Umgebungen ausgeführt?
   - Wie wird die Software dokumentiert?
6. Verwaltung von Abhängigkeiten
   - Welche Software, Modelle, Tools, Bibliotheken und Dienste von Dritten werden genutzt?
   - Welche Datensätze und Onlinedatenbanken von Dritten werden genutzt?
   - Welche Kommunikationsprotokolle und Datenformate werden verwendet?
   - Wie werden die Abhängigkeiten dokumentiert?
   - Wie werden Änderungen an Abhängigkeiten verfolgt?
   - Werden Tools für das Abhängigkeitsmanagement verwendet?

[1] Die Struktur des Softwaremanagementplans basiert auf dem Template des Software Sustainability Institute's unter https://dmponline.dcc.ac.uk/. Archivierte Version: https://perma.cc/NM8M-UTJU.

7. Verwaltung der Software

    - Welcher Aufwand wird für die Entwicklung der Software zur Verfügung stehen?
    - Wie werden die Rollen bei der Softwareentwicklung verteilt?
    - Wie werden Änderungen am Code dokumentiert?
    - Welches Softwareentwicklungsmodell wird verwendet?
    - Wie werden die Releases der Software oder Aktualisierungen der Dienste verwaltet?
    - Wie wird sichergestellt, dass keine Informationen verloren gehen, wenn Entwickler:innen ausscheiden?
    - Wie oft wird der Softwaremanagementplan überprüft und überarbeitet?
    - In welchem Verhältnis steht der Softwaremanagementplan zum Datenmanagementplan?
    - Ist es geplant, die Software als Open-Source-Projekt weiterzuentwickeln?

8. Nutzung der Software

    - Wie wird die Software den Nutzer:innen zur Verfügung gestellt?
    - Wie ist der Support für die Software sichergestellt?
    - Wie wird die Software hinterlegt, um die langfristige Verfügbarkeit zu gewährleisten?
    - Wie können Nutzer:innen zur Software beitragen?

9. Urheberrecht

    - Wer sind Urheber:innen der Software?
    - Welche Lizenz wird der Software vergeben?

5.6 Ethische Aspekte

Ethische Aspekte des FDM wurden in Kapitel 4.7 aufgezeigt. Im Fokus stehen hier forschungsethische Fragen, die mit dem Umgang mit den Forschungsdaten zusammenhängen. Eine Einschätzung des richtigen Verhaltens ist nicht immer leicht, wobei Richtlinien der Guten Wissenschaftlichen Praxis Hilfestellung leisten. Bei dem folgenden Dokument handelt es sich um eine Diskussionsvorlage, ohne dass eine konkrete Musterlösung angegeben wird, da ethische Problemstellungen oftmals nicht allgemeingültig beantwortet werden können. In Tabelle 5.6 wird eine Übersicht über die Lehrmaterialien zu diesem Thema gegeben.

Tabelle 5.6: Lehrmaterial Ehische Aspekte

Thema	**Seite**
Diskussionsvorlage: Ethische Aspekte	190

Diskussionsvorlage: Ethische Aspekte

In der Mitte der Promotionsphase hat Alex bereits erste Daten erhoben und möchte wissenschaftliche Artikel sowie die dazugehörigen Daten publizieren. Dabei stößt Alex auf verschiedene ethische Dilemmas im Hinblick auf den Umgang mit Kolleg:innen und Guter Wissenschaftlicher Praxis.

1. Diskutieren Sie darüber, wie sich Alex in den folgenden Situationen verhalten sollte[1]:

   (a) Ich bin Nachwuchsforscher:in, der in mühevoller Kleinarbeit eine große Menge an Daten gesammelt hat. Mein erster Artikel, der auf diesen Daten basiert, wurde gerade zur Veröffentlichung angenommen. Ein älterer Kollege aus meiner Abteilung hat mich kontaktiert, um nach den Daten zu fragen. Er hat einen großen Einfluss auf meine berufliche Entwicklung. Was soll ich tun?

   - Ich schicke dem älteren Kollegen die Daten.
   - Ich sage dem älteren Kollegen, dass die Daten zur Verfügung stehen werden, sobald der letzte Artikel, den ich über das Thema schreiben möchte, veröffentlicht ist. Dies kann bis zu einem oder zwei Jahren dauern.
   - Ich sage dem älteren Kollegen, dass ich ihn nicht bevorzugt behandeln möchte.
   - Ich sage dem älteren Kollegen, dass ich bereit bin, die Daten unter der Bedingung zur Verfügung zu stellen, dass ich als Mitautor in allen Veröffentlichungen genannt werde, die auf diesen Daten basieren.

   (b) Ich bin Forscher:in mit unbefristeter Stelle und brauche dringend einen weiteren Artikel, der veröffentlicht werden soll. Die Haupthypothese in dem Artikel, an dem ich arbeite, ist, dass A Einfluss auf B hat. Während meiner Untersuchungen habe ich mehrere Variablen für Kontrollzwecke definiert. Während der Analyse wird deutlich, dass es keinen Einfluss von A auf B gibt, es sei denn, ich entferne eine der Kontrollvariablen. Was soll ich tun?

   - Ich entferne die Variable und erwähne sie nicht in meinem Artikel.
   - Ich entferne die Variable, suche nach wissenschaftlichen Argumenten dafür und erwähne sie in meinem Artikel.
   - Ich reiche meinen Artikel ein, ohne die Variable zu entfernen, auch wenn dies bedeuten könnte, dass mein Artikel nicht veröffentlicht wird.
   - Ich frage eine Kollegin, was sie tun würde und folge ihrer Meinung.

[1] Quelle der Aufgabe: Erasmus University Rotterdam. (2021). Dilemma Game. Professionalism and Integrity in Research. [Archivierte Version: https://perma.cc/75S6-KDU4]. https://www.eur.nl/en/about-eur/policy-and-regulations/integrity/research-integrity/dilemma-game. Dilemma 13, 49 und 32.

(c) Ich bin Doktorand:in und habe gerade mit der Analyse meiner Daten begonnen. Dabei wird mir klar, dass etwas bei der Datenerhebung oder -eingabe schief gelaufen ist, da einige Werte eindeutig falsch sind. Die Organisation, die die Dateneingabe durchgeführt hat, zieht die Möglichkeit in Erwägung, dass bei der Dateneingabe etwas schief gelaufen ist. Ich habe keine Zeit, um neue Daten zu sammeln. Was muss ich tun?

- Ich beschließe, die Daten selbst zu korrigieren; es ist recht klar, wie man das macht.
- Ich beschließe, die Beobachtungen mit den falschen Werten zu löschen und meine Forschung mit weniger Beobachtungen durchzuführen, als ursprünglich geplant.
- Ich bespreche das Problem mit meinen Betreuer:innen und überlasse ihnen die Entscheidung, was zu tun ist.
- Ich bitte das Unternehmen, die Daten zu korrigieren und in einem offiziellen Schreiben zuzugeben, dass sie für die falsche Dateneingabe verantwortlich waren.

5.7 Datenschutz

Das Thema Datenschutz wurde in Kapitel 4.8 ausführlich beschrieben und kann in verschiedene Themen untergliedert werden. Besonders wichtig ist, dass Studierende Kenntnisse darüber aufbauen, bei welchen Daten es sich um personenbezogene Daten handelt. Auch die Einholung von informierten Einwilligungserklärungen ist wichtig, damit die erhobenen Daten verwendet werden dürfen. Die Anonymisierung und Pseudonymisierung von Daten ist für die Datenaufbereitung und Publikation entscheidend. Diese genannten Themen sind im Folgenden durch Arbeitsblätter und Handreichungen für die Lehre aufbereitet. Eine Übersicht über das Material bietet Tabelle 5.7.

Tabelle 5.7: Lehrmaterial Datenschutz

Thema	**Seite**
Arbeitsblatt: Personenbezogene Daten	193
Musterlösung: Personenbezogene Daten	194
Arbeitsblatt: Anonymisierung qualitativer Daten	195
Musterlösung: Anonymisierung qualitativer Daten	197
Arbeitsblatt: Anonymisierung quantitativer Daten	199
Musterlösung: Anonymisierung quantitativer Daten	200
Checkliste: Informierte Einwilligung	202
Muster: Verzeichnis von Verarbeitungstätigkeiten	204

Arbeitsblatt: Personenbezogene Daten

Carla führt Interviews mit Studierenden in Deutschland und Kuba durch. Bevor sie mit den Leitfragen startet, bittet Sie alle Teilnehmenden, noch ein paar persönliche Angaben zu machen, damit sie den demografischen Hintergrund der Kohorte beschreiben kann.

1. Kreuzen Sie an, welche der erhobenen Daten in Carlas Fall personenbezogen sind.

Datum	personenbezogen	nicht personenbezogen
Nachname	☐	☐
Vorname	☐	☐
Wohnort	☐	☐
Universität	☐	☐
Muttersprache	☐	☐
Alter	☐	☐

   Begründen Sie Ihre Entscheidung.

   .....

   .....

   .....

2. Erläutern Sie, welche Daten als personenbezogen bezeichnet werden.

   .....

   .....

   .....

3. Erläutern Sie den Unterschied zwischen direkten und indirekten Identifikatoren.

   .....

   .....

   .....

Musterlösung: Personenbezogene Daten

Carla führt Interviews mit Studierenden in Deutschland und Kuba durch. Bevor sie mit den Leitfragen startet, bittet Sie alle Teilnehmenden, noch ein paar persönliche Angaben zu machen, damit sie den demografischen Hintergrund der Kohorte beschreiben kann.

1. Kreuzen Sie an, welche der erhobenen Daten in Carlas Fall personenbezogen sind.

Datum	personenbezogen	nicht personenbezogen
Nachname	☒	☐
Vorname	☒	☐
Wohnort	☒	☐
Universität	☒	☐
Muttersprache	☒	☐
Alter	☒	☐

   Begründen Sie Ihre Entscheidung.

   Je nach Zusammensetzung der Gruppe können alle dieser Angaben eine Person eindeutig identifizieren oder den Personenkreis stark eingrenzen. Bei qualitativen Interviews im Rahmen einer Bachelorarbeit, ist davon auszugehen, dass die Kohorte zu gering ist.

2. Erläutern Sie, welche Daten als personenbezogen bezeichnet werden.

   Informationen über natürliche Personen, die direkt oder indirekt zur Identifikation der Person führen können.

3. Erläutern Sie den Unterschied zwischen direkten und indirekten Identifikatoren.

   Direkte Identifikatoren lassen es zu, dass eine Person eindeutig oder nahezu eindeutig ohne zusätzliche Informationen identifizierbar ist. Indirekte Identifikatoren (auch: Quasi-Identifikatoren) können durch die Kombination mit anderen Daten die Identifikation ermöglichen. Sie allein lassen jedoch keine Identifikation zu.

Arbeitsblatt: Anonymisierung qualitativer Daten

Carla führt leitfadengestützte Interviews mit Studierenden in Deutschland und Kuba durch. Die Interviews werden transkribiert und sollen für die Nachnutzung und Transparenz in der Forschung publiziert werden. Zunächst müssen die Transkripte jedoch anonymisiert werden.

1. Erklären Sie, warum eine gründliche Anonymisierung von qualitativen Daten wichtig ist.

   .....

   .....

   .....

   .....

Gegeben ist der folgende Auszug von Carlas Transkript:

Carla: Guten Tag Pia, ich freue mich, dass du an dem Interview teilnimmst.

Pia: Hallo, das mache ich gern.

Carla: Zunächst würde ich gern wissen inwieweit du mit anderen Kommiliton:innen Tools benutzt, um kollaborativ zu arbeiten.

Pia: Meistens gründen wir innerhalb der Übungsgruppe eine Chatgruppe bei Threema, in der wir uns darüber absprechen, wer welche Aufgabe übernimmt. Ich weiß nicht, ob das so richtig als Tool zählt. Könntest du mir ein paar Beispiele nennen?

Carla: Gerne. Es gibt viele unterschiedliche Kollaborationstools mit diversen Funktionen. Manche dienen eher der Verwaltung von Inhalten und unterstützen das gemeinsame Arbeiten an Dokumenten, wie Moodle oder OneDrive. Andere Kollaborationstools können Aufgaben den beteiligten Personen zuweisen und es kann mitverfolgt werden, wie weit die Bearbeitung von Projektteilen vorangeschritten ist. Beispiele wären Asana, Slack oder auch zum Teil MS Teams.

Pia: Ah, okay. Moodle nutzen wir eher, um unsere Übungsaufgaben von Herrn Fischer zu bekommen, aber nicht innerhalb der Gruppe. Ich weiß, dass die Universität Hamburg auch MS Teams zur Verfügung stellt, aber in der Gruppe nutzen wir am liebsten Asana. [...]

2. Tragen Sie die direkten Identifikatoren aus dem Transkript in die Tabelle (in der Spalte Vorkommen im Transkript) ein.

3. Tragen Sie die indirekten Identifikatoren aus dem Transkript in die Tabelle (in der Spalte Vorkommen im Transkript) ein.

4. Tragen Sie zu allen Identifikatoren eine mögliche Form der Anonymisierung (in der Spalte Anonymisierung) ein.

	Vorkommen im Transkript	Anonymisierung
direkter Identifikator	1.	
	2.	
	3.	
	4.	
indirekter Identifikator	5.	
	6.	
	7.	
	8.	

Musterlösung: Anonymisierung qualitativer Daten

Carla führt leitfadengestützte Interviews mit Studierenden in Deutschland und Kuba durch. Die Interviews werden transkribiert uns sollen für die Nachnutzung und Transparenz in der Forschung publiziert werden. Zunächst müssen die Transkripte jedoch anonymisiert werden.

1. Erklären Sie, warum eine gründliche Anonymisierung von qualitativen Daten wichtig ist.

   Insbesondere bei qualitativen Daten ist die Rückführung auf die interviewte Person durch die Inhalte, aber auch ihre Sprache möglich. Bei der Publikation der Daten kann zum Publikationszeitpunkt nicht abgesehen werden, mit welchen weiteren Daten das Interview in Zusammenhang gebracht werden kann. Da teilweise Meinungen, sensible Daten, etc. erfragt oder in einem Gespräch erwähnt werden, muss zum Schutz der interviewten Person gründlich anonymisiert werden.

   Gegeben ist der folgende Auszug von Carlas Transkript. (**Fett** gedruckte Wörter markieren zu anonymisierende Angaben.)

Carla: Guten Tag **Pia**, ich freue mich, dass du an dem Interview teilnehmen.

Pia: Hallo, das mache ich gern.

Carla: Zunächst würde ich gern wissen, inwieweit du mit anderen Kommiliton:innen Tools benutzt, um kollaborativ zu arbeiten.

Pia: Meistens gründen wir innerhalb der Übungsgruppe eine Chatgruppe bei Threema, in der wir uns darüber absprechen, wer welche Aufgabe übernimmt. Ich weiß nicht, ob das so richtig als Tool zählt. Könntest du mir ein paar Beispiele nennen?

Carla: Gerne. Es gibt viele unterschiedliche Kollaborationstools mit diversen Funktionen. Manche dienen eher der Verwaltung von Inhalten und unterstützen das gemeinsame Arbeiten an Dokumenten, wie Moodle oder OneDrive. Andere Kollaborationstools können Aufgaben den beteiligten Personen zuweisen und es kann mitverfolgt werden, wie weit die Bearbeitung von Projektteilen vorangeschritten ist. Beispiele wären Asana, Slack oder auch zum Teil MS Teams.

Pia: Ah, okay. **Moodle** nutzen wir eher um unsere Übungsaufgaben von **Herrn Fischer** zu bekommen, aber nicht innerhalb der Gruppe. Ich weiß, dass die **Universität Hamburg** auch **MS Teams** zur Verfügung stellt, aber in der Gruppe nutzen wir am liebsten Asana. [...]

