Technology
Arts Sciences
TH Köln

Implementierung einer Machine Learning-Kollektion als Open Educational Resources

Masterarbeit

im Studiengang

Bibliotheks- und Informationswissenschaft (MALIS)
Fakultät für Informations- und Kommunikationswissenschaften
Technische Hochschule Köln

vorgelegt von
Ines Schmahl

Köln, 02.05.2022

Erstgutachter: Prof. Dr. Konrad Förstner
Zweitgutachter: Prof. Dr. Philipp Schaer

Kurzfassung

Die Entwicklungen im Bereich der Informations- und Kommunikationstechnologie haben völlig neue Möglichkeiten des Datenaustausches und der Zusammenarbeit geschaffen. Das zeigt sich auch in der Lehre. Hier hat sich der Begriff der Open Educational Resources (OER) entwickelt, womit frei zugängliche Bildungsmaterialien bezeichnet werden. Um dieses Potential ausschöpfen zu können, braucht es innovative Herangehensweisen.

In dieser Arbeit wird ein neuer Ansatz vorgestellt. Anstatt Lehrmaterialien als Teil eines didaktischen Konzeptes zu betrachten, werden sie als Objekte einer digitalen Sammlung verstanden. Das erleichtert die Adaptierung der Materialien an die spezifischen Anforderungen von Lehrveranstaltungen. Konkret wird der Ansatz auf den Aufbau einer OER-Kollektion für Machine Learning angewendet. Denn gerade in diesem Bereich zeichnet sich ein hoher Bedarf an Kompetenzvermittlung ab, um Forschende auf die Anforderungen einer immer datenintensiveren Wissenschaft vorzubereiten.

Die gewonnenen Erfahrungen werden im Fazit als Lessons learned zusammengefasst, um damit andere bei der Umsetzung ähnlicher Vorhaben zu unterstützen.

Die OER-Kollektion ist zugänglich über die Online-Plattform GitHub unter dem Link:

`https://github.com/Machine-Learning-OER-Collection/Machine-Learning-OER-Basics`

Schlagwörter:
Open Educational Resources, OER, Machine Learning, Nachnutzbarkeit, Sammlung

Abkürzungsverzeichnis

API	Application Programming Interface / Programmierschnittstelle
DFG	Deutsche Forschungsgemeinschaft
DMP	Datenmanagementplan
DOI	Digital Object Identifier
DS	Data Science
IPYNB	IPython Notebook Format
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MOOCs	Massive Open Online Courses
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OCW	OpenCourseWare
OER	Open Educational Resources
OERR	Open Educational Resources Repositorium
REST	Representational State Transfer
SVG	Scalable vector graphics
XML	Extensible Markup Language

Inhaltsverzeichnis

1 Einleitung

Open Educational Resources (OER) sind ein Teilbereich von Open Science, welches zum Ziel hat, die während des wissenschaftlichen Forschungsprozesses entstehenden Informationen zugänglich zu machen [1]. Initiiert wurde dies durch die Digitalisierung, die den Austausch von Daten enorm vereinfacht hat. Für die Lehre als Bestandteil von Wissenschaft bieten sich dadurch neue Möglichkeiten. Das Bereitstellen von Bildungsmaterialien über das Internet ist ein Zugewinn für Lehrende, die sie zur Erstellung eigener Kurse verwenden können.

In der Praxis stehen diesem Workflow allerdings noch einige Hürden entgegen. Didaktische Konzepte sind auf eine jeweilige Zielgruppe ausgerichtet. Auch die Lernziele können im Einzelfall sehr unterschiedlich sein. Dem stehen viele OER-Materialien gegenüber, bei denen es sich um abgeschlossene Lehreinheiten handelt, wie im Kapitel 3.2 gezeigt werden wird.

Um die Nachnutzbarkeit zu verbessern, wird ein neuer Lösungsansatz eingeführt. Anstatt einen Kurs zu entwickeln, werden die Materialien in Form einer digitalen Sammlung präsentiert. Sie erlaubt das einfache Herausnehmen einzelner Objekte, die ohne großen Aufwand in neue Kurse eingebettet werden können.

Da insbesondere im Bereich Machine Learning (ML) aufgrund des Wandels hin zu einer immer datenintensiveren Forschung ein hoher Bedarf an Kompetenzvermittlung besteht, soll im Rahmen der Masterarbeit eine OER-Kollektion hierfür implementiert werden.

Die Arbeit lässt sich in drei Abschnitte unterteilen. Zunächst wird der theoretische Hintergrund zu OER ermittelt. Im zweiten Abschnitt werden die einzelnen Schritte im Rahmen der Vorbereitung zur Implementierung der Kollektion ausgeführt. Diese umfassen die Darlegung der Ausgangssituation. Daraus ergibt sich die Zielsetzung, die in dem Aufbau einer OER-Kollektion für ML und erster Materialien für sie besteht. Dann wird die Zielgruppe definiert und eine Einführung in digitale Sammlungen gegeben. Die Planung setzt sich aus mehreren Teilplänen zusammen, wie den Aufbau der Kollektion, Maßnahmen zur Nachnutzbarkeit und einem Zeit- und Ressourcenplan. Bei der Umsetzung wird detailliert auf die Erstellung der Materialien und das Anlegen der Kollektion auf der Online-Plattform GitHub eingegangen. Abschließend werden im Fazit die Leassons learned vorgestellt und ein Ausblick zur Weiterentwicklung erörtert.

2 Open Educational Resources

In diesem Kapitel werden zunächst die unterschiedlichen Ansätze zur Definition des Begriffes Open Educational Resources herausgearbeitet. Daran anschließend wird die Bedeutung von OER auf die Bildungspolitik dargelegt. Dem folgt eine Zusammenfassung der historischen Entwicklung ausgehend von der Open Source Bewegung mit Blick auf die Rolle von OER in den Fachbereichen Data Science und Machine Learning.

2.1 Definition und Bedeutung von OER

Im Allgemeinen wird mit dem Begriff Open Educational Resources frei zugängliches Bildungsmaterial bezeichnet. Die bekannteste Definition ist von der UNESCO aus dem Jahr 2002.

Sie lautet:

„Open Educational Resources (OER) sind Bildungsmaterialien jeglicher Art und in jedem Medium, die unter einer offenen Lizenz stehen. Eine solche Lizenz ermöglicht den kostenlosen Zugang

sowie die kostenlose Nutzung, Bearbeitung und Weiterverbreitung durch Dritte ohne oder mit geringfügigen Einschränkungen. Dabei bestimmen die Urhebenden selbst, welche Nutzungsrechte sie einräumen und welche Rechte sie sich vorbehalten.“ [2]

In den letzten Jahren gab es Bestrebungen, diese Definition zu konkretisieren ausgehend vom Begriff Openness bzw. Offenheit. Materialien können weder den Zustand offen bzw. nicht offen haben, sondern ähnlich wie bei einer Skala gibt es Abstufungen. Entscheidend dabei sind zwei Aspekte. Zum einen muss die Frage nach den Nutzungsrechten und somit nach dem rechtlichen Rahmen, in dem andere auf das Material zugreifen und es verändern dürfen, geklärt werden. Zum anderen spielt die technische Nachnutzbarkeit eine bedeutende Rolle. Daher wurden zwei Ansätze entwickelt, die diese Punkte einbeziehen, die 5 V-Freiheiten (in deutscher Übersetzung) und das ALMS Framework [3]. In der ursprünglich publizierten Version fehlte der Punkt des Verwahrens/Vervielfältigens. Er wurde nachträglich ergänzt, um das Recht auf Eigentum zu unterstreichen, dass nicht automatisch gleichgesetzt werden darf mit dem Zugänglich machen von Materialien. Es muss Nutzenden möglich sein, OER-Materialien kopieren zu können. Diese Kopien gehen dann in den Besitz der Person über [4].

Die 5 V-Freiheiten definieren ein urheberrechtsfähiges Werk als Open Educational Resources, wenn das Recht auf:

- Verwahren/Vervielfältigen,
- Verwenden,
- Verarbeiten,
- Vermischen,
- Verbreiten

besteht. Diese Nutzungsrechte müssen kostenlos und unbegrenzt eingeräumt werden [5].

Das heißt, dass andere das Werk nachnutzen, es an Dritte weitergeben, überarbeiten und vermischen dürfen. Ebenso dürfen sie es verwahren und vervielfältigen. Dieses Recht ist indirekt bereits in den Rechten auf Verwendung und Verbreitung enthalten. Das setzt allerdings voraus, dass nicht nur Zugang zu den OER-Materialien besteht, sondern diese zum Beispiel heruntergeladen werden können.

Die Vergabe der genannten Nutzungsrechte kann durch Wahl einer geeigneten Lizenz gewährleistet werden. Mittlerweile gibt es bereits vorgefertigte Lizenzverträge wie zum Beispiel von der Non-Profit-Organisation Creative Commons [6]. Solche Lizenztypen werden auch als freie Lizenzen bezeichnet [7]. Allerdings kann selbst bei der passenden Wahl die beschriebenen Rechte nicht immer vollumfänglich gewährleistet werden, wenn dem technische Hürden entgegenstehen. Daher war eine Erweiterung um das ALMS-Framework notwendig. Es umfasst vier Fragekategorien:

- Wie sieht es mit dem Zugang zu den Tools aus, die zur Bearbeitung benötigt werden?
- Braucht man zur Bearbeitung der Materialien Fachkenntnisse?
- Wie hoch ist der Aufwand zur Bearbeitung der Materialien?
- Können die Materialien direkt in dem Format bearbeitet werden, indem sie zur Verfügung gestellt wurden? [8]

Die Fragen sollen die urhebende Person von OER-Materialien als Orientierung dienen und dafür sensibilisieren, dass der gesamte Workflow von der Erstellung, dem Zugänglich machen bis hin zur Nachnutzung und Bearbeitung durch andere mitgedacht werden muss.

Somit liefern diese beiden Ansätze, die sowohl rechtliche als auch technische Aspekte einbeziehen, eine wertvolle Ergänzung zu der Definition der UNESCO.

In OER werden weltweit große Hoffnungen gesetzt. Sie sollen bei der Verwirklichung der Sustainable Development Goals (in deutscher Übersetzung Agenda Bildung 2030), welche im September 2015 von der Generalversammlung der Vereinten Nationen verabschiedet wurden, mithelfen. Hierin wurde festgelegt, dass eines der Nachhaltigkeitsziele ist, allen Menschen bis 2030 Zugang zu hochwertiger Bildung zu ermöglichen [9]. Unterstrichen wird dies durch die Verabschiedung einer Empfehlung zu OER von der UNESCO-Generalkonferenz 2019 [2]. Bildung ist abhängig von den zur Verfügung stehenden Ressourcen. Konkret können OER helfen, das Angebot an Lehrmaterialien zu erweitern, Kollaborationen zwischen Lehrenden zu fördern und so die Weiterentwicklung und Verbesserung der Materialien voranzutreiben [10].

Mittlerweile ist OER auch in Deutschland angekommen. Als Beispiel soll hier die OERinfo-Förderrichtlinie vom Bundesministerium für Bildung und Forschung erwähnt werden. Sie ist Teil des Förderprogrammes Digitale Medien in der beruflichen Bildung. Neben dem Aufbau des Online-Portals OERinfo werden auch Projekte gefördert, deren Ziel vor allem die Verankerung von OER in Deutschland ist [11].

Neben der Rolle, die OER auf internationaler und nationaler Ebene spielt, nehmen sie auch im fachspezifischen Kontext von Data Science (DS) und Machine Learning (ML) eine besondere Stellung ein. Dies ist zurückzuführen auf die historische Entwicklung, welche im nächsten Kapitel erörtert wird.

Gleichzeitig können die Vorteile von OER gerade anhand von Machine Learning gut aufgezeigt werden. Der Einfluss digitaler Daten in der Wissenschaft hat immer mehr zugenommen und einen neuen Begriff den der datenintensiven Forschung kreiert. Hier finden ML-Methoden Anwendung zur Erkenntnisgewinnung. Vom Wissenschaftsrat wurde dazu 2020 das Positionspaper »Zum Wandel in den Wissenschaften durch datenintensive Forschung« publiziert. In der Leitlinie 3 zu Kompetenzaufbau und Spezialisierung wird auf den Bedarf innerhalb der Lehre verwiesen [12]. Um diesen zu decken, könnte OER einen Lösungsansatz darstellen. Das Erstellen von Bildungsmaterialien ist zeitintensiv. Eine Veröffentlichung von bereits angefertigten Materialien erlaubt den Austausch und die Nachnutzung. Dadurch können andere Einrichtungen zeitnah Lehrveranstaltungen anbieten und die benötigten Kompetenzen vermitteln. OER hilft Ressourcen zu bündeln.

2.2 Historischer Überblick und aktueller Stand

Die Entwicklungen im Bereich der Informations- und Kommunikationstechnologie in den letzten Jahrzehnten - allen voran des Internets - sind die Basis für die Entstehung von OER. Der Grundgedanke des freien Austauschs von urheberrechtlich geschützten Werken durch die Nutzung der neuen Technologien entstand mit der Gründung des GNU-Projektes 1983 durch Richard Stallman am Massachusetts Institute of Technology (MIT). Diese Idee wurde aufgegriffen und auf Lehrinhalte übertragen. Ein wichtiger Meilenstein war die Einführung einer Open Content Lizenz, initiiert von David Wiley. Daraus gingen 2002 die Creative Commons hervor [13], die sich inzwischen gelungen etabliert haben und über OER hinaus in weiteren Bereichen wie Open Access und Open Data Anwendung finden, wie auch der Appell zur Nutzung offener Lizenzen in der Wissenschaft von der DFG im Jahre 2014 bestätigt [14]. Zeitgleich zu den Creative Commons wurde am MIT die Initia-

tive OpenCourseWare (OCW) gestartet mit dem Ziel, Lehrmaterialien in Form von Kursen online frei zugänglich zu machen. Das war so erfolgreich, dass andere Einrichtungen sich daran beteiligten und schließlich zur Gründung des OpenCourseWare Konsortiums 2008 führte [13]. Bis heute setzt sich das Konsortium mittlerweile unter dem Namen Open Educational Global (OE Global) [15] für die Öffnung von Bildung ein und hat seinen Wirkungskreis immer mehr erweitert. Es sind neue Themen wie zum Beispiel Open Education Technology hinzugekommen. Darunter versteht man die Öffnung der zugrund liegenden Infrastruktur wie Hardware und Software für offene Bildung, indem es ermöglicht werden soll, sich in deren Weiterentwicklung einbringen zu können [16].

Jedoch sind auch viele Bildungsmaterialien, die online zugänglich sind, trotz Bemühungen OER zu etablieren, bis heute nicht nachnutzbar. Ein Beispiel dafür sind die Massive Open Online Courses (MOOCs), die sich parallel zu OCW entwickelt haben. OCW richten sie an Lehrpersonen und Einrichtungen und stehen unter freien Lizenzen. MOOCs hingegen adressieren die Lernenden. Sie sind zwar frei zugänglich, aber die Nutzungsrechte liegen oft bei kommerziellen Anbietern [17]. Hinzu kommt, dass zwar der Zugriff auf den Kurs möglich ist, aber nicht auf die Materialien und dadurch auch die Nachnutzbarkeit selbst bei entsprechender Lizenzierung erheblich erschwert wird. Diese Problematik ist auch in der Machine Learning- und Data Science-Community bekannt. Ein neuer Ansatz für OER wird in dem Paper »Developing Open Source Educational Resources for Machine Learning and Data Science« vorgestellt. Der Begriff OER wird erweitert auf Open Source Educational Resources (OSER) und unterstreicht bereits im Namen die Nähe zu Open Source. Die Besonderheit in Machine Learning und Data Science ist, dass hier sowohl Theorie meist in Form von Folien als auch Quellcode und Daten als Übungsmaterial zum Einsatz kommen. Des Weiteren finden sich im Alltag bei den Lehrenden unterschiedliche Workflows wieder, von der einfachen Nachnutzung von Materialien bis hin zu deren Weiterentwicklung. Um diesen Anforderungen gerecht zu werden, wurden Best-Practices aus dem Open Source Bereich aufgegriffen und auf die Erstellung von Kursen für MS und DS als OER übertragen.

- Empfehlung 1:

  Kollaboratives Arbeiten erhöht die Qualität. Um die Kommunikation zu steuern, bieten sich Tools wie Git an, die auch in der Softwareentwicklung eingesetzt werden.

- Empfehlung 2:

  Es sollte nicht nur der Kurs zugänglich gemacht werden, sondern darüber hinaus auch die Kursmaterialien selbst. Das erlaubt neben einer Erleichterung der Adaptierung auch die Weiterentwicklung einzelner Komponenten. Zusätzlich sollte es eine Qualitätskontrolle durch Maintainer geben. So werden Personen bezeichnet, die ein Open Source Projekt betreuen.

- Empfehlung 3:

  Materialien sollten mit freien Lizenzen versehen werden. Abhängig von der Art des Materials sind entweder Creative Commons Lizenzen geeignet oder eine Open Source Lizenz.