2. Tragen Sie die direkten Identifikatoren aus dem Transkript in die Tabelle (in der Spalte Vorkommen im Transkript) ein.

3. Tragen Sie die indirekten Identifikatoren aus dem Transkript in die Tabelle (in der Spalte Vorkommen im Transkript) ein.

4. Tragen Sie zu allen Identifikatoren eine mögliche Form der Anonymisierung (in der Spalte Anonymisierung) ein.

	Vorkommen im Transkript	Anonymisierung
direkter Identifikator	1. Clara	I oder interviewende Person
	2. Pia	S oder Studierende
	3. Herr Fischer	@@Name Hochschullehrende##
	4. Universität Hamburg	@@Name Universität##
indirekter Identifikator	5. Moodle	@@Contentmanagementsystem##
	6. MS Teams	@@Videokonferenztool##
	7.	
	8.	

Erklärungen zur Tabelle:

Zu 1/2: Die Interviewpartner:innen werden anonymisiert und abgekürzt. Es wird auch die interviewende Person anonymisiert, da des Öfteren mehrere Personen bei der Interviewführung eingebunden sind.

Zu 3: Hochschullehrkräfte sollten anonymisiert werden, da somit ein Rückbezug auf die interviewte Person und die Institution möglich ist. Auch ist es möglich, dass Angaben über diese dritte Person getätigt werden, die zum Nachteil für sie sein können.

Zu 4: Die Angabe der Institution sollte nur in ausgewählten Fällen getätigt werden. Beispielsweise, wenn es sich um eine interne Befragung handelt, die von der Institution in Auftrag gegeben wurde. Werden jedoch Studien, beispielsweise von Studierenden, an mehreren Institutionen vorgenommen, sodass das N der Befragten an den einzelnen Institutionen gering ist, sollte auf die Angabe der Institution verzichtet werden.

Zu 5: Bei der Nennung von Moodle sollte anonymisiert werden, da nicht jede Universität Moodle nutzt und somit leichter nachvollziehbar ist, an welcher Universität das Interview durchgeführt wurde.

Zu 6: Auch MS Teams sollte anonymisiert werden, da das Tool nicht von jeder Universität erworben wird und somit Rückschlüsse auf die Universität möglich sind.

Prinzipiell ist diese Einschätzung davon abhängig, wie groß die Kohorte der Befragten und der zugehörigen Institutionen sind.

Arbeitsblatt: Anonymisierung quantitativer Daten

Carla führt Interviews mit Studierenden in Deutschland und Kuba durch. Bevor sie mit den Leitfragen startet, bittet sie alle Teilnehmenden, noch ein paar persönliche Angaben zu machen, damit sie den demografischen Hintergrund der Kohorte beschreiben kann. Diese Informationen müssen vor der Publikation anonymisiert werden.

1. Die nachstehende Tabelle zeigt einige bewertete Identifikatoren mit festgestelltem Offenlegungsrisiko. Welche Maßnahmen würden Sie ergreifen, um das Offenlegungsrisiko zu verringern?[1]

Identifikator	**Offenlegungsrisiko**	**Gegenmaßnahmen**
Wohnort	Bei kleinen Wohnorten niedrige Bewohner:innenzahlen; Teilnehmende, die auf diese Frage geantwortet haben, können sehr leicht identifiziert werden (insbesondere in Kombination mit anderen Identifikatoren)	
Alter	Geringe Anzahl an Ausreißern (z. B. ältere Schüler:innen, die die Klasse wiederholt haben, oder jüngere Schüler:innen, die die Klasse übersprungen haben)	
Besuchte Schulen in den letzten drei Jahren	Geringe Anzahl an häufig wechselnden Teilnehmenden	
Herkunft	Geringe Anzahl an Teilnehmenden mit anderer Herkunft	
Muttersprache	Geringe Anzahl an Teilnehmenden mit anderer Muttersprache	

[1] Quelle der Aufgabe: basierend auf Van den Eynden, V. (2018). Exercise: De-identification on quantitative data. UK Data Service.

Musterlösung: Anonymisierung quantitativer Daten

Carla führt Interviews mit Studierenden in Deutschland und Kuba durch. Bevor sie mit den Leitfragen startet, bittet sie alle Teilnehmenden, noch ein paar persönliche Angaben zu machen, damit sie den demografischen Hintergrund der Kohorte beschreiben kann. Diese Informationen müssen vor der Publikation anonymisiert werden.

1. Die nachstehende Tabelle zeigt einige bewertete Identifikatoren mit festgestelltem Offenlegungsrisiko. Welche Maßnahmen würden Sie ergreifen, um das Offenlegungsrisiko zu verringern?[1]

Identifikator	**Offenlegungsrisiko**	**Gegenmaßnahmen**
Wohnort	Bei kleinen Wohnorten niedrige Bewohner:innenzahlen; Teilnehmende, die auf diese Frage geantwortet haben, können sehr leicht identifiziert werden (insbesondere in Kombination mit anderen Identifikatoren)	Variable aus dem Datensatz entfernen
Alter	Geringe Anzahl an Ausreißern (z. B. ältere Schüler:innen, die die Klasse wiederholt haben, oder jüngere Schüler:innen, die die Klasse übersprungen haben)	Aggregation (z. B. von 8–12 Jahren)
Besuchte Schulen in den letzten drei Jahren	Geringe Anzahl an häufig schulwechselnden Teilnehmenden	Variable aus dem Datensatz entfernen
Herkunft	Geringe Anzahl an Teilnehmenden mit anderer Herkunft	Je nach Anzahl der Teilnehmenden mit anderer Herkunft, kann die Variable umkodiert werden in „Andere“ oder muss aus dem Datensatz entfernt werden

[1] Quelle der Aufgabe: basierend auf Van den Eynden, V. (2018). Exercise: De-identification on quantitative data. UK Data Service.

Muttersprache	Geringe Anzahl an Teilnehmenden mit anderer Muttersprache	Je nach Anzahl der Teilnehmenden mit anderer Muttersprache, kann die Variable umkodiert werden in „Andere“ oder muss aus dem Datensatz entfernt werden

Checkliste: Informierte Einwilligung (Seite 1)

Checkliste: Anforderungen an eine Einwilligung nach DSGVO

Zu prüfen...	Ja	Nein
Allgemein		
Wird die Einwilligung zeitlich vor der Erhebung und Verwendung von personenbezogenen Daten eingeholt?		
Bezieht sich die Einwilligung nur auf Datenverarbeitungen, die nicht bereits durch gesetzlicher Grundlage legitimiert sind? (Die rechtlichen Konsequenzen einer „überflüssigen“ Einwilligung sind umstritten)		
Form		
Ist die Eindeutigkeit resp. die aktive Handlung der Einwilligung gewährleistet? Die Einwilligung muss eindeutig erfolgen, d.h. durch eine aktive Handlung des Einwilligenden (z.B. durch Ankreuzen eines Auswahlfeldes).		
Ist die Einwilligung „Teil eines größeren Dokuments“? Wenn ja, dann muss sie von den anderen Sachverhalten des Dokuments klar zu unterscheiden sein: Werden Anforderungen an die „optische“ Hervorhebung der datenschutzrechtlichen Einwilligung eingehalten?		
Ist an eine zweifache Ausfertigung des Dokumentes gedacht? (Verbleib des Originals beim Verantwortlichen, Kopie beim/bei der Betroffenen)		
Existiert eine Bestätigung der Gelegenheit für Rückfragen? Diesbezüglich empfehlen sich Formulierungen wie: „...Ich hatte Gelegenheit, Fragen zu stellen. Diese wurden vollständig und umfassend beantwortet. ...“; Benennung desjenigen, der die Fragen beantwortet hat, ggfs. sollte dessen Name handschriftlich auf dem Einwilligungsbogen nachtragen		
Freiwilligkeit		
Hatte der Betroffene eine echte Wahl zwischen Zustimmung und Ablehnung?		
Ist gewährleistet, dass die Erfüllung eines Vertrages oder die Erbringung einer Dienstleistung nicht von der Einwilligung abhängig gemacht wurde, wenn die Einwilligung nicht zwingend zur Erfüllung benötigt wird (Kopplungsverbot)?		
Informiertheit		
Hat der Betroffene alle erforderlichen Informationen (inkl. Vor- und Nachteile) erhalten? Insbesondere: – Datenverwendung (Zweck, Ziel, Nutzen, Chancen und Risiken) – Personenkreis, der auf Daten Zugriff erlangen darf – Art der von der Verarbeitung betroffenen Daten – Datenweitergabe (an wen, ggfs. Speicherung an welchem Ort, Land)		
Werden alle in Art. 13 DS-GVO bzw. Art 14 DS-GVO genannten Informationen bereitgestellt? Insbesondere: – Ansprechpartner sowie Kontaktdaten (Verantwortlicher, Datenschutzbeauftragter, ...) – Rechtsgrundlage der Vereinbarung – Empfänger – Speicherdauer – Rechte des Betroffenen (Einsichtnahme, Korrektur, Löschen, Widerruf Einwilligung)		
Sind der Verantwortliche sowie seine Vertreter eindeutig benannt? Stehen alle benötigten Kontaktdaten dem Betroffenen zur Verfügung?		

FDMentor Erstellt im Rahmen des FDMentor-Projektes
Projektlaufzeit: 1. Mai 2017 bis 30. April 2019
Kontakt: fdmentor@hu-berlin.de
Twitter: @fd_mentor
https://hu.berlin/fdmentor

gefördert vom: Bundesministerium für Bildung und Forschung

Checkliste: Informierte Einwilligung (Seite 2)

Zu prüfen...	Ja	Nein
Gibt es einen (verständlichen) Hinweis auf die Folgen, die die Verweigerung der Einwilligung für den Betroffenen haben kann?		
Bezieht sich bei der Verarbeitung besonderen Kategorien von Daten (Art. 9 DS-GVO) die Einwilligungserklärung ausdrücklich auch auf diese Daten?		
Bestimmtheit		
Bezieht sich die Einwilligung auf einen konkret benannten Fall? Generaleinwilligungen sind unwirksam; für verschiedene Zwecke müssen separate Einwilligungen eingeholt / abgegeben werden		
Ist die Einwilligungserklärung von etwaigen sonstigen (datenschutzrelevanten) Hinweisen deutlich getrennt? Es ist zu vermeiden, dass der Betroffene auf Grund Unübersichtlichkeit des Dokumentes nicht erkennt, ob und gegebenenfalls in was er eigentlich einwilligt bzw. einwilligen soll.		
Widerrufbarkeit		
Ist auf den jederzeit möglichen Widerruf der Einwilligung im Einwilligungsformular hingewiesen?		
Ist im Einwilligungsformular darauf hingewiesen, dass ein Widerruf immer nur für die nach dem Widerruf erfolgende geplante Verarbeitung gilt?		
Ist der Widerruf der Einwilligung (mindestens) so einfach möglich wie das Erteilen der Einwilligung selbst?		
Gibt es einen (verständlichen) Hinweis auf die Folgen des Widerrufs?		
Einwilligung Minderjähriger		
Bei der Verarbeitung mittels „Dienste der Informationsgesellschaft“ - Art. 8 beachtet?		
Wenn Einwilligung der Eltern vorliegt: spätestens bei Volljährigkeit des Betroffenen ist weitere Verarbeitung nur mit Einwilligung des Betroffenen selbst möglich. Gibt es Mechanismus um die Verarbeitung der Daten zum Zeitpunkt „x“ zu stoppen?		
Nachweisbarkeit		
Ist der Nachweis gegeben, dass die Einwilligung vom Betroffenen abgegeben wurde?		
Ist der Nachweis gegeben, dass die Einwilligung den Anforderungen der Ds-GVO genügend abgegeben wurde?		
Werden erteilte Einwilligungen protokolliert? Wenn ja: Sind ausreichende technische und organisatorische Maßnahmen zum Schutz der Protokolle getroffen? (Beweisfestigkeit)		
Sind erteilte Einwilligungen jederzeit abrufbar?		

Quelle:
Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS) (2016). EU DS-GVO: Anforderungen an eine Einwilligung. Online verfügbar: https://www.gesundheitsdatenschutz.org/download/einwilligung.pdf. [letzter Zugriff: 2021-10-24]. Lizenziert unter der Creative Commons-Lizenz Share-Alike (4.0 Deutschland Lizenzvertrag).

Muster: Verzeichnis von Verarbeitungstätigkeiten (Seite 1)[1]

Verzeichnis von Verarbeitungstätigkeiten **Verantwortlicher** **gem. Artikel 30 Abs. 1 DSGVO**	Vorblatt

Angaben zum Verantwortlichen

Name und Kontaktdaten natürliche Person/juristische Person/Behörde/Einrichtung etc.

Name

Straße

Postleitzahl

Ort

Telefon

E-Mail-Adresse

Internet-Adresse

Angaben zum ggf. gemeinsam mit diesem Verantwortlichen

Name

Straße

Postleitzahl

Ort

Telefon

E-Mail-Adresse

Angaben zum Vertreter des Verantwortlichen

Name und Kontaktdaten natürliche Person/juristische Person/Behörde/Einrichtung etc.

Name

Straße

Postleitzahl

Ort

Telefon

E-Mail-Adresse

Angaben zur Person des Datenschutzbeauftragten * (extern mit Anschrift)

* sofern gem. Artikel 37 DS-GVO benannt

Anrede Titel

Name, Vorname

Straße

Postleitzahl

Ort

Telefon

E-Mail-Adresse

[1] Quelle: Der Bundesbeauftragte für den Datenschutz und Informationsfreiheit (BfDI). (2018). Verzeichnis von Verarbeitungstätigkeiten. [Archivierte Version: https://perma.cc/P42U-KEBJ]. https://www.datenschutzkonferenz-online.de/media/ah/201802_ah_muster_verantwortliche.pdf.

Muster: Verzeichnis von Verarbeitungstätigkeiten (Seite 2)

Verarbeitungstätigkeit: Benennung: _____		**lfd. Nr.:** _____
Datum der Einführung:	Datum der letzten Änderung:	
Verantwortliche Fachabteilung Ansprechpartner Telefon E-Mail-Adresse (Art. 30 Abs. 1 S. 2 lit a)		
Zwecke der Verarbeitung (Art. 30 Abs. 1 S. 2 lit b)		
Optional: Name des eingesetzten Verfahrens		
Beschreibung der Kategorien betroffener Personen (Art. 30 Abs. 1 S. 2 lit. c)	☐ Beschäftigte ☐ Interessenten ☐ Lieferanten ☐ Kunden ☐ Patienten ☐ ☐ ☐	
Beschreibung der Kategorien von personenbezogenen Daten (Art. 30 Abs. 1 S. 2 lit. c)	☐ ☐ ☐ Besondere Kategorien personenbezogener Daten (Art. 9): ☐	

Muster: Verzeichnis von Verarbeitungstätigkeiten (Seite 3)

Kategorien von Empfängern, gegenüber denen die personen-bezogenen Daten offen gelegt worden sind oder noch werden (Art. 30 Abs. 1 S. 2 lit. d)	☐ intern (Zugriffsberechtigte) Abteilung/ Funktion
	☐ extern Empfängerkategorie
	☐ Drittland oder internationale Organisation (Kategorie)
ggf. Übermittlungen von personenbezogenen Daten an ein Drittland oder an eine inter-nationale Organisation (Art. 30 Abs. 1 S. 2 lit. e)	☐ Datenübermittlung findet nicht statt und ist auch nicht geplant ☐ Datenübermittlung findet wie folgt statt:
Nennung der konkreten Datenempfänger	☐ Drittland oder internationale Organisation (Name)
Sofern es sich um eine in Art. 49 Abs. 1 Unterabsatz 2 DS-GVO genannte Datenübermittlung handelt.	Dokumentation geeigneter Garantien
Fristen für die Löschung der verschiedenen Datenkategorien (Art. 30 Abs. 1 S. 2 lit. f)	

Technische und organisatorische Maßnahmen (TOM) gemäß Art. 32 Abs.1 DSGVO (Art. 30 Abs. 1 S. 2 lit. g) *Siehe TOM-Beschreibung in den „Hinweisen zum Verzeichnis von Verarbeitungstätigkeiten", Ziff. 6.7. und 6.8*

…… …… ……
Verantwortlicher Datum Unterschrift

5.8 Ordnung und Struktur

In Kapitel 4.9 werden verschiedene Maßnahmen zur Erzeugung von Ordnung und Struktur vorgestellt. Die Versionierung und das Arbeiten mit Namenskonventionen sind dabei zentrale Bestandteile. Im Folgenden sind Lehrmaterialien (siehe Tabelle 5.9) für diese Themen gegeben.

Tabelle 5.9: Lehrmaterial Ordnung und Struktur

Thema	**Seite**
Arbeitsblatt: Versionierung	208
Musterlösung: Versionierung	209
Arbeitsblatt: Namenskonventionen	210
Musterlösung: Namenskonventionen	211

Arbeitsblatt: Versionierung

Timo möchte sich an dem Open-Source-Projekt *Digital Public Health for All* beteiligen. Es existiert bereits ein Git-Repository des Projekts, bei dem Timo Schreibzugriff erhält. Timo muss sich zunächst in die Versionierung mit Git einarbeiten.

1. Nennen Sie zwei Gründe für die Nutzung von Versionierung, auch wenn man allein an einem Projekt arbeitet.

   1. .....
   2. .....

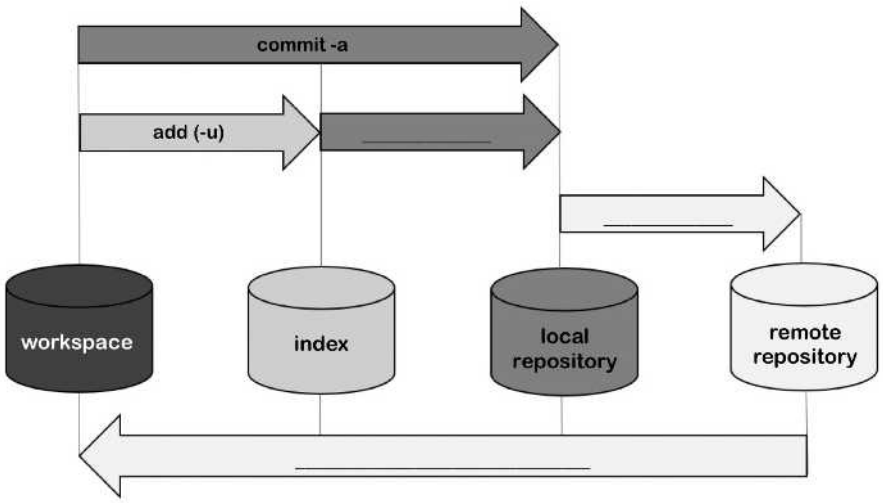
2. Für eine erfolgreiche Arbeit mit Git werden verschiedene Komponenten benötigt, die unterschiedliche Aufgaben übernehmen. Nennen Sie konkrete Tools, die die folgenden Komponenten bilden:

   GUI: .....

   Software: .....

   Browsernutzung: .....

3. Git beinhaltet vier verschiedene Speicherorte. Beschriften Sie die Pfeile mit den zugehörigen Kommandos, um auf die Speicherorte zuzugreifen.



Musterlösung: Versionierung

Timo möchte sich an dem Open-Source-Projekt *Digital Public Health for All* beteiligen. Es existiert bereits ein Git-Repository des Projekts, bei dem Timo Schreibzugriff erhält. Timo muss sich zunächst in die Versionierung mit Git einarbeiten.

1. Nennen Sie zwei Gründe für die Nutzung von Versionierung, auch wenn man allein an einem Projekt arbeitet.

    1. Es ist klar ersichtlich, welche Version die aktuellste für die weitere Arbeit ist.
    2. Ältere Versionen können wiederhergestellt werden und Änderungen können nachverfolgt werden.

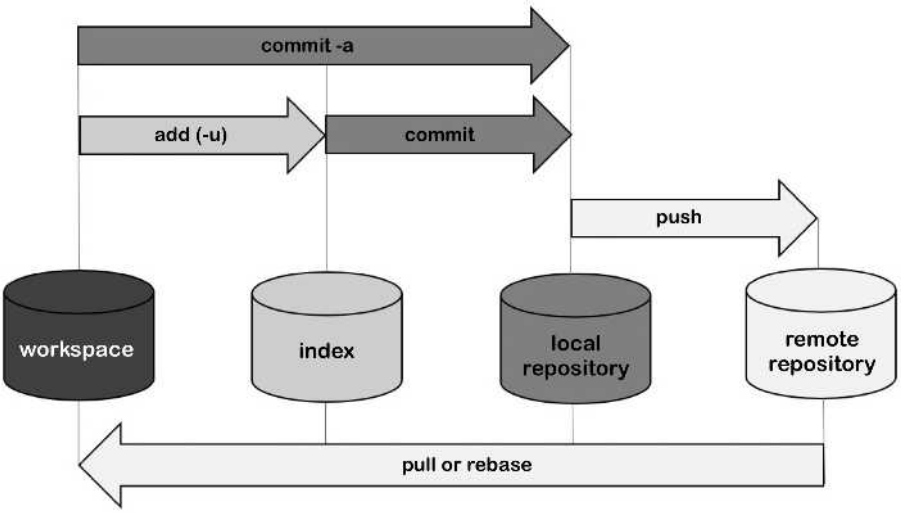
2. Für eine erfolgreiche Arbeit mit Git werden verschiedene Komponenten benötigt, die unterschiedliche Aufgaben übernehmen. Nennen Sie konkrete Tools, die die folgenden Komponenten bilden:

    GUI: GitHab Desktop, SourceTree, Fork

    Software: git

    Browsernutzung: GitLab, GitHub

3. Git beinhaltet vier verschiedene Speicherorte. Beschriften Sie die Pfeile mit den zugehörigen Kommandos, um auf die Speicherorte zuzugreifen.



Arbeitsblatt: Namenskonventionen

Timo möchte in einem Open-Source-Projekt gemeinsam mit Kayla eine App programmieren. Für die Entwicklung nutzen sie ein gemeinsames Git-Repository und müssen vorab einige organisatorische Fragen zum kollaborativen Programmieren klären.

1. Kreuzen Sie an, auf welche Benennungen sich beide verständigen sollten.

   - ☐ Variablennamen
   - ☐ Datentypen
   - ☐ Appname
   - ☐ Methodennamen
   - ☐ Klassennamen

2. Kreuze Sie an, welche der folgenden Beispiele gute Namenskonventionen sind.

   - ☐ Computer Science Timo Kayla final!.pdf
   - ☐ aufgabe2_17022022.pdf
   - ☐ 17022022+Timo+Kayla_v1.csv
   - ☐ 7yK8s890Jnshxy.tex
   - ☐ 20220519_KI_AB3_Kayla.txt
   - ☐ computerScience-Timo-Kayla-fertig-überarbeitet
   - ☐ Tagung-Artikel.tex
   - ☐ Aufgabe1_Timo_Kayla-v1.tex

3. Begründen Sie Ihre Einschätzung.

   .....

   .....

   .....

   .....

   .....

Musterlösung: Namenskonventionen

Timo möchte in einem Open-Source-Projekt gemeinsam mit Kayla eine App programmieren. Für die Entwicklung nutzen sie ein gemeinsames Git-Repository und müssen vorab einige organisatorische Fragen zum kollaborativen Programmieren klären.

1. Kreuzen Sie an, auf welche Benennungen sich beide verständigen sollten.

   ☒ Variablennamen
   ☒ Datentypen
   ☒ Appname
   ☒ Methodennamen
   ☒ Klassennamen

2. Kreuze Sie an, welche der folgenden Beispiele gute Namenskonventionen sind.

   ☐ Computer Science Timo Kayla final!.pdf
   ☐ aufgabe2_17022022.pdf
   ☐ 17022022+Timo+Kayla_v1.csv
   ☐ 7yK8s890Jnshxy.tex
   ☒ 20220519_KI_AB3_Kayla.txt
   ☐ computerScience-Timo-Kayla-fertig-überarbeitet
   ☐ Tagung-Artikel.tex
   ☒ Aufgabe1_Timo_Kayla-v1.tex

3. Begründen Sie Ihre Einschätzung.

   Dateinamen sollten die wichtigsten Informationen enthalten und möglichst präzise sein, um sich von anderen, ähnlichen Dateien abzuheben. Die Autor:innen sollten in diesem Fall aus dem Namen hervorgehen und auf Sonderzeichen sollte verzichtet werden. Auch Bezeichnungen wie „final“ sollten durch eine klare Versionsnummer und/oder ein Datum ersetzt werden, da Dokumente oftmals zu einem späteren Zeitpunkt erneut überarbeitet werden.

5.9 Speicher und Back-up

Speicher und Back-up (Kapitel 4.10) ist ein essentielles Thema, um sich vor Datenverlust zu schützen. Im Fokus steht dabei die Auswahl geeigneter Speichermedien. Dem Thema entsprechendes Lehrmaterial wird in Tabelle 5.10 zusammengefasst.

Tabelle 5.10: Lehrmaterial Speicher und Back-up

Thema	**Seite**
Arbeitsblatt: Vergleich von Speichermedien	213
Musterlösung: Vergleich von Speichermedien	214
Checkliste: Speicher	216

Arbeitsblatt: Vergleich von Speichermedien

Carla möchte die erhobenen Forschungsdaten während der Bearbeitung ihrer Bachelorarbeit speichern und wägt ab, welche Speichermedien die geeigneten für sie sind. Bei ihr kommt erschwerend hinzu, dass sie auch unterwegs Daten erhebt.

1. Vergleichen Sie die Vor- und Nachteile der genannten Speichermedien[1].

	Vorteile	**Nachteile**
Eigener PC		
Mobile Speichermedien (USB-Sticks, externe Festplatten etc.)		
Institutionelle Speicherorte (Cloud-Services, virtuelle Laufwerke etc.		
Externe Speicherorte (Cloud-Services eines kostenlosen oder kostenpflichtigen Anbieters etc.		

[1] Quelle der Aufgabe: basierend auf Biernacka, K., Buchholz, P., Danker, S. A., Dolzycka, D., Engelhardt, C., Helbig, K., Jacob, J., Neumann, J., Odebrecht, C., Petersen, B., Slowig, B., Trautwein-Bruns, U., Wiljes, C. & Wuttke, U. (2021). T*rain-the-Trainer Konzept zum Thema Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.5773203.

Musterlösung: Vergleich von Speichermedien

Carla möchte die erhobenen Forschungsdaten während der Bearbeitung ihrer Bachelorarbeit speichern und wägt ab, welche Speichermedien die geeigneten für sie sind. Bei ihr kommt erschwerend hinzu, dass sie auch unterwegs Daten erhebt.

1. Vergleichen Sie die Vor- und Nachteile der genannten Speichermedien[1].

	Vorteile	**Nachteile**
Eigener PC	selbstverantwortlich für Sicherheit und Back-up	Einzellösungen aufwendig
	maximale Kontrolle	evtl. fehlende Ressourcen und Know-how
Mobile Speichermedien (USB-Sticks, externe Festplatten etc.)	einfach zu transportieren	Verlust, Diebstahl etc. (besonders unsicher)
	kann im verschließbaren Schrank oder Safe aufbewahrt werden	bei Verlust: Inhalte u. U. ungeschützt
		externe Festplatte: stoß- und verschleißanfällig
Institutionelle Speicherorte (Cloud-Services, virtuelle Laufwerke etc.)	Back-up der Daten ist sichergestellt	Geschwindigkeit evtl. vom Netzwerk abhängig
	professionelle Durchführung und Wartung	Zugriff auf Back-ups evtl. verzögert durch Dienstweg
		unbekannt, welche Sicherheitskriterien und -strategien eingesetzt werden

[1] Quelle der Aufgabe: basierend auf Biernacka, K., Buchholz, P., Danker, S. A., Dolzycka, D., Engelhardt, C., Helbig, K., Jacob, J., Neumann, J., Odebrecht, C., Petersen, B., Slowig, B., Trautwein-Bruns, U., Wiljes, C. & Wuttke, U. (2021). T*rain-the-Trainer Konzept zum Thema Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.5773203.

Externe Speicherorte (Cloud-Services eines kostenlosen oder kostenpflichtigen Anbieters etc.)	einfach zu nutzen und zu verwalten	je nach Anbieter kann die Verbindung auch unsicher sein
	Back-up der Daten ist sichergestellt	abhängig vom Zugang zum Internet (Up- und Downloadgeschwindigkeiten)
	für mobiles Arbeiten nutzbar	Zugriff auf Back-ups evtl. verzögert
	professionelle Durchführung und Wartung	je nach Standort der Server fragwürdiger Datenschutz

Checkliste: Speicher

Leitfragen: Was ist bei der Speicherwahl zu beachten?

- ☐ Wie viel Speicherplatz benötige ich?

  .....

- ☐ Welche Datentypen habe ich und wie häufig werde ich diese ersetzen?

  .....

- ☐ Wer benötigt Zugang und welche Zugriffsrechte soll die Person erhalten?

  .....

- ☐ Ist es notwendig Remote-Zugang zu den Daten zu haben?

  .....

- ☐ Wie wichtig ist schneller Zugriff?

  .....

- ☐ Ist simultaner und synchronischer Zugriff benötigt?

  .....

- ☐ Welche Schritte sollte ich vornehmen um meine Daten vor Verlust zu schützen? (Passwort, Verschlüsselung, physischer Schutz u. A.)

  .....

- ☐ Welche Speicherlösungen sind für personenbezogene Daten geeignet? (falls zutreffend)

  .....

- ☐ Wie häufig werde ich ein Backup machen und wo wird dieser gespeichert?

  .....

- ☐ Wie viel finanzielle Mittel stehen mir zur Verfügung?

  .....

Quelle:
CESSDA Training Working Group. CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC, 2017-2018, https://www.cessda.eu/DMGuide. Das Werk ist lizensiert unter der Creative Commons Attribution-ShareAlike 4.0 International Lizenz.

5.10 Dokumentation und Metadaten

Dokumentation und Metadaten (Kapitel 4.11) sind im FDM, aber auch in der Informatik im Allgemeinen unverzichtbar. Eine ausführliche Dokumentation sollte nicht nur für Software, sondern auch für Forschungsdaten vorgenommen werden und zum Standard gehören. Der adäquate Umgang mit Metadaten sowie eine Nutzung zur Beschreibung von Daten ist insbesondere im FDM eine Herausforderung. Mit dem folgenden Lehrmaterial (siehe Tabelle 5.12) sollen Studierende beim Umgang mit Metadaten unterstützt werden.

Tabelle 5.12: Lehrmaterial Dokumentation und Metadaten

Thema	**Seite**
Arbeitsblatt: Dokumentation	218
Musterlösung: Dokumentation	219
Arbeitsblatt: Metadaten	221
Musterlösung: Metadaten	222

Arbeitsblatt: Dokumentation

Carla führt Interviews mit Studierenden an zwei verschiedenen Universität durch und möchte die Interviewdaten im Anschluss publizieren. Um eine Nachnutzung zu ermöglichen, möchte sie eine ausführliche Dokumentation zu den Daten und ihres Forschungsvorgehens erstellen.