- Empfehlung 4:

  Bei der Erstellung und Weiterentwicklung von Materialien entstehen oft verschiedene Versionen. Es sollte klar erkennbar sein, ob es sich um die Endversion handelt oder um Versionen, die noch in der Bearbeitung sind. Außerdem sollte im Sinne der Nachvollziehbarkeit der Prozess dokumentiert werden, zum Beispiel in einer Change Log Datei.

- Empfehlung 5:

  Inhalte und Lernziele sollten klar definiert sein.

- Empfehlung 6:

  Es sollten Optionen angeboten werden, die eine Selbstlernkontrolle für die Teilnehmenden des Kurses ermöglichen, wie zum Beispiel Musterlösungen.

- Empfehlung 7:

  Anstatt großer Einheiten sollte der Kurs in viele kleine Abschnitte unterteilt sein. Das vereinfacht das Überarbeiten des Kurses.

- Empfehlung 8:

  Es muss die Anpassungsfähigkeit des Kurses an verschiedene Programmiersprachen bedacht werden. Daher ist eine Trennung von Theorie und Codebeispielen sinnvoll.

- Empfehlung 9:

  Literate programming sollte wohlüberlegt eingesetzt werden. Besser ist es, Text und Code voneinander zu trennen, um Fehleranfälligkeiten zu reduzieren.

- Empfehlung 10:

  Feedback ist immer wertvoll, unabhängig von der jeweiligen Erfahrung, die Menschen mitbringen. Daher sollte die Möglichkeit, Feedback zu geben, so weit wie möglich geöffnet werden. [18]

Dieser Ansatz zeigt auf, dass neben der allgemeinen Definition von OER, die im vorherigen Kapitel ausgeführt wurde, disziplinspezifische Erweiterungen einen vielversprechenden Beitrag leisten können. Denn gerade sie beziehen Besonderheiten oder auch bestehende Praktiken mit ein. Das erleichtert Lehrenden die Umsetzung und hilft so bei der Etablierung von OER.

3 Vorbereitung

Ausgehend vom konkreten Bedarf an Lehrmaterialien für die Erstellung eigener Lehrveranstaltungen in den Forschungsgruppen aus dem Programmbereich Data Science and Services an der ZB MED entstand die Idee, eigenständig ein neues Konzept zu initiieren [19]. Bevor mit der Planung begonnen wird, wird sich zuerst durch eine Recherche nach OER-Materialien ein Überblick über den aktuellen Stand verschafft. Die Ergebnisse daraus fließen in die Zielsetzung dieser Masterarbeit ein. Es soll eine OER-Kollektion für Machine Learning implementiert und erste Materialien hierfür erstellt werden. Dem folgen die Festlegung der Zielgruppe und eine Einführung in digitale Sammlungen. Die Planung setzt sich aus mehreren Teilplänen zusammen, die sich auf verschiedene Aspekte, wie Bestimmung des Inhalts der Sammlung, Maßnahmen zur Nachnutzbarkeit oder Ressourcen, beziehen.

3.1 Motivation

Wie schon in Kapitel 2.2 ausgeführt wurde, spielt OER in den Bereichen Data Science und Machine Learning alleine schon aufgrund der historischen Nähe eine relevante Rolle. Darüber hinaus entstand die Idee einer OER-Kollektion speziell für ML-Materialien als Masterarbeitsthema aus den konkreten Erfahrungen, die in den Forschungsgruppen aus dem Programmbereich Data Science and

Services unter Leitung von Herrn Prof. Dr. Konrad Förstner an der ZB MED gemacht wurden. Die Notwendigkeit von ML-Kompetenzen und daraus resultierend deren Vermittlung für den Bereich der Molekularbiologie wird in dem Artikel »A lesson for teaching fundamental Machine Learning concepts and skills to molecular biologists« dargelegt [20]. Für den Kurs »Systems biology: From large datasets to biological insight« im Rahmen des EMBL-EBI Trainingsprogramms von 2021 [21] wurde eine Machine Learning Session entwickelt, die über das generische Repositorium Zenodo [22] und über die Online-Plattform GitHub [23] zugänglich ist. Mit Repositorium ist hier ein Speicherort für digitale Objekte gemeint, die darüber Personen zugänglich gemacht werden können [24]. Im Kontext von GitHub werden Dateien mitsamt ihrer Versionierungshistorie in Form eines Git-Repositoriums zur Verfügung gestellt [25]. In diesem Beispiel kann auf Rohmaterialien wie Code, Folien und Bilder zugegriffen werden. Außerdem sind die Materialien in verschiedenen Formaten herunterladbar. Die Folien wurden als .pdf- und .tex-Datei bereitgestellt, die Bilder zusätzlich als .svg-Datei. Somit ist eine direkte Bearbeitung möglich.

Das Materialien häufig nicht in dieser Art und Weise verfügbar sind, zeigt die Recherche nach OER-Materialien im nächsten Kapitel.

3.2 Recherche nach OER-Materialien

Open Educational Resources gehen Hand in Hand mit dem Aufbau von Infrastrukturen wie zum Beispiel OER-Repositorien (OERR), über die Materialien auffindbar und zugänglich gemacht werden können. Zunächst wird die Suche in verschiedenen OER-Repositorien vorgestellt. Das Informationsportal infoOER bietet einen ersten Einstieg. Eine differenzierte Suche lässt sich über die OER.de-Karte durchführen. Hier kann zum Beispiel unter der Rubrik Service gezielt nach Typ Repositorium, Land Deutschland und Zielgruppe »Tertiärer Bereich bis Master« gefiltert werden [26].

Folgende Treffer wurden erzielt:

- The Virtual Linguistics Campus (OER),
- DuEPublico,
- TIB AV-Portal,
- Digitale Sammlungen der Bayrischen Staatsbibliothek,
- HHU Mediathek,
- rpi-virtuell,
- OpenLearnWare (OLW),
- Freiburger Multimedia Object Repository (FREIMORE).

In diesen wurde mit den Suchbegriffen künstliche Intelligenz, Artificial Intelligence, maschinelles Lernen und Machine Learning gesucht.

Bei der E-Learning-Plattform The Virtual Linguistics Campus für Kurse aus dem Fachgebiet der Linguistik und den Sprachstudien konnte für keinen der Suchbegriffe ein Treffer erzielt werden. Von der Universität Duisburg-Essen wird der zentrale Dokumenten- und Publikationsserver DuEPublico betrieben, welcher von Studierenden und Mitarbeitenden als Plattform für Lehr- und Lernmaterialien genutzt werden kann. Für den Suchbegriff künstliche Intelligenz wurden zwei Treffer erzielt. Dabei handelte es sich um Videodateien, die indes bei ihren Metadaten keine Lizenzangaben enthielten. Das TIB AV-Portal ist, wie der Name schon sagt, spezialisiert auf audiovisuelle

Medien. Hier wurden für die vier Suchbegriffe eine Vielzahl von Treffern gefunden, die wiederum über diverse Filter eingeschränkt werden können. Zumindest gibt es einen Filter für Wiederverwendung, über den sich die Option Open Access auswählen lässt. Allerdings muss dann durch die einzelnen Ergebnistreffer geklickt werden, um Informationen zu den Lizenzangaben zu erhalten. Ähnlich verhält es sich bei der HUU Mediathek der Heinrich-Heine-Universität Düsseldorf. Zwar konnten für die genannten Suchbegriffe insgesamt 37 Ergebnistreffer erhalten werden. Es lässt sich jedoch nicht gezielt nach den Lizenzangaben filtern. Das Ziel der Digitalen Sammlung der Bayrischen Staatsbibliothek ist die Digitalisierung von Printmedien und nicht die Bereitstellung von OER-Materialien. Daher wurde sie übersprungen. Das virtuelle religionspädagogische Institut rpi-virtuell vereint verschiedene Angebote. Zusätzlich zur Bereitstellung von Materialien geht es auch um die Vernetzung der Fachcommunity. Für den Begriff künstliche Intelligenz lassen sich immerhin vier Treffer finden, nachdem zusätzlich über den Filter Lizenz »Zur nicht kommerziellen Wiederverwendung und Veränderung gekennzeichnet« aktiviert wurde und für Artificial Intelligence ein Treffer. Diese Materialien richten sich gezielt an Lehrpersonen. Die Online-Plattform OpenLearnWare (OLW) der TU Darmstadt greift in ihrem Namen die OpenCourseWare-Initiative des MIT auf. Sie bietet Vorlesungen an, die unter CC-Lizenzen stehen. Unter dem Begriff machine learning findet sich die höchste Trefferzahl mit 16 Ergebnissen. Das Freiburger Multimedia Object Repository (FREIMORE) hat seinen Dienst eingestellt.

Diese Art der Suche ist sehr zeitaufwendig. Bevor mit der Suche nach OER-Materialien begonnen werden kann, steht zuerst die Suche nach geeigneten Plattformen an.

Es wird nun eine zweite Suchoption vorgestellt, die eine Alternative zum Durchsuchen eines Repositoriums nach dem anderen anbietet. Dabei handelt es sich um den Open Educational Resources Search Index (OERSI). Dieser bietet einen zentralen Sucheinstieg an und vereint 15 Datenquellen in einer Metasuche (Stand: März 2022). Ins Leben gerufen wurde das Projekt von der Hochschulbibliothek des Landes Nordrhein-Westfalen (hbz) und der Technischen Informationsbibliothek (TIB) 2020. Erwähnenswert ist es, dass es sich dabei um eine Open-Source-Entwicklung handelt [27]. Das ermöglicht Nutzenden zum Beispiel Wünsche für neue Features oder Fehlermeldungen aktiv einzubringen und greift den Punkt Open Education Technology aus dem Programm des OE Global auf [16]. Die Suche findet bei Machine Learning die höchste Treffermenge mit 371 Einträgen. Angenehm für die Nutzenden ist die Vereinheitlichung der verschiedenen Metadaten aus den einzelnen Datenquellen. Ebenso ist die Übersichtlichkeit der Informationen auf der Ergebnisseite mit direkter Angabe des Materialtyps wie Präsentation oder Video und der Lizenzangaben eine Erleichterung bei der Suche. Zusätzlich stehen diverse Filter zur Verfügung wie Fachgebiet, Material, Lizenz und weitere. Was deutlich auffällt, ist, dass der Typ Video für den Suchbegriff Machine Learning mit 362 Einträgen am höchsten liegt.

Auf internationaler Ebene gibt es diesen Ansatz bereits. Beispielhaft werden hier die Metasuchdienste OER Commons (seit 2007) [28] und Merlot (seit 1997) [29] ausgewählt.

OER Commons bezeichnet sich als öffentliche digitale Bibliothek für OER [30]. Es ermöglicht die Suche durch zahlreiche Provider. Für den Begriff Machine Learning mit der Filteroption Graduate/Professional in der Rubrik Education Level wurden 15 Treffer erhalten. Für den Begriff Artificial Intelligence lag der Wert mit derselben Einstellung sogar bei 19 Treffer. Interessant ist ein Blick in die Filteroptionen bei Material Type. Es ist eine Auswahl möglich zwischen Activity/Lab, Full Course, Lecture, Lecture Notes und weiteren. Das zeigt, dass der Fokus auf Kursen liegt.

MERLOT steht für Multimedia Education Resource for Learning and Online Teaching. Es enthält kuratierte Bildungsmaterialien von zahlreichen Mitgliedern. Der Sucheinstieg erfolgt über die

Smart Search, welche den Wechsel zwischen der Suche in der MERLOT Collection, anderen Bibliotheken und dem Web erlaubt [31]. In der MERLOT Collection wurden mit den Einstellungen Graduate School und Professional bei Audience und zusätzlich dem Filter »Has a Creative Commons License« 42 Treffer erhalten. Auch hier wird deutlich, wenn sich die Verteilung der Treffer in der Rubrik Material Type angeschaut wird, dass der Schwerpunkt auf Kursen liegt. Für den Filter Online Course gibt es 21 Einträge. Für den Suchbegriff Artificial Intelligence und derselben Filterauswahl waren es 72 Treffer. Hier fällt die Bilanz sogar noch deutlicher aus mit 43 Treffern unter der Filteroption Online Course und 15 Treffern bei Open (Access) Textbook.

Zusammenfassend kann man sagen, dass durch Metasuchdienste die Suche nach OER-Materialien deutlich vereinfacht wird. In den deutschsprachigen Suchdiensten überwogen eindeutig die Treffer für Videomaterial aus dem TIB AV-Portal und über OERSI. In den internationalen Metasuch-Portalen fanden sich mehrheitlich Kurse und Literatur. Bei der Lizenzvergabe fällt auf, dass die Materialien oft nicht den Vorgaben der 5 V-Freiheiten entsprechen. Auch die technische Nachnutzbarkeit, die sowohl im ALMS-Framework als auch in dem Artikel »Developing Open Source Educational Resources for Machine Learning and Data Science« thematisiert wird, findet oft wenig Berücksichtigung. Zugriff wird auf die fertigerstellten Materialien gewährt, nicht auf das Rohmaterial. Gerade das wäre aber für Lehrkräfte wichtig, die nicht den gesamten Kurs übernehmen möchten, sondern nur Interesse an einzelnen Elementen daraus haben.

3.3 Zielgruppe und Zielsetzung

Für Data Science und Machine Learning gibt es einen disziplinübergreifenden Bedarf an Unterricht und damit einhergehend an Unterrichtsmaterial. Gleichzeitig muss bei der Umsetzung von Lehrkonzepten auf den individuellen Bedarf der Teilnehmenden solcher Veranstaltungen Rücksicht genommen werden. Es gibt somit nicht den einen Kurs, der fachübergreifend allen angeboten werden kann. Es gibt jedoch Ansätze, wie das Beispiel der Machine Learning Session für den Kurs »Systems biology: From large datasets to biological insight« zeigt, Materialien so bereitzustellen, dass eine hohe Adaptierung realisierbar ist [20].

Daraus entstand die Idee eine OER-Kollektion aufzubauen. Losgelöst von einem konkreten Kurskonzept bietet sie eine weitaus größere Flexibilität bezüglich der Zusammenstellung einzelner Materialien an. Ein besonderer Fokus wird auf der Bereitstellung von Bildern liegen. Visualisierungen unterstützen den Lernprozess und helfen, komplexe Themen ansprechend und verständlich zu vermitteln. Ihr Gebrauch ist oft sowohl fachübergreifend und als auch sprachübergreifend möglich. Gleichzeitig ist aber die Erstellung zeitaufwendiger als die von Texten. Daher ist davon auszugehen, dass sie für Lehrkräfte von besonderem Interesse sind.

Konkret wird es in der Masterarbeit darum gehen, die Kollektion zu implementieren und erste Materialien über sie zur Verfügung zu stellen.

Die primäre Zielgruppe für die Nutzung einer solchen Kollektion sind Lehrkräfte aus dem Hochschulbereich. Dies können Hochschulprofessor:innen und Dozent:innen sein, aber auch Doktorand:innen und Habilitand:innen, die oft in die Vorbereitung und Durchführung von Lehrveranstaltungen involviert sind, sowie studentische Hilfskräfte. Neben dem Hochschulbereich gibt es den Weiterbildungssektor. Da die Materialien in der Kollektion unter eine CC-BY 4.0 Lizenz gestellt werden, können sie auch von nicht-öffentlichen Bildungseinrichtungen nachgenutzt werden. Denn diese Lizenz erlaubt die kommerzielle Nutzung [32]. Ebenfalls kommen als eine weitere Zielgruppe Lehrkräfte für Ausbildungsberufe infrage.

3.4 Merkmale einer digitalen Sammlung

Eine Sammlung lässt sich durch verschiedene Merkmale charakterisieren. Zum einen definiert sie sich über ihren jeweiligen Gegenstand zum Beispiel ein Themengebiet. Es können aber auch andere Spezifika sein wie eine Person, Zeitepoche oder Region. Analoge Sammlungen haben als weiteres Merkmal, dass sie sich im Besitz befinden. Außerdem wird eine Sammlung fachlich betreut und weiterentwickelt.

Digitale Sammlungen unterscheiden sich davon. Zwar gibt es auch hier einen Gegenstand der Sammlung und sie werden betreut und weiterentwickelt. Darüber hinaus bieten sie aber völlig neue Chancen an. Sie können viel mehr Objekte aggregieren. Der Zugriff ist orts- und zeitunabhängig und es sind völlig neue Formen des Arbeitens mit ihren Objekten durchführbar [33].

In dem Artikel »Die Sammlung ist tot, es lebe die Sammlung!« wird darauf verwiesen, dass dies aber nur umgesetzt werden kann, wenn bestimmte Voraussetzungen erfüllt sind. Es werden die FAIR-Prinzipien aufgegriffen [34]. Sie wurden 2016 publiziert und sollen Forschenden helfen ihre Forschungsdaten sowohl technisch als auch für Menschen nachnutzbar zu machen. Dabei beinhalten sie die vier Forderungen nach Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit [35]. Aufgrund ihrer allgemeinen Formulierung und ihrem Fokus auf Metadaten lassen sie sich jedoch generell auf digitale Objekte anwenden.