Carla geht in ihrem Forschungsvorhaben rein qualitativ vor und führt Interviews mit Studierenden der Universität Hamburg (UHH) und der University of Cuba (UC) durch. Für jede Universität wird eine gesonderte informierte Einwilligung verfasst. Bei dem Transkriptionsleitfaden handelt es sich um den gleichen Inhalt, aber in deutscher und spanischer Sprache. Carla führt mit den Studierenden beider Universitäten jeweils fünf Interviews durch. Alle Interviews an der UHH finden am 22.04.2022 statt. Die Interviews an der CU werden am 29.04.2022 durchgeführt. Für das Gesamtprojekt wählt Carla den Namen *Collab2022*. Die Dateinamen bezeichnet sie auf Englisch.

1. Kreuzen Sie an, welche Formen der Dokumentation Carla nutzen sollte:

   - ☐ ReadMe-Datei
   - ☐ Data Dictionary
   - ☐ Codebook
   - ☐ Electronic Lab Notebooks bzw. Elektronische Laborbücher
   - ☐ Software-Dokumentation

2. Geben Sie ein beispielhaftes Data Dictionary für Carlas Forschungsvorhaben an.

Dateiname	**Beschreibung**	**Datum**

Musterlösung: Dokumentation

Carla führt Interviews mit Studierenden an zwei verschiedenen Universität durch und möchte die Interviewdaten im Anschluss publizieren. Um eine Nachnutzung zu ermöglichen, möchte sie eine ausführliche Dokumentation zu den Daten und ihres Forschungsvorgehens erstellen.

Carla geht in ihrem Forschungsvorhaben rein qualitativ vor und führt Interviews mit Studierenden der Universität Hamburg (UHH) und der University of Cuba (UC) durch. Für jede Universität wird eine gesonderte informierte Einwilligung verfasst. Bei dem Transkriptionsleitfaden handelt es sich um den gleichen Inhalt, aber in deutscher und spanischer Sprache. Carla führt mit den Studierenden beider Universitäten jeweils fünf Interviews durch. Alle Interviews an der UHH finden am 22.04.2022 statt. Die Interviews an der CU werden am 29.04.2022 durchgeführt. Für das Gesamtprojekt wählt Carla den Namen *Collab2022*. Die Dateinamen bezeichnet sie auf Englisch.

1. Kreuzen Sie an, welche Formen der Dokumentation Carla nutzen sollte:

   ☒ ReadMe-Datei

   ☒ Data Dictionary

   ☒ Codebook

   ☐ Electronic Lab Notebooks bzw. Elektronische Laborbücher

   ☐ Software-Dokumentation

2. Geben Sie ein beispielhaftes Data Dictionary für Carlas Forschungsvorhaben an.

Dateiname	**Beschreibung**	**Datum**
Collab2022_Informed_Consent_UHH	Vorlage: Informierte Einwilligung für die UHH	2022-04-01
Collab2022_Informed_Consent_UC	Vorlage: Informierte Einwilligung für die UC	2022-04-01
Collab2022_Codebook	Erläuterung der verwendeten Variablen und Labels (Codebook)	2022-04-01
Collab2022_ReadMe	ReadMe	22.04.2022
Collab2022_Transcription_Guideline_UHH	Transkriptionsleitfaden UHH	2022-04-10
Collab2022_Transcription_Guideline_UC	Transkriptionsleitfaden UC	2022-04-1
Collab2022_Interview_Guideline_UHH	Interviewleitfaden UHH (auf Deutsch)	2022-04-01

Collab2022_Interview_Guideline_UC	Interviewleitfaden UC (auf Spanisch)	2022-04-01
Collab2022_UHH_t[*Nummer 1-5*]	Interviewtranskripte UHH	2022-04-22
Collab2022_UC_t[*Nummer 1-5*]	Interviewtranskripte UC	2022-04-29

Arbeitsblatt: Metadaten

Carla führt Interviews mit Studierenden durch und möchte neben den gesprochenen Inhalten auch die Stimmung der Studierenden erheben. Es soll ein Algorithmus mit diesen Daten trainiert werden. Dabei vergibt sie verschiedene Tags in den Videos, die Stimmungen repräsentieren.

1. Nennen Sie Funktionen von Metadaten.

   .....

   .....

   .....

2. Nennen Sie die Metadaten, die in diesem Fall vergeben werden können.

   .....

   .....

   .....

3. Erstellen Sie ein codemeta.json-File für eine Software Ihrer Wahl.[1]

[1] Quelle der Aufgabe: TU9-FDM. (2019). Software als Forschungsdaten. https://doi.org/10.5281/zenodo.2611303

Musterlösung: Metadaten

Carla führt Interviews mit Studierenden durch und möchte neben den gesprochenen Inhalten auch die Stimmung der Studierenden erheben. Es soll ein Algorithmus mit diesen Daten trainiert werden. Dabei vergibt sie verschiedene Tags in den Videos, die Stimmungen repräsentieren.

1. Nennen Sie Funktionen von Metadaten.

   - dienen der Beschreibung der Daten und erhöhen somit die Auffindbarkeit
   - ermöglichen Maschinenlesbarkeit
   - vereinfachen die Verwaltung der Daten
   - erhöhen die Auswertbarkeit und Verständlichkeit der Daten

2. Nennen Sie die Metadaten, die in diesem Fall vergeben werden können.

   - Datum der Interviewdurchführung
   - Alter
   - Geschlecht
   - Hochschule
   - Stimmung
   - besprochene Themen in Stichworten
   - Aufmerksamkeit
   - Motivation

3. Erstellen Sie ein codemeta.json-File für eine Software Ihrer Wahl.[1]

   Ein Beispiel für ein codemeta.json-File kann auf https://github.com/codemeta/codemeta/blob/master/examples/example-code-jsonld.json (archivierte Version: https://perma.cc/3VEB-QRU4) *gefunden werden.*

[1] Quelle der Aufgabe: U9-FDM. (2019). Software als Forschungsdaten. https://doi.org/10.5281/zenodo.2611303

5.11 Zugriffssicherheit

Das Thema Zugriffssicherheit ist von besonderer Bedeutung für den Schutz von Forschungsdaten. Relevante Inhalte werden in Kapitel 4.12 erläutert und umfassen beispielsweise den Passwortschutz sowie die Regelung von Zugriffsrechten. Zu beiden Themen werden Arbeitsblätter und ihre Musterlösungen präsentiert (siehe Tabelle 5.14).

Tabelle 5.14: Lehrmaterial Zugriffssicherheit

Thema	**Seite**
Arbeitsblatt: Passwortschutz	224
Musterlösung: Passwortschutz	225
Arbeitsblatt: Zugriffsrechte	226
Musterlösung: Zugriffsrechte	228

Arbeitsblatt: Passwortschutz

Timo arbeitet in seinem Open-Source-Projekt gemeinsam mit Kayla sowie mit beteiligten Krankenkassen, um eine App für Patient:innen zu implementieren. Teilweise müssen alle Akteur:innen auf die gleichen Daten zugreifen können, wobei abgesichert werden muss, dass keine Dritten auf die Daten zugreifen können. Zunächst müssen zum Schutz der Daten sichere Passwörter vergeben werden.

1. Erläutern Sie bei den folgenden Passwörtern, inwieweit sie laut Bundesamt für Sicherheit in der Informationstechnik (BSI) als sicher eingestuft werden können.

   (a) kSm03um,#x299H1+xwiwoh28ß´d2s0ßwelaßi!

   .....

   .....

   (b) OurWholeUniverseWasInAHot,DenseState2007

   .....

   .....

   (c) OWUWIAHDS2007

   .....

   .....

   (d) 0wUw14Hd5+2007:

   .....

   .....

Musterlösung: Passwortschutz

Timo arbeitet in seinem Open-Source-Projekt gemeinsam mit Kayla sowie mit beteiligten Krankenkassen, um eine App für Patient:innen zu implementieren. Teilweise müssen alle Akteur:innen auf die gleichen Daten zugreifen können, wobei abgesichert werden muss, dass keine Dritten auf die Daten zugreifen können. Zunächst müssen zum Schutz der Daten sichere Passwörter vergeben werden.

1. Erläutern Sie bei den folgenden Passwörtern, inwieweit sie laut Bundesamt für Sicherheit in der Informationstechnik (BSI) als sicher eingestuft werden können.

   (a) kSm03um,#x299H1+xwiwoh28ß´d2s0ßwelaßi!

   + kein einheitliches Passwort; Verwendung aller verfügbarer Zeichen; Länge; kommt nicht im Wörterbuch vor
   - nicht gut zu merken

   (b) OurWholeUniverseWasInAHot,DenseState2007

   + kein einheitliches Passwort; gut zu merken; Länge
   - keine Verwendung aller verfügbaren Zeichen; einzelne Wörter kommen im Wörterbuch vor

   (c) OWUWIAHDS2007

   + kein einheitliches Passwort; (als Passsatz) gut zu merken; Länge; Wörter kommen nicht im Wörterbuch vor
   - keine Verwendung aller verfügbaren Zeichen

   (d) 0wUw14Hd5+2007:

   + kein einheitliches Passwort; (als Passsatz) gut zu merken; Länge; Wörter kommen nicht im Wörterbuch vor; Verwendung aller verfügbaren Zeichen
   - Ersetzungsregeln für Buchstaben müssen merkbar sein

Arbeitsblatt: Zugriffsrechte

Timo arbeitet in seinem Ope-Source-Projekt gemeinsam mit Kayla sowie mit beteiligten Krankenkassen. Teilweise müssen alle Akteur:innen auf die gleichen Daten zugreifen können, wobei das nicht für alle Daten gilt. Bevor die Arbeit beginnen kann, muss entschieden werden, welche Akteur:innen Zugriff auf welche Daten bekommen.

1. Timo und Kayla entwickeln eine App, die Patient:innendaten verarbeiten soll. Die App wird speziell für Krankenkassen entwickelt, die ihren Kund:innen weitere Services bieten möchten. Die Patient:innendaten werden bereits während der Entwicklung benötigt, um einige Funktionen testen zu können. Beispielsweise soll ein Ärzt:innenwechsel erleichtert werden, indem die Patient:innendaten übertragen werden. Es handelt sich dabei um personenbezogene und sensible Daten. Kreuzen Sie an, welche Zugriffsrechte auf die Patient:innendaten gestattet werden sollte.

	Timo	Kayla	Krankenkasse
Lesezugriff	☐	☐	☐
Schreibzugriff	☐	☐	☐
Ausführungszugriff	☐	☐	☐

   Begründen Sie Ihre Einschätzung.

   .....

   .....

   .....

   .....

   .....

2. Die Patient:innendaten sollen auf einem Repositorium abgelegt und aufbewahrt werden. Welcher Zugang kann hierfür gewährt werden?

   ☐ offen

   ☐ eingeschränkt

   ☐ geschlossen

   Begründen Sie Ihre Einschätzung.

   .....

   .....

   .....

   .....

3. Timo und Kayla entwickeln eine App, die Patient:innendaten verarbeiten soll. Beide arbeiten an derselben (bereits existierenden) App, kümmern sich jedoch um unterschiedliche Schnittstellen. Die App wird im Rahmen eines Open-Source-Projektes weiterentwickelt und nach der ersten Implementation auf GitHub veröffentlicht. In den beteiligten Krankenkassen sind ebenfalls Softwareentwickler:innen tätig, die die Schnittstellen zu ihren Datenbanken und Systemen programmieren sollen. Kreuzen Sie an, welche Zugriffsrechte den folgenden Akteur:innen auf den Quellcode von Timos App gegeben werden sollte.

	Timo	Kayla	Krankenkasse	Öffentlichkeit
Lesezugriff	☐	☐	☐	☐
Schreibzugriff	☐	☐	☐	☐
Ausführungszugriff	☐	☐	☐	☐

Begründen Sie Ihre Einschätzung.

.....

.....

.....

.....

Musterlösung: Zugriffsrechte

Timo arbeitet in seinem Open-Source-Projekt gemeinsam mit Kayla sowie mit beteiligten Krankenkassen. Teilweise müssen alle Akteur:innen auf die gleichen Daten zugreifen können, wobei das nicht für alle Daten gilt. Bevor die Arbeit beginnen kann, muss entschieden werden, welche Akteur:innen Zugriff auf welche Daten bekommen.

1. Timo und Kayla entwickeln eine App, die Patient:innendaten verarbeiten soll. Die App wird speziell für Krankenkassen entwickelt, die ihren Kund:innen weitere Services bieten möchten. Die Patient:innendaten werden bereits während der Entwicklung benötigt, um einige Funktionen testen zu können. Beispielsweise soll ein Ärzt:innenwechsel erleichtert werden, indem die Patient:innendaten übertragen werden. Es handelt sich dabei um personenbezogene und sensible Daten. Kreuzen Sie an, welche Zugriffsrechte auf die Patient:innendaten gestattet werden sollte.

	Timo	Kayla	Krankenkasse
Lesezugriff	☒	☒	☒
Schreibzugriff	☐	☐	☒
Ausführungszugriff	nicht zutreffend	nicht zutreffend	nicht zutreffend

   Begründen Sie Ihre Einschätzung.

   Da die Daten selbst von der Krankenkasse gesammelt und dann Timo und Kayla zur Verfügung gestellt werden, hat die Krankenkasse automatisch Lese- und Schreibzugriff auf die Daten. Da Timo und Kayla diese Daten benötigen, um die App zu entwickeln, werden ihnen Leserechte gewährt. Der Schreibzugriff ist nicht notwendig, da die Daten nicht verändert werden sollen. Da es sich bei den Patient:innendaten nicht um eine ausführbare Software handelt, ist der Ausführungszugriff für diesen Fall irrelevant.

2. Die Patient:innendaten sollen auf einem Repositorium abgelegt und aufbewahrt werden. Welcher Zugang kann hierfür gewährt werden?

   - ☐ offen
   - ☐ eingeschränkt
   - ☒ geschlossen

   Begründen Sie Ihre Einschätzung.

   Der Zugriff auf die Patient:innendaten darf der breiten Öffentlichkeit nicht gewährt werden, da es sich um sensible Daten handelt.

3. Timo und Kayla entwickeln eine App, die Patient:innendaten verarbeiten soll. Beide arbeiten an derselben (bereits existierenden) App, kümmern sich jedoch um unterschiedliche Schnittstellen. Die App wird im Rahmen eines Open-Source-Projektes weiterentwickelt und nach der ersten Implementation auf GitHub veröffentlicht. In den beteiligten Krankenkassen sind ebenfalls Softwareentwickler:innen tätig, die die Schnittstellen zu ihren Datenbanken und Systemen programmieren sollen. Kreuzen Sie an, welche Zugriffsrechte den folgenden Akteur:innen auf den Quellcode von Timos App gegeben werden sollte.

	Timo	Kayla	Krankenkasse	Öffentlichkeit
Lesezugriff	☒	☒	☒	☒
Schreibzugriff	☒	☒	☒	☒
Ausführungszugriff	☒	☒	☒	☒

Begründen Sie Ihre Einschätzung.

Die Antwort hängt von der gewählten Open-Source-Lizenz ab. Angenommen, es wird ein permissive Lizenz gewählt, dürfen alle Beteiligten die App sowohl lesen, verändern als auch ausführen.

5.12 Publikation von Forschungsdaten

Die Publikation von Forschungsdaten ist für eine transparente Forschung und für eine mögliche Nachnutzung essenziell. In Kapitel 4.13 werden beispielsweise Publikationswege und geeignete Repositorien vorgestellt. Auch auf Spezifika für die Informatik wird eingegangen. Anhand des folgenden Materials (siehe Tabelle 5.15) können die Studierenden konkrete Aufgaben zur Publikation von Forschungsdaten bearbeiten.

Tabelle 5.15: Lehrmaterial Publikation von Forschungsdaten

Arbeitsblatt: Publikation von Forschungsdaten

Carla möchte ihre Interviewdaten publizieren, damit weitere Forschende die Daten nutzen können. Sie ist sich einerseits unsicher, wie sie an die Datenpublikation herangehen soll und welche Konsequenzen das andererseits für sie und ihre zukünftige Arbeit in diesem Bereich haben könnte.

1. Nennen und begründen Sie, welches Repositorium Sie Carla empfehlen würden. Nutzen Sie für Ihre Suche re3data.org.

   .....

   .....

2. Erläutern Sie den Nutzen, den ein persistenter Identifikator für die Referenzierung eines Datensatzes bietet.

   .....

   .....

3. Formulieren Sie Gegenargumente.

Gründe, um Forschungsdaten nicht zu publizieren	**Gegenargumente**
Wenn ich meine Forschungsdaten publiziere, dann könnte jemand mir zuvorkommen und vor mir Erkenntnisse veröffentlichen, die auf meinen Daten basieren.	
Meine Daten sind sehr sensibel. Ich kann sie nicht für andere verfügbar machen.	
Wenn ich meine Daten publiziere, dann wird meine Forschung völlig transparent und selbst kleinste Fehler werden offenbart.	
Veröffentlichte Daten bringen keinen weiteren Nutzen.	
Ich habe versprochen, meine Daten nach Projektende zu vernichten.	

Musterlösung: Publikation von Forschungsdaten

Carla möchte ihre Interviewdaten publizieren, damit weitere Forschende die Daten nutzen können. Sie ist sich einerseits unsicher, wie sie an die Datenpublikation herangehen soll und welche Konsequenzen das andererseits für sie und ihre zukünftige Arbeit in diesem Bereich haben könnte.

1. Nennen und begründen Sie, welches Repositorium Sie Carla empfehlen würden. Nutzen Sie für Ihre Suche re3data.org.

   Da die von Carla bearbeitete Thematik nicht nur in den Bereich der Informatik fällt und es sich bei den Daten um qualitative Daten handelt, wäre z. B. *Qualiservice*[1] für die Publikation dieser Daten geeignet. Neben der persönlichen Beratung, der Kuratierung und Aufbereitung der Daten wird auch die Langzeitarchivierung angeboten. Zu den Services gehört unter anderem die Beratung zu personenbezogenen Daten und den Anonymisierungsverfahren bzw. den verschiedenen Möglichkeiten, diese Daten abzulegen und anderen Forschenden zur Verfügung zu stellen.

2. Erläutern Sie den Nutzen, den ein persistenter Identifikator für die Referenzierung eines Datensatzes bietet.

   - Referenzierbarkeit
   - Zitierfähigkeit
   - Persistenz
   - Neutralität (enthalten nicht zwingend semantische Hinweise auf die Domain)

[1] https://www.qualiservice.org/de/. Archivierte Version: https://perma.cc/3JZ8-LYGX.

3. Formulieren Sie Gegenargumente.

Gründe, um Forschungsdaten nicht zu publizieren	**Gegenargumente**
Wenn ich meine Forschungsdaten publiziere, dann könnte jemand mir zuvorkommen und vor mir Erkenntnisse veröffentlichen, die auf meinen Daten basieren.	Die Datenpublikation hindert Sie nicht daran, zuerst Ihre Erkenntnisse zu veröffentlichen. Die meisten Forschungsförderer erlauben Ihnen eine gewisse Zeit der alleinigen Nutzung, wollen aber auch eine zeitnahe Publikation. Denken Sie auch daran, dass Sie bereits seit einiger Zeit mit Ihren Daten arbeiten und somit die Daten besser kennen als jeder andere, der sie neu verwenden möchte. Wenn Sie dennoch Zweifel haben, können Sie ein Embargo für Ihre Daten setzen. Darüber hinaus können Sie Ihr Forschungsvorhaben beim Open Science Framework[1] registrieren und somit immer nachweisen, dass Sie Ideengebende sind.
Meine Daten sind sehr sensibel. Ich kann Sie nicht für andere verfügbar machen.	Für die Publikation von sensiblen Daten kann im Vorfeld eine Einverständniserklärung von den Betroffenen eingeholt werden. Auch eine Anonymisierung kann dabei behilflich sein, um personenbezogene Daten zu schützen. Falls es dennoch unmöglich sein sollte, die sensiblen Daten zu publizieren, kann der Zugriff zu den Daten kontrolliert oder ein Embargo aufgesetzt werden. Es können auch nur die Metadaten öffentlich zugänglich gemacht werden und der Zugriff zu den eigentlichen Daten kann eingeschränkt werden.

[1] https://osf.io/. Archivierte Version: https://perma.cc/9BLL-YEKV.

Wenn ich meine Daten publiziere, dann wird meine Forschung völlig transparent und selbst kleinste Fehler werden offenbart.	In der Tat wird Ihre Forschung durch die Datenpublikation transparenter und eventuelle Fehler können so auch frühzeitig erkannt werden. Auf diese Weise kann Ihre Forschung an Qualität gewinnen, und vielleicht ergeben sich auch neue Kooperationen.
Veröffentlichte Daten bringen keinen weiteren Nutzen.	Wissenschaftler:innen möchten auf Daten aus verschiedenen Studien, Methodenbereichen und Publikationen zugreifen. Es ist schwierig vorherzusagen, welche Daten für die zukünftige Forschung wichtig sein können. Hätten Sie gedacht, dass Aufzeichnungen eines Amateurgärtners eines Tages essenzielle Erkenntnisse für die Forschung zum Thema Klimawandel beitragen könnten? Ein weiteres Beispiel für fachübergreifende Nachnutzung von Daten ist: Daniel Steiner, Heinz J. Zumbühl, und Andreas Bauder. *Two Alpine glaciers over the past two centuries: A scientific view based on pictorial sources*. 2008. Solange Sie Ihre Daten gut dokumentieren und Kontextinformationen dazu liefern, können Ihre Daten verstanden und nachgenutzt werden.
Ich habe versprochen, meine Daten nach Projektende zu vernichten.	Warum geben Sie solche Versprechen? Vermeiden Sie diese Art von Versprechen. Üblicherweise besteht weder eine rechtliche noch ethische Notwendigkeit, die Daten zu vernichten, mit Ausnahme von personenbezogenen Daten.

Checkliste: Wie publiziere ich Forschungsdaten?

WIE PUBLIZIERE ICH FORSCHUNGSDATEN?

DOKUMENTIERE DIE DATEN

Dokumentiere stets Deine Daten vom Beginn der Forschungsarbeit an, um die Daten nachvollziehbar zu gestalten und vergebe darüber hinaus relevante Metadaten. Halte Dich dabei an fachspezifische Metadatenstandards.

Weitere Informationen:
https://tinyurl.com/FDdoku

WÄHLE EIN REPOSITORIUM

Suche nach einem geeigneten, fachspezifischen und für Deine Community relevanten Repositorium. Falls Du nicht fündig wirst, wähle ein fachübegreifendes oder ein institutionelles Repositorium.

Weitere Informationen:
www.re3data.org

VERGEBE EINE LIZENZ

Wähle eine für Deine Forschungsdaten geeignete Lizenz (z. B. Creative Commons). Versuche dabei die Nachnutzungsbedingungen so offen wie möglich und so geschlossen wie nötig zu halten.

Weitere Informationen:
https://creativecommons.org
https://choosealicense.com

PERSISTENTE IDENTIFIER

Achte darauf, dass Deine Daten eine DOI erhalten, um sie langfristig auffindbar zu machen. Erstelle für Dich eine ORCID damit Dir Deine wissenschaftlichen Arbeiten eindeutig zugewiesen werden können.

Weitere Informationen:
www.doi.org
https://orcid.org

RECHTLICHE ASPEKTE

Der Veröffentlichung von Forschungsdaten können verschiedene rechtliche und/oder ethische Aspekte entgegenstehen. Überprüfe dies vor der Publikation.

Weitere Informationen:
https://tinyurl.com/FDrecht

PUBLIZIERE

Lade Deine Forschungsdaten in einem geeigneten Dateiformat auf das gewählte Repositorium hoch und lass es die Welt wissen!
Bei Fragen stehen Dir die Mitarbeiter des Repositoriums gerne zur Verfügung.

Weitere Informationen:
https://tinyurl.com/dateiformate

FDMentor
Erstellt im Rahmen des FDMentor-Projektes
Projektlaufzeit: 1. Mai 2017 bis 30. April 2019
Idee und Gestaltung: Katarzyna Biernacka, Dr. Dominika Dolzycka, Petra Buchholz
Kontakt: fdmentor@hu-berlin.de
Twitter: @fd_mentor
https://hu.berlin/fdmentor

gefördert vom: Bundesministerium für Bildung und Forschung

Checkliste: Datenqualität (Seite 1)[1]

iRights.Lab
Think Tank für die digitale Welt

Checkliste für eine bessere Datenqualität

Was können Sie mit der Checkliste erreichen?

- Häufige Datenqualitätsprobleme vorausschauend erkennen
- Anregung für eine Datenqualitätsstrategie erhalten
- Strategisches Vorgehen bei Datenqualitätsproblemen als Teil einer Data Governance implementieren

15 Dimensionen zur Messung von Daten- bzw. Informationsqualität

Diese 15 Dimensionen können Sie dabei unterstützen, Datenqualität systematisch zu beurteilen.

☐	**Zugänglichkeit** (accessibility) Kann ich die Daten auf einfache Weise abrufen?	System
☐	**Bearbeitbarkeit** (ease of manipulation) Ist es möglich, dass ich die Daten ändern/transformieren kann?	
☐	**Glaubwürdigkeit** (believability) Sind den Daten weitere Informationen (z.B. Zertifikate oder ausführliche Berichte über Datengewinnung und -aufbereitung) beigefügt, die Vertrauen in die Datenqualität erwecken?	Inhalt
☐	**Hohes Ansehen** (reputation) Steht die Datenquelle für kontinuierlich qualitativ hochwertige Daten? (ähnlich Glaubwürdigkeit)	
☐	**Objektivität** (objectivity) Enthalten die Informationen keine Wertungen und sind sachlich?	
☐	**Fehlerfreiheit** (free of error) Entsprechen die Informationen dem, was sie abbilden sollen?	
☐	**Aktualität** (timeliness) Sind die Informationen so aktuell, wie es der Zweck erfordert?	Nutzung
☐	**Wertschöpfung** (value-added) Kann ich aus den Daten einen gesellschaftlichen oder wirtschaftlichen Wert schöpfen?	
☐	**Vollständigkeit** (completeness) Sind alle notwendigen Datenfelder vollständig gefüllt?	
☐	**Angemessener Umfang** (appropriate amount of data) Reicht mir der Umfang der Daten aus, um meine Frage zu beantworten?	
☐	**Relevanz** (relevancy) Sind die Informationen für meinen Zweck notwendig?	
☐	**Übersichtlichkeit** (concise representation) Kann ich auf die benötigten Informationen innerhalb eines Datensatzes einfach zugreifen? (Zugänglichkeit innerhalb eines Datensatzes)	Darstellung
☐	**Einheitliche Darstellung** (consistent representation) Entsprechen die Informationen fortlaufend einem festgelegten Format?	
☐	**Verständlichkeit** (understandability) Sind die Informationen inhaltlich einfach zugänglich?	
☐	**Eindeutige Auslegbarkeit** (interpretability) Wenn verschiedene Personen mit den Daten unabhängig voneinander arbeiten, werden sie die Daten in der gleichen (korrekten) Weise interpretieren?	

Checkliste für eine bessere Datenqualität mFUND-Begleitforschung zu Data Governance

[1] Quelle: iRights-Lab. (2019). Checkliste für eine bessere Datenqualität.

Checkliste: Datenqualität (Seite 2)

iRights.Lab
think tank für die digitale Welt

Anmerkungen zu den 15 Dimensionen

- Die vorgestellten 15 Dimensionen basieren auf einer Studie der MIT-Wissenschaftler*innen Wang und Strong von 1996[1]. Sie sind eine gängige Form, Datenqualität zu bestimmen. Natürlich existieren auch noch andere Herangehensweisen.
- In den seltensten Fällen sind alle Dimensionen relevant. Vielmehr sollten Sie eine Auswahl aus den 15 Dimensionen entsprechend der Sinnhaftigkeit und dem Zweck treffen.

Von 15 Dimensionen zu einer Strategie

- Sie können die 15 Dimensionen zum losen Reflektieren von vorliegender Datenqualität oder möglichen Datenqualitätsproblemen bei noch unbekannten Datenquellen nutzen. Sie können die Dimensionen auch noch näher definieren und daraufhin Datenqualität messen und eine Bestandsaufnahme durchführen.
- Eine Messung ist empfehlenswert, weil Sie damit Veränderungen nachverfolgen können und entsprechende Maßnahmen treffen können.
- Je mehr Sie von ad-hoc-Maßnahmen zu einer ganzheitlichen Strategie wechseln, desto mehr müssen Sie Prozesse zur Datenqualitätssicherung verstetigen. Damit sind Sie auf dem besten Wege, eine Data Governance zu implementieren.
- Eine Data Governance bezeichnet ein klar definiertes Konzept, dass den Umgang mit Daten innerhalb einer Organisation regelt. Das können zum einen Prozesse zur Datenqualität, aber eben auch Datenschutz und Datensicherheit sein.
- Folgende Schritte empfehlen wir zu einer Data Governance für eine höhere Datenqualität:
    - Ein innerhalb der Organisation einheitliches Verständnis des Kernprozesses bei Datenqualitätsproblemen festlegen und verstetigen (standardisieren)
    - Vorhandene Datenstandards verwenden und Metadaten pflegen
    - Alle betroffenen Stellen in der Organisation einbeziehen und Verantwortlichkeiten klar benennen
    - Prozesse transparent und nachvollziehbar machen (z.B. durch eine Versionskontrolle bei Datensatzänderungen)
    - Auswahl der Bewertungsdimensionen stetig hinterfragen und ggf. anpassen
    - Regelmäßig Datenqualität messen

- Metadaten: Metadaten geben mehr Informationen zum Datensatz, werden jedoch leider häufig vernachlässigt. Metadaten für eigene Datensätze zu pflegen, sollten Sie hoch priorisieren. Damit sichern Sie ab, dass man auch nach einiger Zeit einen qualitativ hochwertigen Datensatz ohne große Mehrarbeit verwenden kann. Bei externen Datensätzen müssen Sie davon ausgehen, dass diese nicht ihren eigenen Qualitätsstandards entsprechen. Das gilt auch für die entsprechenden Metadaten.

mFUND-Begleitforschung zu Data Governance
Das iRights.Lab führt ein Begleitforschungsprojekt zu „Data-Governance im Innovationsprozess“ durch. Ziel ist es, dass sich Organisationen, die an datengetriebenen Innovationen arbeiten, selbstständig und frühzeitig eine Data-Governance auferlegen, um Aspekte wie Datenschutz, Datensicherheit und Datenqualität zu gewährleisten und als Wettbewerbsvorteil nutzen zu können.
www.irights-lab.de

Gefördert durch:
Bundesministerium für Verkehr und digitale Infrastruktur
aufgrund eines Beschlusses des Deutschen Bundestages

[1] Richard Y. Wang, Diane M. Strong: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12:4, 5–34, 1996.

5.13 Urheberrecht und Lizenzierung

In Kapitel 4.14 wurde auf verschiedene Aspekte des Urheberrechts und der Lizenzierung eingegangen. Insbesondere, aber nicht ausschließlich, ist die Lizenzierung von Software für die Informatik relevant und sollte von Studierenden angewandt werden können. Übungen für diesen Themenbereich sind in Tabelle 5.17 zusammengefasst.