Die letzte Forderung nach der Wiederverwendbarkeit legt ihren Fokus auf die Nachnutzbarkeit durch den Menschen. Hier geht es nicht nur um die Verständlichkeit des Inhalts durch Beschreibung mit Metadaten und Dokumentationen, sondern des Weiteren um den rechtlichen Rahmen. Objekte müssen mit eindeutigen Nutzungsrechten versehen sein, um ihre Nachnutzbarkeit gewährleisten zu können. Übertragen auf den Sammlungsbegriff bedeutet dies, dass die Vorstellung vom Besitz einer Sammlung für die digitale Form revidiert werden muss. Gerade die Möglichkeiten des Teilens und der Verbreitung stellen den Mehrwert dar. Dadurch nimmt die Sammlung eine neue Rolle in den digitalen Informationsinfrastrukturen ein [34].

3.5 Planung

Die Planung kann in mehrere Schwerpunkte gegliedert werden. Die Hauptaufgaben sind der Aufbau der Kollektion und die Erstellung von OER-Materialien. Dies benötigt einen Zeit- und einen Ressourcenplan. Für die Materialien muss eine inhaltliche Eingrenzung vorgenommen werden und die Nutzungsbedingungen müssen festgelegt werden. Da bei OER die Nachnutzbarkeit gegeben sein muss, wird ähnlich zu einem Datenmanagementplan bei Forschungsprojekten ein zusätzlicher Plan angelegt, der diesen Punkt einbindet. Zuletzt müssen Überlegungen getroffen werden, wie die Kollektion aufgebaut und nach Beendigung der Masterarbeit weitergeführt werden soll.

3.5.1 Planung der Kollektion

Begonnen wird mit der Planung der Kollektion. Allgemein kann der Aufbau einer Sammlung in mehrere Schritte unterteilt werden:

- Bestimmung des Gegenstandes,
- Auswahl der Objekte,
- Metadaten erheben,
- Aufnahme. [36]

Schritt 1: Bestimmung des Gegenstandes

Zuerst erfolgt eine Festlegung des Gegenstandes. Die Kollektion, die aufgebaut werden soll, hat zwei Merkmale. Die Objekte sollen vom Typ her OER-Materialien sein. Das zweite Merkmal schränkt die Objekte thematisch auf das Gebiet Machine Learning ein.

Schritt 2: Auswahl der Objekte

Wie im Kapitel 3.2 gezeigt, ist es zurzeit sehr schwierig im Internet OER-Materialien zu finden, die sich gut adaptieren lassen. Daher muss in diesem Fall von der Vorgehensweise der Auswahl von Objekten Abstand genommen werden. Die Objekte werden selbst hergestellt. Dies führt dazu, dass sich Aufgaben, die ansonsten von verschiedenen Personen ausgeführt werden, vereinen. Die Person, welche die Kollektion aufbaut, ist auch die urhebende Person.

Trotzdem müssen Kriterien definiert werden, die eine einheitliche Struktur der Kollektion gewährleisten. Das ist auch unter dem Gesichtspunkt, dass nach Beendigung der Masterarbeit die Kollektion kollaborativ weitergeführt werden soll, relevant. Thematisch lässt sich Machine Learning in drei Kategorien unterteilen: überwachtes Lernen, unüberwachtes Lernen und verstärkendes Lernen. Jede der Kategorien lässt sich weiter differenzieren gemäß der jeweiligen Algorithmen. Das Ziel ist es, den Lehrstoff in möglichst elementare Einheiten zu zerlegen. Für jede dieser Einheiten sollen drei Typen von Material erstellt werden: Grafiken, Beispielcode und Erklärtexten zu den Grafiken. Neben diesen Aspekten muss die Nachnutzbarkeit berücksichtigt werden. Darunter fallen zum Beispiel Nutzungsrechte. Der Inhalt der Kollektion wird unter eine CC-BY 4.0 Lizenz gestellt. Eine rechtliche Besonderheit, auf die im Kapitel 3.5.4 eingegangen wird, sind die im ML verwendeten Daten, die ebenfalls kontrolliert werden müssen. Die Objekte müssen folgenden Kriterien entsprechen:

- Sie müssen inhaltlich zum Gegenstand der Kollektion passen.
- Sie müssen sich vom Materialtyp her in die Kollektion einordnen lassen.
- Die Metadaten müssen vollständig sein.
- Sie müssen unter einer CC-BY 4.0 Lizenz stehen.
- Wenn externe Quellen verwendet wurden, so müssen diese unter einer CC-BY 4.0 Lizenz stehen oder vergleichbar.

Schritt 3: Metadaten erheben

Für jedes Objekt werden Metadaten erhoben. Sie orientieren sich an der Nutzungslizenz und umfassen Angaben zum Titel, Autor:in, Quelle und der Lizenz. Des Weiteren werden durch Tools wie die Online-Plattform GitHub automatisiert Metadaten erfasst.

Schritt 4: Aufnahme

Die Aufnahme in die Kollektion erfolgt nach Prüfung des Objektes. Dazu wird ein Qualitätssicherungsprozess implementiert, welcher im Kapitel 4.3.1 beschrieben wird.

3.5.2 Ermittlung inhaltlich relevanter Themen

Als Nächstes folgt die inhaltliche Eingrenzung. Die Kollektion startet bei den Grundlagen im Machine Learning. Zur Orientierung wird das Buch »Einführung in Machine Learning mit Python« von Andreas C. Müller und Sarah Guido genommen, welches mit dem Bereich des überwachten Lernens einsteigt. Es werden die wichtigsten Algorithmen vorgestellt, die sich entsprechend der

Fragestellung zum Teil in ein Klassifikations- und ein Regressionsproblem einteilen lassen [37]. Somit ergeben sich als Einheiten:

- k-nächste-Nachbarn:
  - k-nächste-Nachbarn-Klassifikation,
  - k-nächste Nachbarn-Regression,
- lineare Modelle:
  - lineare Modelle zur Regression,
  - Ridge-Regression,
  - lineare Modelle zur Klassifikation,
- naive Bayes-Klassifikatoren,
- Entscheidungsbäume:
  - Eigenschaften,
  - Random Forests,
  - Gradient Boosting Machines,
  - Support Vector Machines mit Kernel,
- Neuronale Netze (Deep Learning).

Da das Ziel der Masterarbeit in der Implementierung der Kollektion besteht, geht es nicht darum alle Algorithmen aus der Liste abzuarbeiten, sondern den Anfang bei den Materialien zu machen. Daher wird der Umfang offengelassen.

3.5.3 Plan zur Nachnutzbarkeit

Die Nachnutzbarkeit hat bei OER-Materialien Priorität. Im ersten Kapitel wurden bereits die 5 V-Freiheiten und das ALMS-Framework eingeführt [8]. Zusätzlich wurde auf fachspezifische Empfehlungen eingegangen [18]. Die FAIR-Prinzipien sind im Kapitel 3.4 hinzugekommen [34]. Bevor in die konkrete Planung hineingegangen wird, werden die unterschiedlichen Ansätze miteinander verglichen.

FAIR-Prinzipien	5 V-Freiheiten	ALMS-Framework	fachspezifische Empfehlungen
Auffindbarkeit: • „F1. (Meta)Daten wird ein global eindeutiger und dauerhaft persistenter Identifier zugewiesen“[38] • „F2. Daten werden mit umfangreichen Metadaten (vergl. R.1) beschrieben“[38] • „F3. Metadaten enthalten klar und eindeutig den Identifier, der die Daten referenziert“[38] • „F4. Metadaten werden in einem durchsuchbaren Verzeichnis registriert oder indiziert “[38]			
Zugänglichkeit: • „A1. (Meta)Daten sind über ihren Identifier mithilfe eines standardisierten Kommunikationsprotokolls auffindbar“[38] • „A1.1 Das Protokoll ist offen, frei und universell implementierbar“[38]			

FAIR-Prinzipien	5 V-Freiheiten	ALMS-Framework	fachspezifische Empfehlungen
• „A1.2 Das Protokoll unterstützt, wo notwendig, die Authentifizierung und Rechteverwaltung“[38] • „A2. Metadaten sind/bleiben verfügbar, auch für den Fall, dass die zugehörigen Forschungsdaten nicht mehr vorhanden sind “[38]			
Interoperabilität: • „I1. (Meta)-Daten nutzen eine formale, zugängliche, gemeinsam genutzte und breit anwendbare Sprache für die Wissensrepräsentation“[38] • „I2. (Meta)Daten benutzten Vokabulare, welche den FAIR Prinzipien folgen“[38] • „I3. (Meta)Daten enthalten qualifizierte Referenzen auf andere (Meta)Daten“[38]			

FAIR-Prinzipien	5 V-Freiheiten	ALMS-Framework	fachspezifische Empfehlungen
Wiederverwendbarkeit: • „R1. (Meta)Daten sind detailliert beschrieben und enthalten präzise, relevante Attribute“[38] • „R1.1. (Meta)Daten enthalten eine eindeutige, zugreifbare Angabe einer Nutzungslizenz“[38] • „R1.2. (Meta)Daten enthalten detaillierte Provenienz-Informationen“[38] • „R1.3. (Meta)Daten entsprechen den fachgebietsrelevanten Community Standards“[38]	• Verwahren und Vervielfältigen • Verwenden • Verarbeiten • Vermischen • Verbreiten [5]	• Wie sieht es mit dem Zugang zu den Tools aus, die zur Bearbeitung benötigt werden? • Braucht man zur Bearbeitung der Materialien Fachkenntnisse? • Wie hoch ist der Aufwand zur Bearbeitung der Materialien? • Können die Materialien direkt in dem Format bearbeitet werden, indem sie zur Verfügung gestellt wurden? [8]	• Empfehlung 1: kollaboratives Arbeiten • Empfehlung 2: einzelne Kursmaterialien zugänglich machen • Empfehlung 3: Lizenz • Empfehlung 4: Kennzeichnung der Endversion • Empfehlung 5: Angaben zu Inhalt und Lernzielen • Empfehlung 6: Selbstlernkontrolle • Empfehlung 7: kleine Einheiten • Empfehlung 8: Trennung Theorie und Code • Empfehlung 9: Trennung Code und Text empfohlen • Empfehlung 10: Feedbackfunktion [18]

Aus der Tabelle ist sofort ersichtlich, dass die ersten drei Prinzipien, die Auffindbarkeit, die Zugänglichkeit und die Interoperabilität in den bisherigen Ansätzen zu OER nicht eingegangen sind. Hingegen greifen die 5 V-Freiheiten den Punkt R1.1 aus der Wiederverwendbarkeit auf und führen ihn weiter aus, indem sie die zu vergebenden Nutzungsrechte aufzählen. Sie verlangen aber nicht die Zugreifbarkeit auf die Lizenz. Das ALMS-Framework lässt sich ebenfalls unter dem letzten Prinzip einordnen. Allerdings gibt es hier keinen eindeutigen Bezug zwischen den Fragestellungen und den Punkten R1 bis R1.3. Das kann so erklärt werden, dass OER nicht so spezialisiert sind wie Forschungsdaten. Zuletzt finden sich die fachspezifischen Empfehlungen gleichfalls bei der Wiederverwendbarkeit. Hier gibt es durchaus Übereinstimmungen. Die Angabe einer Lizenz in der

Empfehlung 3 ist in Punkt R1.1 enthalten. Die Empfehlungen 4 und 5 knüpfen an R1 an und beziehen sich auf die Verständlichkeit. Zuletzt kann der Punkt R1.3, der sich auf die Einhaltung fachspezifischer Standards bezieht, auf alle 10 Empfehlungen übertragen werden.

Der Vergleich zeigt, dass allein die Berücksichtigung der Ansätze zur Nachnutzbarkeit von OER nicht ausreichend sind für die Kollektion. Denn sie konzentrieren sich nur auf die Wiederverwendbarkeit. Dabei könnten sich die anderen drei Prinzipien durch die Nutzung von Repositorien, welche die FAIR-Prinzipien unterstützen, umsetzen lassen. So helfen Eingabemasken bei der Vergabe von Metadaten konform zu etablierten Standards. Oft werden auch persistente Identifikatoren automatisiert vergeben.

Als Fazit aus dem Vergleich könnte man für OER zwei weitere Empfehlungen formulieren.

- Materialien sollten in einem Repositorium veröffentlicht werden.
- OER-Repositorien sollten die FAIR-Prinzipien unterstützen.

Nach diesen Überlegungen wird im nächsten Schritt ein Plan zur Nachnutzbarkeit angelegt. Als Orientierung wird die Vorlage eines Datenmanagementplans für im Rahmen des EU-Forschungsprogrammes Horizon Europe geförderter Forschungsprojekte genommen [39]. Diese bietet sich insofern an, weil der Fokus auf den FAIR-Prinzipien liegt. Anhand der Leitfragen aus dem Abschnitt 2 zu FAIR data werden Maßnahmen hergeleitet.

Auffindbarkeit:

Es ist geplant, die Materialien zuerst über die Online-Plattform GitHub in einem öffentlichen Git-Repositorium bereitzustellen. Im nächsten Schritt soll die Integration zwischen GitHub und dem Repositorium Zenodo zur Veröffentlichung genutzt werden. Dadurch erhält das Git-Repositorium eine DOI (Digital Object Identifier) [40]. Bei der Erstellung der Materialien werden für jedes einzelne Objekte folgende Metadaten erhoben: Titel, Autor:in, Quelle und Lizenz. Zusätzlich werden automatisiert über GitHub Metadaten gesammelt bezüglich der Versionierung. Des Weiteren wird die Kollektion mit Schlagwörtern versehen. Die Metadaten in Zenodo stimmen mit dem generischen DataCite-Metadatenschema überein [41]. Die Einträge in Zenodo sind über die Suchfunktion in Zenodo auffindbar. Darüber hinaus werden sie durch die DOI in dem Index des internationalen Konsortiums und der DOI-Registrierungsagentur DataCite geführt [42].

Zugänglichkeit:

Repositorium:

Die OER-Materialien werden über ein öffentliches Git-Repositorium auf GitHub und über Zenodo zugänglich gemacht. Die Benutzung von Zenodo steht jedem frei unter Beachtung der Nutzungsbedingungen. Unter anderem ist es erlaubt, Bildungsinhalte zu publizieren [43]. Jeder Eintrag wird automatisch mit einer DOI versehen [44]. Außerdem plant Zenodo eine CoreTrustSeal-Zertifizierung zu beantragen [45]. GitHub bietet drei Zugänge an. In diesem Fall wird der Zugang GitHubFree gewählt, der eine kostenlose Nutzung vorsieht [46].

Daten:

Die Materialien werden alle öffentlich zugänglich gemacht. Sie sind für jeden frei herunterladbar. Von der technischen Seite her stellt Zenodo über Protokolle wie OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) [47] und REST-API (Representational State Transfer

Programmierschnittstelle) [48] eine freie Zugänglichkeit her.

Metadaten:

Ebenso werden die Metadaten frei zugänglich gemacht. Entsprechend der Nutzungsbedingungen basierend auf der Version v1.2 ist in Punkt 7 festgelegt, dass Metadaten, die über Zenodo veröffentlicht werden, automatisch unter einer CC0 1.0 Lizenz stehen, außer anderes ist vermerkt [43]. Die langfristige Zugänglichkeit wird bei Zenodo vom Betreiber CERN auf mindestens 20 Jahre zugesagt [41]. Darüber hinaus gibt es Kooperationen zwischen GitHub und zum Beispiel dem Software Heritage Archive, die es ermöglichen, öffentliche Repositorien archivieren zu lassen [49]. Diese angebotene Funktion soll genutzt werden.

Interoperabilität:

Zum einem wird durch die Nutzung von Zenodo gewährleistet, dass die Metadaten dem DataCite-Metadatenschema entsprechen. Darüber hinaus ist es möglich, sie in dem Dateiformat JSON herunterzuladen [41]. Es wird bewusst darauf geachtet, dass die Objekte in offenen Dateiformaten wie SVG, welches auf XML basiert und IPYNB, welches auf JSON basiert, abgespeichert werden. Die Texte werden in Markdown geschrieben. Alle Objekte lassen sich mit Open-Source-Software bearbeiten.

JSON, XML und Markdown sind Auszeichungssprachen. Das heißt, dass die Daten strukturiert erfasst werden und dadurch maschinenlesbar sind.

Nachnutzbarkeit:

Informationen zu den Materialien werden über das Git-Repositorium bereitgestellt in README-

Dateien. Die Versionierungskontrolle erlaubt zudem die Rückverfolgung der einzelnen Bearbeitungsschritte jedes der Objekte. Es wird ein Qualitätssicherungsprozess implementiert, welcher im Kapitel 4.3.1 ausführlich beschrieben wird. Die Objekte werden unter eine CC-BY 4.0 Lizenz gestellt.

3.5.4 Auwahl der Lizenz

Die Auswahl der Lizenz sollte mit den 5 V-Freiheiten übereinstimmen. Es bietet sich an, auf die Standard-Lizenzverträge der Creative Commons zurückzugreifen. Es gibt vier Bedingungen:

- by (Attribution) für Namensnennung,
- sa (share alike) für die Weitergabe unter gleichen Bedingungen,
- nc (non commercial) für nicht-kommerziell,
- nd (no derivates) für keine Bearbeitung.