Tabelle 5.17: Lehrmaterial Urheberrecht und Lizenzierung

Arbeitsblatt: Softwarelizenzierung

Alex stellt fest, dass existierende Algorithmen zur Analyse von Twitterdaten nicht akkurat genug für seinen Anwendungsfall sind. Deshalb entwickelt Alex einen eigenen Programmcode. Des Weiteren entwickelt Alex Erklärvideos, die die Funktionsweise des Algorithmus erklären. Nun ist sich Alex unsicher, unter welcher Lizenz die erstellten Materialien zur Verfügung gestellt werden könnten.

1. Erklären Sie, was man unter einer Open-Source-Lizenz versteht.

   .....

   .....

2. Nennen Sie, welche Lizenzen Alex bei den verschiedenen Szenarien vergeben kann.

   (a) Welche Lizenz kann Alex vergeben, wenn der selbstentwickelte Programmcode einem bestimmten Kunden verkauft werden soll?

   .....

   (b) Welche Lizenz kann Alex vergeben, um den selbstentwickelten Programmcode für alle offen verfügbar und nachnutzbar zu machen?

   .....

   (c) Welche Lizenz kann Alex vergeben, wenn Alex einen bereits existierenden Code weiterentwickeln möchte, der unter der GPL-Lizenz steht?

   .....

3. Kreuzen Sie an, welche Lizenz gewählt werden sollte, um[1]

   (a) die Ausbreitung zu maximieren und Restriktionen zu minimieren.

   - ☐ Permissive Lizenz
   - ☐ Copyleft-Lizenz
   - ☐ Open-Source-Lizenz
   - ☐ Proprietäre Lizenz
   - ☐ mehrere Lizenzen

[1] Quelle der Aufgabe: TU9-FDM. (2019). Software als Forschungsdaten. https://doi.org/10.5281/zenodo.2611303

(b) sicherzustellen, dass alle Ableger auch Open Source sind.

- ☐ Permissive Lizenz
- ☐ Copyleft-Lizenz
- ☐ Open-Source-Lizenz
- ☐ Proprietäre Lizenz
- ☐ mehrere Lizenzen

(c) Open Source für die akademische Welt zu ermöglichen, jedoch Geld an der Wirtschaft zu verdienen.

- ☐ Permissive Lizenz
- ☐ Copyleft-Lizenz
- ☐ Open-Source-Lizenz
- ☐ Proprietäre Lizenz
- ☐ mehrere Lizenzen

(d) den Quellcode zu schützen (maximale Kontrolle).

- ☐ Permissive Lizenz
- ☐ Copyleft-Lizenz
- ☐ Open-Source-Lizenz
- ☐ Proprietäre Lizenz
- ☐ mehrere Lizenzen

Musterlösung: Software-Lizenzierung

Alex stellt fest, dass existierende Algorithmen zur Analyse von Twitterdaten nicht akkurat genug für seinen Anwendungsfall sind. Deshalb entwickelt Alex einen eigenen Programmcode. Des Weiteren entwickelt Alex Erklärvideos, die die Funktionsweise des Algorithmus erklären. Nun ist sich Alex unsicher, unter welcher Lizenz die erstellten Materialien zur Verfügung gestellt werden könnten.

1. Erklären Sie, was man unter einer Open-Source-Lizenz versteht.

   Unter einer Open-Source-Lizenz versteht man Nutzungsbedingungen, die einen Quellcode frei zugänglich machen und somit den Nachnutzenden ermöglichen, diesen zu nutzen und zu verändern.

2. Nennen Sie, welche Lizenzen Alex bei den verschiedenen Szenarien vergeben kann.

   (a) Welche Lizenz kann Alex vergeben, wenn der selbstentwickelte Programmcode einem bestimmten Kunden verkauft werden soll?

   Alex kann in diesem Fall eine proprietäre Lizenz vergeben.

   (b) Welche Lizenz kann Alex vergeben, um den selbstentwickelten Programmcode für alle offen verfügbar und nachnutzbar zu machen?

   Um den Programmcode möglichst offen und nachnutzbar zu machen, kann er Open-Source-Lizenzen wählen, z. B. die MIT-Lizenz oder die Apache License 2.0.

   (c) Welche Lizenz kann Alex vergeben, wenn Alex einen bereits existierenden Code weiterentwickeln möchte, der unter der GPL-Lizenz steht?

   In diesem Fall würde Alex auch eine GPL-Lizenz vergeben dürfen.

3. Kreuzen Sie an, welche Lizenz gewählt werden sollte, um[1]

   (a) die Ausbreitung zu maximieren und Restriktionen zu minimieren.

   - ☒ Permissive Lizenz
   - ☐ Copyleft-Lizenz
   - ☐ Open-Source-Lizenz
   - ☐ Proprietäre Lizenz
   - ☐ mehrere Lizenzen

[1] Quelle der Aufgabe: TU9-FDM. (2019). Software als Forschungsdaten. https://doi.org/10.5281/zenodo.2611303

(b) sicherzustellen, dass alle Ableger auch Open Source sind.

- ☐ Permissive Lizenz
- ☒ Copyleft-Lizenz
- ☐ Open-Source-Lizenz
- ☐ Proprietäre Lizenz
- ☐ mehrere Lizenzen

(c) Open Source für die akademische Welt zu ermöglichen, jedoch Geld an der Wirtschaft zu verdienen.

- ☐ Permissive Lizenz
- ☐ Copyleft-Lizenz
- ☐ Open-Source-Lizenz
- ☐ Proprietäre Lizenz
- ☒ mehrere Lizenzen

(d) den Quellcode zu schützen (maximale Kontrolle).

- ☐ Permissive Lizenz
- ☐ Copyleft-Lizenz
- ☐ Open-Source-Lizenz
- ☒ Proprietäre Lizenz
- ☐ mehrere Lizenzen

Arbeitsblatt: Creative-Commons-Lizenzierung

Alex hat bereits erste Forschungsdaten gesammelt und möchte seine Ergebnisse auf einer Konferenz publizieren. Auch für die Anwendung der Erkenntnisse für Lehrzwecke hat Alex einen ausführlichen Leitfaden geschrieben und möchte diesen auf der Webseite des Lehrstuhls veröffentlichen. Nun muss sich Alex für geeignete Lizenzen entscheiden.

1. Für die Publikation des Forschungsbeitrags dürfen die Autor:innen zwischen verschiedenen CC-Lizenzen wählen. Alex ist es dabei wichtig, dass die Arbeit nicht kommerziell genutzt werden darf, sie darf aber gern unter der Namensnennung der Urheber:innen verändert werden. Kreuzen Sie an, welche Lizenz dafür gewählt werden muss.

   - ☐ CC BY
   - ☐ CC BY-SA
   - ☐ CC BY-NC-SA
   - ☐ CC BY-ND
   - ☐ CC BY-NC

2. Das erstellte Lehrmaterial soll von allen Interessierten genutzt werden können, jedoch möchte Alex nicht auf die Namensnennung verzichten. Kreuzen Sie an, welche CC-Lizenz dafür geeignet ist.

   - ☐ CC BY
   - ☐ CC BY-SA
   - ☐ CC BY-NC-SA
   - ☐ CC BY-ND
   - ☐ CC BY-NC

3. Kreuzen Sie an, welche der CC-Lizenzen als *offen* im Sinne von Open Science gelten.

   - ☐ CC0
   - ☐ CC BY-NC
   - ☐ CC BY-ND
   - ☐ CC BY-NC-SA
   - ☐ CC BY-SA
   - ☐ CC BY

4. Erklären Sie den Unterschied zwischen *lizenzfrei* und *freie Lizenz*.

   .....

   .....

Musterlösung: Creative-Commons-Lizenzierung

Alex hat bereits erste Forschungsdaten gesammelt und möchte seine Ergebnisse auf einer Konferenz publizieren. Auch für die Anwendung der Erkenntnisse für Lehrzwecke hat Alex einen ausführlichen Leitfaden geschrieben und möchte diesen auf der Webseite des Lehrstuhls veröffentlichen. Nun muss sich Alex für geeignete Lizenzen entscheiden.

1. Für die Publikation des Forschungsbeitrags dürfen die Autor:innen zwischen verschiedenen CC-Lizenzen wählen. Alex ist es dabei wichtig, dass die Arbeit nicht kommerziell genutzt werden darf, sie darf aber gern unter der Namensnennung der Urheber:innen verändert werden. Kreuzen Sie an, welche Lizenz dafür gewählt werden muss.

   - ☐ CC BY
   - ☐ CC BY-SA
   - ☐ CC BY-NC-SA
   - ☐ CC BY-ND
   - ☒ CC BY-NC

2. Das erstellte Lehrmaterial soll von allen Interessierten genutzt werden können, jedoch möchte Alex nicht auf die Namensnennung verzichten. Kreuzen Sie an, welche CC-Lizenz dafür geeignet ist.

   - ☒ CC BY
   - ☐ CC BY-SA
   - ☐ CC BY-NC-SA
   - ☐ CC BY-ND
   - ☐ CC BY-NC

3. Kreuzen Sie an, welche der CC-Lizenzen als *offen* im Sinne von Open Science gelten.

   - ☒ CC0
   - ☐ CC BY-NC
   - ☐ CC BY-ND
   - ☐ CC BY-NC-SA
   - ☒ CC BY-SA
   - ☒ CC BY

4. Erklären Sie den Unterschied zwischen *lizenzfrei* und *freie Lizenz*.

   Lizenzfrei bedeutet, dass keine Lizenz vergeben wurde bzw. unauffindbar ist. Falls diese Daten dem Urheberrecht unterliegen, gilt in diesem Fall das Urheberrechtsgesetz. Dieses ist in Deutschland ziemlich strikt, wodurch eine Nachnutzung der Daten erschwert sein könnte. Eine freie Lizenz hingegen wird frei von den Ersteller:innen der Daten vergeben und öffnet somit potenziell die Nachnutzungsmöglichkeiten der Daten. Im besten Fall erlaubt sie eine freie Nutzung der Daten ohne weitere Einschränkungen.

Arbeitsblatt: Urheberrecht

Timo schreibt im Zuge seiner Masterarbeit Software, die er nach der Fertigstellung veröffentlichen möchte. Ihm fällt auf, dass nicht nur die Softwarelizenzierung eine wichtige Rolle spielt. Er fragt sich unter anderem, wem die Daten aus dem Open-Source-Projekt gehören.

1. Timo schreibt seine Masterarbeit an der Technischen Universität Dresden und engagiert sich ehrenamtlich bei einem Open-Source-Projekt. Kreuzen Sie an, bei wem das Urheberrecht von Timos Code liegt.

   ☐ Timo

   ☐ Technsiche Universität Dresden

   ☐ Projektleiter:in des Open-Source-Projekts

2. Erläutern Sie, ob es möglich ist das Urheberrecht zu übertragen bzw. abzutreten.

   .....

   .....

3. Kreuzen Sie an, in welchem Fall das Nutzungsrecht bei der Universität liegen würde.

   ☐ Nie

   ☐ Wenn Timo es ausdrücklich überträgt

   ☐ Bei einem Arbeitsverhältnis mit der Universität

4. Kreuzen Sie an, in welchem Fall das Nutzungsrecht bei der:dem Projektleiter:in des Open-Source-Projekts liegen würde.

   ☐ Nie

   ☐ Wenn Timo es ausdrücklich überträgt

   ☐ Bei einem Arbeitsverhältnis in dem Projekt

5. Kreuzen Sie an, welche Daten in aller Regel nicht urheberrechtlich geschützt sind.

   ☐ einfache Computerprogramme

   ☐ Messdaten aus den Naturwissenschaften

   ☐ Interviewdaten

Musterlösung: Urheberrecht

Timo schreibt im Zuge seiner Masterarbeit Software, die er nach der Fertigstellung veröffentlichen möchte. Ihm fällt auf, dass nicht nur die Softwarelizenzierung eine wichtige Rolle spielt. Er fragt sich unter anderem, wem die Daten aus dem Open-Source-Projekt gehören.

1. Timo schreibt seine Masterarbeit an der Technischen Universität Dresden und engagiert sich ehrenamtlich bei einem Open-Source-Projekt. Kreuzen Sie an, bei wem das Urheberrecht von Timos Code liegt.

   ☒ Timo

   ☐ Technsiche Universität Dresden

   ☐ Projektleiter:in des Open-Source-Projekts

2. Erläutern Sie, ob es möglich ist das Urheberrecht zu übertragen bzw. abzutreten.

   Es ist in Deutschland nicht möglich, das Urheberrecht abzutreten. Es kann maximal ein Nutzungsrecht übertragen werden.

3. Kreuzen Sie an, in welchem Fall das Nutzungsrecht bei der Universität liegen würde.

   ☐ Nie

   ☒ Wenn Timo es ausdrücklich überträgt

   ☒ Bei einem Arbeitsverhältnis mit der Universität

4. Kreuzen Sie an, in welchem Fall das Nutzungsrecht bei der:dem Projektleiter:in des Open-Source-Projekts liegen würde.

   ☐ Nie

   ☒ Wenn Timo es ausdrücklich überträgt

   ☒ Bei einem Arbeitsverhältnis in dem Projekt

5. Kreuzen Sie an, welche Daten in aller Regel nicht urheberrechtlich geschützt sind (Ausnahmen sind möglich).

   ☒ einfache Computerprogramme

   ☒ Messdaten aus den Naturwissenschaften

   ☐ Interviewdaten

5.14 Langzeitarchivierung

Das Thema Langzeitarchivierung wurde in Kapitel 4.15 vorgestellt und ist ein essenzieller Bestandteil des FDM. Um ein Verständnis über Langzeitarchivierung zu entwickeln, ist es wichtig, einige Begriffe voneinander abzugrenzen. Auch die Wahl von nachhaltigen Formaten ist ein wichtiger Bestandteil. In Tabelle 5.18 wird entsprechendes Arbeitsmaterial aufgelistet.

Tabelle 5.18: Lehrmaterial Langzeitarchivierung

Thema	**Seite**
Arbeitsblatt: Langzeitarchivierung	250
Musterlösung: Langzeitarchivierung	251
Checkliste: Wahl von Langzeitarchiven	253

Arbeitsblatt: Langzeitarchivierung

Carla möchte die geführten Interviews nach dem Abschluss ihrer Bachelorarbeit langzeitarchivieren. Es handelt sich um ein internationales Projekt. Sie muss ihre Daten nun ggf. für die Langzeitarchivierung aufbereiten und geeignete Repositorien auswählen.

1. Die Tonaufnahmen der Interviews liegen derzeit in einem .mp4-Format vor. Welche Maßnahmen müssen für eine Langzeitarchivierung vorgenommen werden? Begründen Sie Ihre Aussage.

   - ☐ Es müssen keine Maßnahmen ergriffen werden.
   - ☐ Das Format sollte beibehalten, aber die Dateigröße komprimiert werden.
   - ☐ Das Format sollte in .mp3 konvertiert werden.

   .....

2. Nennen Sie Dateiformate, die für die Langzeitarchivierung geeignet sind.

   .....

3. Was ist der Unterschied zwischen einem Repositorium und einem Archiv?

   .....

4. Was ist der Unterschied zwischen Back-up und Archivierung?

   .....

5. Welchen der folgenden Aussagen würden Sie für die Suche nach einem geeigneten Archiv zustimmen?

   - ☐ Das Archiv sollte eine konkrete Zugehörigkeit zu einer Fachdisziplin aufweisen.
   - ☐ Das Archiv sollte durch ein Gütesiegel ausgezeichnet sein.
   - ☐ Die Langlebigkeit des Archivs ist wichtiger als der Bekanntheitsgrad des Anbieters.
   - ☐ Es sollte eine möglichst lange Speicherdauer garantiert werden.