Aus diesen können durch Kombination 6 Lizenzen gebildet werden:

- CC-BY,
- CC-BY-SA,
- CC-BY-NC,
- CC-BY-ND,
- CC-BY-NC-SA,

- CC-BY-NC-ND.

Überdies gibt es noch die Möglichkeit, Werke in die Public Domain zu überführen mit einer CC0 Lizenz. Damit werden alle Nutzungsrechte abgetreten [50]. Um eine größtmögliche Nachnutzbarkeit zu begünstigen, wird die CC-BY Lizenz empfohlen. Sie erlaubt das Teilen und Bearbeiten auch zur kommerziellen Nutzung [51]. Daher wurde sich für diese entschieden in der Version 4.0.

Bei den Materialien handelt es sich auch um Codebeispiele. In den fachspezifischen Empfehlungen wurde das Thema der unterschiedlichen Materialtypen angesprochen und neben den Creative Commons Lizenzen auf Open-Source-Lizenzen verwiesen [18]. Hier werden alle Objekte unabhängig vom Typ unter eine CC-BY 4.0 Lizenz gestellt, weil die Codebeispiel als Lehrmaterial angedacht sind.

Eine weiterer wichtiger Punkt sind die Daten, die gebraucht werden, um maschinelle Lernmodelle zu entwickeln. Daten können dem Urheberrecht oder verwandten Schutzrechten unterliegen. Dabei stellt die Datenanalyse an sich keinen Verstoß dar. Problematisch kann jedoch die Datenbeschaffung und -aufbereitung sein. Das könnte als Vervielfältigung gelten [52]. Somit muss darauf geachtet werden, dass die Daten, die in den Codebeispielen verwendet werden, unter einer CC-BY 4.0 Lizenz oder vergleichbar stehen.

3.5.5 Auswahl der Tools

Die Auswahl der Tools richtet sich nach den Aktivitäten unter Einbeziehung des ALMS Frameworks [8]. Es lassen sich 6 Aktivitäten identifizieren:

- die Erstellung des Codes,
- die Erstellung der Grafiken,
- die Erstellung der Erklärtexte,
- die Materialien öffentlich zugänglich machen,
- die Materialien kollaborativ bearbeiten,
- die Materialien publizieren.

Alle Materialien sollen mit Open-Source-Software erstellt und in offenen Dateiformaten abgespeichert werden, in denen sie auch direkt weiterbearbeitet werden können. Bezüglich der benötigten Fachkenntnisse muss hier berücksichtigt werden, dass sich die Materialien an die ML- und DS-Community richten. Daher wird bei der Auswahl darauf geachtet, dass die Tools in diesen Communities bekannt sind.

Für die Erstellung des Codes wird die Open-Source-Software Jupyter Notebooks [53] verwendet, die für wissenschaftliche Zwecke entwickelt wurde. Die Jupyter-Notebook-Dokumente bestehen aus Eingabe- und Ausgabezellen. Diese können Code, Texte und Plots enthalten. Das Dateiformat ist IPYNB. Inzwischen werden zahlreiche Programmiersprachen unterstützt [54]. Für die Codebeispiele wird Python genommen, wie auch in dem Buch »Einführung in Machine Learning mit Python«[37].

Die Grafiken werden mit der Open-Source-Software Inkscape angefertigt [55]. Sie dient der Erstellung von Vektorgrafiken. Im Unterschied zu Rastergrafiken, in denen ein Bild aus Bildpunkten in einem Raster besteht, setzt sich in einer Vektorgrafik ein Bild aus Objekten zusammen. Das Dateiformat ist SVG [56].

Die Erklärtexte werden in Markdown geschrieben. Es gibt zahlreiche Markdown-Editoren. Einer

davon ist Atom, welcher Open-Source ist und in der Masterarbeit zum Einsatz kommt [57].

Neben der Erstellung der Materialien spielt auch die Bereitstellung eine Rolle. Hier fiehl die Wahl auf die Online-Plattform GitHub. Die Materialien können über ein öffentliches Git-Repositorium zur Verfügung gestellt werden und sind frei herunterladbar.

Dies ist aber nicht der alleinige Grund, warum sich für GitHub entschieden wurde. Ein weiterer sehr wichtiger Punkt ist die Fortführung der Kollektion nach Beendigung der Masterarbeit. GitHub bietet Features zum kollaborativen Arbeiten und zum Projektmanagement an [58].

Um die Materialien darüber hinaus als Publikation verfügbar zu machen, wurde sich für das Repositorium Zenodo entschieden. Die Wahl ergab sich aus den folgenden Gründen. Es existiert eine Integration zwischen GitHub und Zenodo, die die technische Umsetzung vereinfacht [40]. Das Repositorium unterstützt die FAIR-Prinzipien [41]. Es erlaubt die Publikation von Bildungsmaterialien [43] und die Machine Learning Session aus dem Kurs »Systems biology: From large datasets to biological insight« wurde hierüber ebenfalls publiziert [22].

3.5.6 Fachspezifische Ergänzungen

Zuletzt fließen in die Planung die fachspezifischen Empfehlungen mit ein [18]. Einige wurden schon aufgegriffen. So wird durch GitHub das kollaborative Arbeiten [58] und eine Versionskontrolle [59] zur Dokumentation ermöglicht. Über README-Dateien, die auf der Ebene der einzelnen Ordner angelegt werden können, lassen sich Informationen zum Inhalt angeben. Es werden verschiedene Materialtypen erstellt, die so die Trennung von Theorie und Code umsetzen. GitHub bietet Issues an [60]. Sie können zur Kommunikation wie zum Beispiel für Feedback verwendet werden. Durch die Vergabe von Labels lassen sie sich in eine Vielzahl von Arbeitsabläufen integrieren [61].

Auf zwei Empfehlungen wird nicht eingegangen. Da es sich um eine Kollektion und nicht um einen Kurs handelt, entfällt die Selbstlernkontrolle. Auch wird durch Verwendung von Jupyter-Notebooks-Dokumenten auf eine Trennung von Code und Text verzichtet. Dies begründet sich dadurch, dass die Codebeispiele Unterrichtsmaterial sind. Die Texte in dem Code dienen als Erklärhilfen.

3.5.7 Zeitplan

Insgesamt stehen für die Implementierung der Kollektion und die Erstellung erster Materialien 6 Monate zur Verfügung. Grob wird der Zeitrahmen eingeteilt in: die Einarbeitungsphase, die Implementierungsphase und das Verfassen der wissenschaftlichen Arbeit.

Zur Einarbeitung werden 2 Monate eingeplant. Die Umsetzung der Aufgaben soll nicht länger als maximal 3 Monate in Anspruch nehmen. Dabei ist es anvisiert, dass der Hauptteil bestehend aus dem Aufbau der Kollektion und die Erstellung der Materialien nach 2 Monaten abzuschließen sind. Das Zugänglich machen der Materialien und das Publizieren über Zenodo können begleitend zur Verfassung der wissenschaftlichen Arbeit vorgenommen werden, in der verbleibenden Zeit von 2 Monaten.

4 Umsetzung

Beginnend mit einer Darlegung, der in der Masterarbeit angewandten Vorgehensweise, werden im Weiteren die beiden Hauptaufgaben, die Erstellung der verschiedenen Lernmaterialien und der Aufbau der Kollektion als Git-Repositorium auf der Online-Plattform GitHub ausgeführt. Zuletzt wird eine Organisation auf GitHub angelegt, die ein flexibles Zusammenarbeiten ermöglicht.

4.1 Vorgehensweise

Im ersten Schritt wurde eine Literaturrecherche durchgeführt, um die zentralen Konzepte und fachspezifische Ansätze im OER zu ermitteln. Des Weiteren ging es um die Eruierung der historischen Entwicklung von OER mit Blick auf den bildungspolitischen Einfluss in Deutschland und auf internationaler Ebene.

Dem folgte eine Recherche nach OER-Materialien für Machine Learning in mehreren ausgewählten Online-Verzeichnissen, um die persönlichen Erfahrungen, die in den Forschungsgruppen aus dem Programmbereich Data Science and Services an der ZB MED gemacht wurden, zu belegen. Es zeigte sich, dass die Suche nach geeignetem Material zeitintensiv ist und wenig brauchbare Treffer liefert.