Musterlösung: Langzeitarchivierung

Carla möchte die geführten Interviews nach dem Abschluss ihrer Bachelorarbeit langzeitarchivieren. Es handelt sich um ein internationales Projekt. Sie muss ihre Daten nun ggf. für die Langzeitarchivierung aufbereiten und geeignete Repositorien auswählen.

1. Die Tonaufnahmen der Interviews liegen derzeit in einem .mp4-Format vor. Welche Maßnahmen müssen für eine Langzeitarchivierung vorgenommen werden? Begründen Sie Ihre Aussage.

   ☒ Es müssen keine Maßnahmen ergriffen werden.

   ☐ Das Format sollte beibehalten, aber die Dateigröße komprimiert werden.

   ☒ Das Format sollte in .mp3 konvertiert werden.

   Da es sich bei .mp4 um ein Video-Container-Format handelt, ist es empfehlenswert, das Format in ein reines Audioformat zu konvertieren, welches von dem gewählten Archiv als empfehlenswert gelistet wird (z. B. .mp3). Grundsätzlich wäre jedoch auch die Archivierung in dem Originalformat (.mp4) möglich.

2. Nennen Sie Dateiformate, die für die Langzeitarchivierung geeignet sind.

   Zu den Dateiformaten, die für eine Langzeitarchivierung geeignet sind, zählen je nach Dateityp: .txt, .odf, .tiff, .png, .csv, .xml, aber auch .docx oder .xlsx und pdf/a.

3. Was ist der Unterschied zwischen einem Repositorium und einem Archiv?

   Sowohl Repositorien als auch Langzeitarchive sind dafür gedacht, Daten zu halten. Der Unterschied besteht jedoch in der primären Zielsetzung der Infrastruktur: Repositorien sollen die Daten für die Öffentlichkeit verfügbar machen (mit möglichen Zugangsbeschränkungen), während die Langzeitarchive die Daten für unbestimmte (bzw. eine bestimmte, lange) Zeit aufbewahren und lesbar halten sollen.

4. Was ist der Unterschied zwischen Back-up und Archivierung?

   Ein Back-up ist eine regelmäßige (automatische) Sicherung aller aktuellen Daten und aller Versionen, um den Datenverlust vorzubeugen. Im Gegensatz dazu werden bei der Langzeitarchivierung nur endgültige Daten und Versionen (bzw. Meilensteinversionen) langfristig archiviert.

5. Welchen der folgenden Aussagen würden Sie für die Suche nach einem geeigneten Archiv zustimmen?

   ☐ Das Archiv sollte eine konkrete Zugehörigkeit zu einer Fachdisziplin aufweisen.

   ☒ Das Archiv sollte durch ein Gütesiegel ausgezeichnet sein.

☒ Die Langlebigkeit des Archivs ist wichtiger als der Bekanntheitsgrad des Anbieters.

☒ Es sollte eine möglichst lange Speicherdauer garantiert werden.

Checkliste: Wahl von Langzeitarchiven

Leitfragen: Was ist bei der Wahl eines Langzeitarchivs zu beachten?

- ☐ Wie lange sollen die Daten aufbewahrt werden?

  .....

- ☐ Wie viel Speicherplatz benötige ich?

  .....

- ☐ Welche Datenformate habe ich? Müssen sie in nachhaltige Formate umgewandelt werden?

  .....

- ☐ Wer benötigt Zugang?

  .....

- ☐ Wo werden die Daten und deren Dokumentation nach Projektende aufbewahrt?

  .....

- ☐ Hat der Dienstleister eine Strategie zur Datenkonvertierung und Migration?

  .....

- ☐ Wird die Integrität der Daten regelmäßig überprüft?

  .....

- ☐ Ist das Langzeitarchiv vertrauenswürdig? Besitzt es ein Siegel?

  .....

- ☐ Wie langlebig ist der Dienstleister?

  .....

- ☐ Wie häufig wird ein Backup gemacht und wo wird dieser gespeichert?

  .....

Quelle:
CESSDA Training Working Group. CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC, 2017-2018, https://www.cessda.eu/DMGuide. Das Werk ist lizensiert unter der Creative Commons Attribution-ShareAlike 4.0 International Lizenz.

gefördert vom: Bundesministerium für Bildung und Forschung

5.15 Nachnutzung

Die Nachnutzung ist ein weiterer wichtiger Bestandteil von FDM und wird in Kapitel 4.16 beschrieben. Insbesondere durch das Ziel, Forschungsdaten zu publizieren, soll erreicht werden, dass Forschungsdaten nachgenutzt werden können. Dabei muss jedoch beachtet werden, welche Daten wie nachgenutzt werden dürfen. In Tabelle 5.19 wird das folgende Lehrmaterial zum Thema Nachnutzung zusammengefasst.

Tabelle 5.19: Lehrmaterial Nachnutzung

Thema	**Seite**
Arbeitsblatt: Nachnutzung	255
Musterlösung: Nachnutzung	257

Arbeitsblatt: Nachnutzung

Alex hat einige Algorithmen für die Analyse von Twitterdaten entwickelt. Nun möchte Alex diese auf einem bestehenden Datensatz testen und bestimmen, inwieweit sie auch für die Analyse weiterer Nachrichtendienste verwendet werden können. Dafür möchte Alex bestehende Datensätze nutzen.

1. Kreuzen Sie an, welche CC-Lizenz Alex vergeben kann, wenn Alex zwei Datensätze nachnutzt, die jeweils unter den folgenden Lizenzen stehen[1]:

   (a) CC BY und CC BY-SA?

   - ☐ CC BY
   - ☐ CC BY-SA
   - ☐ andere, welche? .....
   - ☐ unzulässig

   (b) CC BY-SA und CC BY-NC?

   - ☐ CC BY-SA
   - ☐ CC BY-NC
   - ☐ CC BY-NC-SA
   - ☐ andere, welche? .....
   - ☐ unzulässig

   (c) CC BY und CC BY-ND?

   - ☐ CC BY
   - ☐ CC BY-ND
   - ☐ andere, welche? .....
   - ☐ unzulässig

[1] Quelle der Aufgabe: Biernacka, K., Buchholz, P., Danker, S. A., Dolzycka, D., Engelhardt, C., Helbig, K., Jacob, J., Neumann, J., Odebrecht, C., Petersen, B., Slowig, B., Trautwein-Bruns, U., Wiljes, C. & Wuttke, U. (2021). *Train-the-Trainer Konzept zum Thema Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.5773203

2. Kreuzen Sie an, welche der genannten Softwarelizenzen miteinander kompatibel sind.

	GPLv2	LGPLv3	Apache 2.0	BSD 2/3 clause	MIT
GPLv2	☐	☐	☐	☐	☐
LGPLv3	☐	☐	☐	☐	☐
Apache 2.0	☐	☐	☐	☐	☐
BSD 2/3 clause	☐	☐	☐	☐	☐
MIT	☐	☐	☐	☐	☐

3. Gegeben sind die folgenden Daten: https://doi.org/10.17632/z9zw7nt5h2.1.

   (a) Beurteilen Sie, ob Alex diese Daten nachnutzen darf.

   .....

   (b) Beurteilen Sie, ob Alex die Daten mit den eigenen verbinden darf.

   .....

   (c) Zitieren Sie den Datensatz.

   .....

Musterlösung: Nachnutzung

Alex hat einige Algorithmen für die Analyse von Twitterdaten entwickelt. Nun möchte Alex diese auf einem bestehenden Datensatz testen und bestimmen, inwieweit sie auch für die Analyse weiterer Nachrichtendienste verwendet werden können. Dafür möchte Alex bestehende Datensätze nutzen.

1. Kreuzen Sie an, welche CC-Lizenz Alex vergeben kann, wenn Alex zwei Datensätze nachnutzt, die jeweils unter den folgenden Lizenzen stehen[1]:

   (a) CC BY und CC BY-SA?

   - ☐ CC BY
   - ☒ CC BY-SA
   - ☐ andere, welche? .....
   - ☐ unzulässig

   (b) CC BY-SA und CC BY-NC?

   - ☐ CC BY-SA
   - ☐ CC BY-NC
   - ☐ CC BY-NC-SA
   - ☐ andere, welche? .....
   - ☒ unzulässig

   (c) CC BY und CC BY-ND?

   - ☐ CC BY
   - ☐ CC BY-ND
   - ☐ andere, welche? .....
   - ☒ unzulässig

[1] Quelle der Aufgabe: Biernacka, K., Buchholz, P., Danker, S. A., Dolzycka, D., Engelhardt, C., Helbig, K., Jacob, J., Neumann, J., Odebrecht, C., Petersen, B., Slowig, B., Trautwein-Bruns, U., Wiljes, C. & Wuttke, U. (2021). *Train-the-Trainer Konzept zum Thema Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.5773203

2. Kreuzen Sie an, welche der genannten Softwarelizenzen miteinander kompatibel sind.

	GPLv2	LGPLv3	Apache 2.0	BSD 2/3 clause	MIT
GPLv2	✓	-	-	✓	✓
LGPLv3	-	✓	✓	✓	✓
Apache 2.0	-	✓	✓	✓	✓
BSD 2/3 clause	✓	✓	✓	✓	✓
MIT	✓	✓	✓	✓	✓

3. Gegeben sind die folgenden Daten: https://doi.org/10.17632/z9zw7nt5h2.1.

   (a) Beurteilen Sie, ob Alex diese Daten nachnutzen darf.
   Ja, dank der CC BY 4.0 Lizenz.

   (b) Beurteilen Sie, ob Alex die Daten mit den eigenen verbinden darf.
   Ja, die CC BY 4.0 Lizenz erlaubt die Veränderungen und somit auch das Verbinden von Daten.

   (c) Zitieren Sie den Datensatz.
   Hussein, S. (2021). Twitter Sentiments Dataset. Mendeley Data. https://doi.org/10.17632/z9zw7nt5h2.1

5.16 Weitere rechtliche Aspekte

In Kapitel 4.17 werden weitere rechtliche Aspekte beleuchtet, die im Rahmen des FDM wichtig sind. Insbesondere das Patentrecht ist hier für das Fachgebiet Informatik zu nennen. Um grundlegende Regelungen des Patentrechts zu lehren, werden im Folgenden ein Arbeitsblatt mit Musterlösung vorgestellt (siehe Tabelle 5.20). Für eine Vertiefung der Thematik sollte sich zusätzlich juristischer Rat eingeholt werden.

Tabelle 5.20: Lehrmaterial Weitere rechtliche Aspekte

Thema	**Seite**
Arbeitsblatt: Patentrecht	260
Musterlösung: Patentrecht	261

Arbeitsblatt: Patentrecht

Alex hat einen Algorithmus entwickelt, um Twitterdatenströme analysieren zu können. Das Projekt ist vom BMBF gefördert und der Algorithmus soll flächendeckend in der Lehre eingesetzt werden, um die Stimmung von Studierenden zu erheben.

1. Begründen Sie, ob sich Alex den Algorithmus patentieren lassen kann.

   .....

   .....

   .....

   .....

2. Begründen Sie, ob der Algorithmus von Alex als Diensterfindung erklärt werden kann.

   .....

   .....

   .....

   .....

Musterlösung: Patentrecht

Alex hat einen Algorithmus entwickelt, um Twitterdatenströme analysieren zu können. Das Projekt ist vom BMBF gefördert und der Algorithmus soll flächendeckend in der Lehre eingesetzt werden, um die Stimmung von Studierenden zu erheben.

1. Begründen Sie, ob sich Alex den Algorithmus patentieren lassen kann.
   Alex kann sich den Algorithmus im Rahmen eines Softwarepatents patentieren lassen, da es sich um eine Erfindung handelt, die nicht zum derzeitigen Stand der Technik gehört. Da ein kommerzielles Interesse der Softwarenutzung bei Dritten bestehen kann, ist die Patentierung ein Weg, die gewerbliche Nutzung auszuschließen.

2. Begründen Sie, ob der Algorithmus von Alex als Diensterfindung erklärt werden kann.
   Da Alex im Rahmen des Projektes einen Arbeitsvertrag mit der Humboldt-Universität zu Berlin geschlossen hat, kann die Software als Diensterfindung deklariert werden. Aufgrund der Wissenschaftsfreiheit darf Alex die Erfindung jedoch für eigene Lehr- und Forschungszwecke nutzen.

Literatur

Adam, B. & Lindstädt, B. (2019). *ELN-Wegweiser. Elektronische Laborbücher im Kontext von Forschungsdatenmanagement und guter wissenschaftlicher Praxis – ein Wegweiser für die Lebenswissenschaften.* ZB MED – Informationszentrum Lebenswissenschaften. https://doi.org/10.4126/FRL01-006415715

Amt der Europäischen Union für geistiges Eigentum. (2021). Häufig gestellte Fragen (FAQ) zum Urheberrecht. [Zuletzt geprüft 2022-06-13]. https://euipo.europa.eu/ohimportal/de/web/observatory/faqs-on-copyright-de

Baumann, P., Krahn, P. & Lauber-Rönsberg, A. (2021). *Forschungsdatenmanagement und Recht. Datenschutz-, Urheber- und Vertragsrecht.* W. Neugebauer.

Beurskens, M. (2013). Legal Questions of Twitter Research. In K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (Hrsg.), *Twitter and Society* (S. 123–133). Peter Lang.

Biernacka, K., Buchholz, P., Danker, S. A., Dolzycka, D., Engelhardt, C., Helbig, K., Jacob, J., Neumann, J., Odebrecht, C., Petersen, B., Slowig, B., Trautwein-Bruns, U., Wiljes, C. & Wuttke, U. (2021a). *Train-the-Trainer Konzept zum Thema Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.5773203

Biernacka, K., Halbherr, V., Lange, M., Martin, L., Mieck, C. & Reimer, N. (2022). *Open Access und wissenschaftliches Publizieren: Train-the-Trainer-Konzept.* https://doi.org/10.5281/zenodo.6034407

Biernacka, K., Helbig, K. & Buchholz, P. (2021b). Adaptable Methods for Training in Research Data Management. *Data Science Journal.* https://doi.org/10.5334/dsj-2021-014

Biernacka, K., Helbig, K., Senft, M. & Trautwein-Bruns, U. (2020). Datendokumentation leicht gemacht! Ein interaktiver Online-Workshop. https://doi.org/10.5281/zenodo.4037151

Brettschneider, P. (2021a). Checkliste - Ist Text- und Data-Mining zulässig? [Archivierte Version: https://perma.cc/JPZ5-3ETR]. https://www.forschungsdaten.info/typo3temp/secure_downloads/108220/0/d2e38ad44680e4d9c72bfb7e67de72ffac428115/csm_Checkliste_ca2f1bbe0f.png

Brettschneider, P. (2021b). Text- und Data-Mining nach dem Verständnis des Gesetzgebers. [Archivierte Version: https://perma.cc/578J-PFJN]. https://www.forschungsdaten.info/typo3temp/secure_downloads/108220/0/72ff76ad09936db0310a9275bb17472b95355f45/csm_2020.1.2_Text_Data-Mining_v6_3a4ca4f5c3.png

Brettschneider, P., Biernacka, K., Böker, E., Danker, S. A., Jacob, J., Perry, A., Wiljes, C. & Wuttke, U. (2021). Urheberrecht und Lizenzierung bei Forschungsdaten. https://doi.org/10.5281/zenodo.5243232

Bundesamt für Sicherheit in der Informationstechnik. (2021). Sichere Passwörter erstellen. [Archivierte Version: https://perma.cc/P2LK-W6DE]. https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Informationen-und-Empfehlungen/Cyber-Sicherheitsempfehlungen/Accountschutz/Sichere-Passwoerter-erstellen/sichere-passwoerter-erstellen_node.html

Bundesamt für Sicherheit in der Informationstechnik. (2022). Glossar. [Archivierte Version: https://perma.cc/DQ2G-JG5Y]. https://www.bsi.bund.de/DE/Service-Navi/Cyber-Glossar/Functions/glossar.html?nn=520190&cms_lv2=132772

Carroll, S. R., Hudson, M., Chapman, J., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R. & Walker, J. (2019). Die CARE-Prinzipien für indigene Data Governance. https://doi.org/10.5281/zenodo.5995059

CASRAI. (2021). CRediT – Contributor Roles Taxonomy. [Archivierte Version: https://perma.cc/L4VR-73FP]. https://casrai.org/credit/

CESSDA ERIC. (2020). Data Management Expert Guide. Documentation and metadata. [Archivierte Version: https://perma.cc/E83H-WJ7N]. https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

Chue Hong, N. (2021). https://software.ac.uk/choosing-repository-your-software-project

Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., Honeyman, T., Struck, A., Lee, A., Loewe, A., van Werkhoven, B., Jones, C., Garijo, D., Plomp, E., Genova, F., ... WG, R. F. (2022). FAIR Principles for Research Software (FAIR4RS Principles). https://doi.org/10.15497/RDA00068

Clément-Fontaine, M., Di Cosmo, R., Guerry, B., Moreau, P. & Pellegrini, F. (2019). Encouraging a wider usage of software derived from research: Opportunity Note. https://doi.org/10.52949/4

CoreTrustSeal Standards and Certification Board. (2019). CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020–2022. https://doi.org/10.5281/zenodo.3632533

Crawford, K. (2017). The Trouble with Bias. [Zuletzt geprüft 2022-02-11]. https://www.youtube.com/watch?v=fMym_BKWQzk

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles*. https://doi.org/10.25490/A97F-EGYK

Deppe, A. (2020). FAIR, CARE und mehr. Prinzipien für einen verantwortungsvollen Umgang mit Forschungsdaten. In M. Schulze (Hrsg.), *Historisches Erbe und zeitgemäße Informationsinfrastrukturen: Bibliotheken am Anfang des 21. Jahrhunderts* (S. 299–312). kassel university press.

Deutsche Forschungsgemeinschaft. (2019). Leitlinien zur Sicherung guter wissenschaftlicher Praxis, 29. https://doi.org/10.5281/zenodo.3923602

Döring, K. W. (2008). *Handbuch Lehren und Trainieren in der Weiterbildung*. Beltz.

Engelhardt, C., Biernacka, K., Coffey, A., Cornet, R., Danciu, A., Demchenko, Y., Downes, S., Erdmann, C., Garbuglia, F., Germer, K., Helbig, K., Hellström, M., Hettne, K., Hibbert, D., Jetten, M., Karimova, Y., Kryger Hansen, K., Kuusniemi, M. E., Letizia, V., … Zhou, B. (2022). *How to be FAIR with your data. A teaching and training handbook for higher education institutions*. Zenodo. https://doi.org/10.5281/zenodo.5837500

forschungsdaten.info. (2021). Text- und Data-Mining. Automatisierte Auswertung von Forschungsdaten. [Archivierte Version: https://perma.cc/7BFR-WES3]. https://www.forschungsdaten.info/themen/rechte-und-pflichten/text-und-data-mining/

forschungsdaten.org. (2017). Persistent Identifier. [Archivierte Version: https://perma.cc/8D2F-BYKX]. https://www.forschungsdaten.org/index.php/Persistent_Identifier

forschungsdaten.org. (2019). Langzeitarchivierung. [Archivierte Version: https://perma.cc/7P84-2DYB]. https://www.forschungsdaten.org/index.php/Langzeitarchivierung

forschungsdaten.org. (2021). Data Journals. [Archivierte Version: https://perma.cc/YGC7-2PE2]. https://www.forschungsdaten.org/index.php/Data_Journals

Funk, S. E. (2010a). 8.3 Emulation. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann & K. Huth (Hrsg.), *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. http://d-nb.info/117802752X/34

Funk, S. E. (2010b). 8.4 Migration. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann & K. Huth (Hrsg.), *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. http://d-nb.info/117802752X/34

Galetzka, C., Jun, C.-j. & Roßmann, Y. (2021). *Praxishandbuch Open Source. Technische und rechtliche Rahmenbedingungen für einen lizenzkonformen Einsatz von FOSS im Unternehmen*.

Gerlach, R., Rex, J., Lang, K., Neute, N. & Schwartze, V. (2020). Fact Sheet: Research Data Repositories. https://doi.org/10.5281/zenodo.3900922

Gesellschaft für Informatik e. V. (2004). Unsere Ethischen Leitlinien. [Archivierte Version: https://perma.cc/8JUU-BKJ7]. https://gi.de/fileadmin/GI/Allgemein/PDF/GI_Ethische_Leitlinien_2004.pdf

Gesellschaft für Informatik e. V. (2016). Empfehlungen für Bachelor- und Masterprogramme im Studienfach Informatik an Hochschulen. [Archivierte Version: https://perma.cc/SR3N-Q2TR]. https://gi.de/fileadmin/GI/Hauptseite/Aktuelles/Meldungen/2016/GI-Empfehlungen_Bachelor-Master-Informatik2016.pdf

Götting, H.-P., Hetmank, S. & Schwipps, K. (2014). *Patentrecht*. Verlag C.H. Beck.

Gruber, T. (2016). Ontology. In L. Liu & Ö. M. (Hrsg.), *Encyclopedia of Database Systems*. Springer. https://doi.org/10.1007/978-1-4899-7993-3_1318-2

Gruninger, M. & Lee, J. (2002). Ontology applications and design. *Communications of the ACM*, *45*(2), 39.

Hartmann, T. (2018). "Terra incognita – digitale Forschungsdaten auf der Suche nach einer rechtlichen Heimat. [Archivierte Version: https://perma.cc/8GJK-249U]. https://www.forschungsdaten.org/index.php/Datei:Hartmann_TerraIncognita-Forschungsdaten-RechtlicheHeimat.pdf

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V. & Zwitter, A. (2017). Digitale Demokratie statt Datendiktatur. In C. Könneker (Hrsg.), *Unsere digitale Zukunft. In welcher Welt wollen wir leben?* (S. 3–21). Springer. https://doi.org/10.1007/978-3-662-53836-4_1

Hiemenz, B. M. & Kuberek, M. (2018). Leitlinie? Grundsätze? Policy? Richtlinie? – Forschungsdaten-Policies an deutschen Universitäten. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, *5*(2), 1–13. https://doi.org/10.5282/o-bib/2018H2S1-13

Humboldt-Universität zu Berlin. (2015). Fachspezifische Studien- und Prüfungsordnung für das Bachelorstudium Informatik. [Archivierte Version: https://perma.cc/888T-P76R]. https://gremien.hu-berlin.de/de/amb/2015/13/13_2015_AMB_Monobachelor_Informatik_DRUCK.pdf

IANUS. (2016). IT-Empfehlungen Für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften. Dateiformate. [Archivierte Version: https://perma.cc/M6BH-RTCY]. https://ianus-fdz.de/it-empfehlungen/dateiformate

IEEE. (2020). IEEE Standard for Learning Object Metadata. *IEEE Std 1484.12.1-2020*, 1–50. https://doi.org/10.1109/IEEESTD.2020.9262118

Jones, M. B., Boettiger, C., Mayes, A. C., Smith, A., Slaughter, P., Niemeyer, K., Gil, Y., Fenner, M., Nowak, K., Hahnel, M., Coy, L., Allen, A., Mercè Crosas, A. S., Hong, N. C., Cruse, P., Katz, D. S. & Goble, C. (2017). CodeMeta: an exchange schema for software metadata. https://doi.org/10.5063/schema/codemeta-2.0

Kindling, M. & Schirmbacher, P. (2013). Die digitale Forschungswelt als Gegenstand der Forschung / Research on Digital Research / Recherche dans la domaine de la recherche numerique. *Information - Wissenschaft & Praxis*, *64*(2-3), 127–136. https://doi.org/10.1515/iwp-2013-0017

Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R. R. & Asmi, A. (2021). Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. *Data Science Journal*, *20*(1), 12. https://doi.org/10.5334/dsj-2021-012

Koordinatiosstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST). (2021). Dateiformate (KaD). Empfehlung. [Archivierte Version: https://perma.cc/AQP7-F2LD]. https://kost-ceco.ch/cms/empfehlung.html

Kranz, G. (2021). Metadaten. [Archivierte Version: https://perma.cc/ZE52-G2C4]. https://whatis.techtarget.com/de/definition/Metadaten

Kreutzer, T. & Lahmann, H. (2021). *Rechtsfragen bei Open Science. Ein Leitfaden.* Hamburg University Press. https://doi.org/10.15460/HUP.211

Lambe, P. (2006). Defining "Taxonomy". [Archivierte Version: https://perma.cc/KFP8-3KRX]. http://www.greenchameleon.com/gc/blog_detail/defining_taxonomy/

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., Van De Sandt, S., Ison, J., Martinez, P. A. et al. (2020).

Towards FAIR principles for research software. *Data Science*, *3*(1), 37–59. https://doi.org/10.3233/DS-190026

Lauber-Rönsberg, A. (2021). Rechtliche Aspekte des Forschungsdatenmanagements. In H. N. Markus Putnings & J. Neumann (Hrsg.), *Praxishandbuch Forschungsdatenmanagement* (S. 89–114). De Gruyter Saur. https://doi.org/10.1515/9783110657807-005

Lucke, U. (2021). Forschungsdatenmanagement aus der Sicht der Informatik. [Zuletzt geprüft 2022-06-13]. https://www.youtube.com/watch?v=2DoBQlH3wR4

Ludwig, J. & Enke, H. (2013). *Leitfaden zum Forschungsdaten-Management*. Verlag Werner Hülsbusch.

Open Knowledge Foundation. (2021). *Open Definition. Defining Open in Open Data, Open Content and Open Knowledge.* [Archivierte Version: https://perma.cc/EE4Z-88SZ] (2.1). http://opendefinition.org/od/2.1/de/

ORCID. (2022). Open Researcher and Contributor ID. [Archivierte Version: https://perma.cc/5KCH-WRP8]. https://orcid.org/

OSF Support. (2022). How to Make a Data Dictionary. [Archivierte Version: https://perma.cc/7F28-GULL]. https://help.osf.io/hc/en-us/articles/360019739054-How-to-Make-a-Data-Dictionary

Piwowar, H. A. & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, (1:e175), 25. https://doi.org/10.7717/peerj.175

Preston-Werner, T. (2013). Semantic Versioning 2.0.0. [Archivierte Version: https://perma.cc/H25D-XK7P]. https://semver.org/lang/de/

Research Data Alliance. (2021). RDA Metadata Standards Directory. [Archivierte Version: https://perma.cc/FW2N-GFZU]. http://rd-alliance.github.io/metadata-directory/subjects/

Research Data Alliance International Indigenous Data Sovereignty Interest Group. (2019). CARE Principles for Indigenous Data Governance. [Archivierte Version: https://perma.cc/Z429-XMUK]. http://gida-global.org/

Rösch, H. (2021). Forschungsethik und Forschungsdaten. In M. Putnings, H. Neuroth & J. Neumann (Hrsg.), *Praxishandbuch Forschungsdatenmanagement* (S. 115–140). De Gruyter Saur. https://doi.org/10.1515/9783110657807-006

Schasche, S. (2018). Haltbarkeit von Speichermedien. [Archivierte Version: https://perma.cc/95UR-8TM8]. https://www.pc-magazin.de/ratgeber/speichermedien-lebensdauer-dvd-festplatte-usb-stick-floppy-disk-1485976.html

Schulz, S. (2019). *Physical Computing als Mittel der wissenschaftlichen Erkenntnisgewinnung in der Informatik und als fächerverbindende MINT-Arbeitsweise*. Logos Verlag Berlin GmbH.

Smith, A. M., Katz, D. S. & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, *2*. https://doi.org/10.7717/peerj-cs.86

Summann, F. (2015). ZDS - Kriterienbereich Normdaten. Kriterienbereich Normdaten / kontorllierte Vokabulare. [Archivierte Version: https://perma.cc/57DC-EUGR]. https://wiki.dnb.de/display/DINIAGKIM/ZDS+-+Kriterienbereich+Normdaten

Technische Universität Dresden. (2016). Studienordnung für den Bachelor-Studiengang Informatik. [Archivierte Version: https://perma.cc/RB2V-PM2L]. https://www.verw.tu-dresden.de/AmtBek/PDF-Dateien/2016-06/11soBA24.04.2016.pdf

The Software Sustainability Institute. (2018). Checklist for a Software Management Plan. https://doi.org/10.5281/zenodo.2159713

The Software Sustainability Institute. (2021). Software Management Plans. [Archivierte Version: https://perma.cc/J5Z5-MSGC]. https://www.software.ac.uk/software-management-plans

TU9-FDM. (2019). Software als Forschungsdaten. https://doi.org/10.5281/zenodo.2611303

UK Data Service. (2020). *2021*(2021-08-04). https://ukdataservice.ac.uk/learning-hub/research-data-management/

Universität Hamburg. (2019). Fachspezifische Bestimmungen für den Studiengang Informatik (B.Sc.). [Archivierte Version: https://perma.cc/84BN-R6YY]. https://www.uni-hamburg.de/campuscenter/studienorganisation/ordnungen-satzungen/pruefungs-studienordnungen/mathematik-informatik-und-naturwissenschaften/20190403-fsb-min-bsc-informatik-38.pdf

VÖRBY. (2012). Beispiel für Festlegungen, die Zugriffsrechte enthalten können. [Archivierte Version: https://perma.cc/BZ3R-HR9G]. https://de.wikipedia.org/wiki/Datei:Zugriffsrecht.png

Weber-Wulff, D., Class, C., Coy, W., Kurz, C. & Zehllhöfer, D. (2009). *Gewissensbisse. Ethische Probleme in der Informatik. Biometrie - Datenschutz - geistiges Eigentum.* transcript.

Wheeler, D. A. (2012). Open Source Software (OSS or FLOSS) and the U.S. Department of Defense (DoD). [Archivierte Version: https://perma.cc/J8MP-EWW3]. https://dwheeler.com/essays/oss-dod-overview-2012-08-15.ppt

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Yang, N. & Li, T. (2020). How stakeholders' data literacy contributes to student success in higher education: a goal-oriented analysis. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–18. https://doi.org/10.1186/s41239-020-00220-3

Dieses Buch bietet einen Überblick über Forschungsdatenmanagement und dessen konkrete Umsetzung in der Informatik. Anhand von Personas und Szenarien wird ein Großteil von informatischen Anwendungsfällen und Fragen abgedeckt, um Lehrende sowie Studierende der Informatik dabei zu unterstützen, Forschungsdatenmanagement adäquat zu realisieren.

Im ersten Teil des Buchs wird Lehrenden der Informatik anhand von Modulen konkret aufgezeigt, welche Themen der Informatik besonders geeignet sind, um Forschungsdatenmanagement in das Informatikstudium zu integrieren. Der zweite Teil des Buchs erläutert Bestandteile des Forschungsdatenmanagements und veranschaulicht deren Anwendung auf ausgewählte Szenarien. Abschließend werden im dritten Teil Lehrmaterialien (Arbeitsblätter, Musterlösungen, Checklisten und weitere) zur Verfügung gestellt, um den direkten und fachgerechten Einsatz in der universitären Lehre zu stärken.

Katarzyna Biernacka ist wissenschaftliche Mitarbeiterin am Lehrstuhl Didaktik der Informatik / Informatik und Gesellschaft an der Humboldt-Universität zu Berlin.

Sandra Schulz ist Juniorprofessorin für Didaktik der Informatik / Computer Science Education an der Universität Hamburg.

Logos Verlag Berlin
ISBN 978-3-8325-5490-3