Nach Untersuchung der Sachlage wurde als neuer Ansatz zur Deckung des Bedarfs der Aufbau einer OER-Kollektion für Machine Learning und das Erstellen erster Materialien dafür beschlossen. Besondere Beachtung wurde bei der Planung auf die Nachnutzbarkeit der Materialien gelegt. Zusätzlich zu einem Zeit- und Ressourcenplan wurde dafür extra ein Maßnahmenplan angelegt, ähnlich einem Datenmanagementplan bei Forschungsprojekten. Ebenso wurde die Vorgehensweise zum Aufbau der Kollektion festgelegt und die Aufnahmekriterien für die Objekte definiert. Im weiteren Verlauf wurde bei der Umsetzung zur Erstellung der Materialien zunächst ein Prototyp entwickelt. Darauf basierend wurden dann die Materialien erstellt. In GitHub ist Git als Software zur Versionskontrolle integriert [59]. Das erlaubt ein iteratives Arbeiten. Die Materialien wurden Schritt für Schritt verbessert. Anderen Personen kann Zugriff auf private Git-Repositorien gegeben werden. Diese werden als Collaborators bezeichnet [62]. So konnte von seitens des Betreuers sowohl der aktuelle Stand als auch vorgenommene Änderungen jederzeit eingesehen werden. Des Weiteren wurden die in GitHub angebotenen Features zum Projektmanagement eingesetzt [58], um Workflows zur Kommunikation, zur Qualitätskontrolle und zum kollaborativen Arbeiten zu implementieren. Dabei wurde sich an der Vorgehensweise bei Open-Source-Projekten orientiert. Zu diesem Zweck wurde während der Implementierungsphase eine zweite Literaturrecherche nach Best-Practices durchgeführt. Im letzten Schritt wurden die OER-Materialien über eine Organisation in GitHub öffentlich zugänglich gemacht.

4.2 Erstellung der OER-Materialien

In der Kollektion werden drei Materialtypen bereitgestellt Bilder, Erklärtexte und Code. Im Folgenden wird die Erstellung dieser detailliert beschrieben.

4.2.1 Erstellung der Grafiken als SVG mit Inkscape

Die Grafiken wurden mit der Open-Source-Software Inkscape angefertigt [55]. Dabei stand der Gedanke der Nachnutzbarkeit für eine möglichst große Gruppe von Menschen im Vordergrund. Inkscape bietet die Option des Arbeitens mit Ebenen an. Diese können ein- und ausgeblendet werden. Das ermöglicht eine einfache Anpassung der Grafiken. Des Weiteren können diese Ebenen noch in Unterebenen gegliedert werden und lassen so noch mehr Raum für individuelle Gestaltung [63]. Es wurde sich folgendes Schema überlegt. Generell sollte es eine Trennung zwischen Grafik- und Textelementen geben. Das erleichtert einen sprachübergreifenden Austausch. Des Weiteren wurden die Metadaten wie Titel, Autor:in, Quelle und Lizenz mit Logo auf eine eigene Ebene gesetzt. So können die Nutzenden selbst entscheiden, ob sie die Quellenangaben im Bild haben möchten oder alternativ die Angaben in ein Bildverzeichnis setzen. Die Logos der Creative Commons Lizenzen lassen sich frei herunterladen [64]. Eine weitere Möglichkeit, Metadaten einzubinden, gestatten die

Dokumenteinstellungen in Inkscape. Es finden sich die Reiter Metadaten und Nutzungsbedingungen - Lizenz, unter denen ebenfalls die entsprechenden Informationen eingetragen wurden. Öffnet man den XML-Editor, so sieht man, dass die Daten durch XML strukturiert erfasst wurden und maschinenlesbar sind. Als letzten Punkt wurde auf das Layout eingegangen. Die Bildgröße wurde der Grafik angepasst. Außerdem wurde darauf geachtet, dass die Farbkombination aus rot und grün vermieden wurde aufgrund der Rot-Grün-Sehschwäche.

4.2.2 Erstellung des Beispielcodes mit Jupyter Notebooks

Die Codebeispiele wurden mit der Open-Source-Software Jupyter Notebooks erstellt. Das erlaubt die Umsetzung von Literate Programming. Bei diesem Ansatz befinden sich der Quellcode und die Dokumentation in einer Datei [65]. Die Verständlichkeit für den Menschen wird in den Vordergrund gestellt. Der Aufbau der Datei unterteilte sich in einen einleitenden Textteil und den Code. Der Text wurde in Markdown geschrieben. Der erste Teil bestand aus dem Titel, der Quellenangabe, den verwendeten Bibliotheken und einer Verlinkung auf die Datenquelle mit den Nutzungsbedingungen. Dem folgte eine stichpunktartige Zusammenfassung der einzelnen Bearbeitungsschritte zur Entwicklung des maschinellen Lernmodells.

Der Quellcode wurde in Python geschrieben. Um die Verständlichkeit zu verbessern, wurde der Code mit Kommentaren versehen. Zusätzlich wurde in den Kommentaren auf weitere Informationen verlinkt. Außerdem wurden immer wieder Plots eingebunden, um Zusammenhänge grafisch darzustellen. Diese wurden mit Titel, Achsenbeschriftung und gegebenenfalls einer Legende versehen.

Modelle im Machine Learning werden mit Daten trainiert. Welche Daten geeignet sind, hängt von der jeweiligen Fragestellung ab, für die ein Modell entwickelt wird. Wie schon angesprochen, spielen auch die Nutzungsbedingungen eine entscheidende Rolle, ob die Daten verwendet werden können oder nicht. Darüber hinaus darf für OER auch der didaktische Aspekt nicht vergessen werden. Darauf wird auch in dem Artikel »A lesson for teaching fundamental Machine Learning concepts and skills to molecular biologists« hingewiesen [20]. Interessante Datensätze, zu denen die Lernenden einen persönlichen Bezug herstellen können, steigern die Lernbereitschaft. Es gibt zwar mittlerweile einige Plattformen, über die nach Daten gesucht werden kann, wie das UC Irvine Machine Learning Repository [66], OpenML [67] oder Kaggle [68]. Allerdings ist die Situation ähnlich wie bei OER-Materialien für Machine Learning unbefriedigend. Letztendlich bleibt einem nach einer erfolglosen Suche nur das Erstellen eines synthetischen Datensatzes übrig. Die Bibliothek Scikit-learn bietet dafür Funktionen an [69].

4.2.3 Erstellung der Erklärtexte in Markdown

Die Erklärtexte zu den Grafiken wurden in Markdown geschrieben. Dazu wurde der Open-Source-Editor Atom verwendet [57]. Jeder Text startete mit einer Überschrift. Dem folgte ein kurzer Erklärtext, in den die jeweiligen Grafiken eingebettet wurden. Optional wurde auch Code als Codeblock [70] eingefügt mit Angabe zur Quelle. Dies entfiel bei den Bildern. Hier ist die Quellenangabe bereits enthalten. Für den Erklärtext an sich wurde sie ans Ende des Dokumentes gesetzt.

Gerade wenn verschiedene Elemente zusammengefügt werden, ist dieses scheinbare Detail von Bedeutung. Nur so können andere auf die Originalquellen zurückgreifen.

Des Weiteren machte sich hier der Vorteil von Markdown bemerkbar. Es ist einfach zu erlernen. Bereits nach einer kurzen Einarbeitungsphase kann ein optisch ansprechender Text erstellt werden.

4.3 Anlegen der Kollektion auf GitHub

Die Kollektion wird als Git-Repositorium über die Online-Plattform GitHub öffentlich zugänglich gemacht. Beim Aufbau wird auf Best-Practices für Open Source Projekte zurückgegriffen. Zusätzlich wird eine Organisation angelegt, welche mehr Optionen zur Zusammenarbeit anbietet als ein Git-Repositorium.

4.3.1 Aufbau des Git-Repositoriums

GitHub wird in der Regel für Open-Source-Projekte genutzt. Mittlerweile haben sich dafür Best-Practices etabliert, wie das Anlegen einer README-Datei. Auch wenn diese sich auf die Entwicklung von Software beziehen, so lassen sie sich ebenso dem Projektmanagement zuordnen. Denn es geht in erster Linie um die Organisation von Aufgaben in Teams, die Kommunikation untereinander und mit der Community oder die Dokumentation von Arbeitsschritten. Dies lässt sich ebenso für die OER-Kollektion nutzen.

Als erstes wurde in GitHub ein privates Repositorium angelegt. Das wurde geklont, indem eine Kopie auf dem lokalen Rechner abgespeichert wurde [71].

Dann wurden die benötigten Prozesse definiert zur:

- Informationsversorgung,
- Kommunikation,
- Dokumentation,
- Mitarbeit,
- Qualitätskontrolle.

Das Bereitstellen von Informationen erfolgte durch README-Dateien. Das ist eine bewährte Methode aus der Softwareentwicklung, um wichtige Informationen den Nutzenden im vorab mitzuteilen [72]. Mittlerweile werden solche Dateien auch als Einstieg für Softwareprojekte genutzt. Sie sollten daher alle nötigen Informationen enthalten, die jemand braucht, um entscheiden zu können, ob das Projekt für die eigene Arbeit relevant ist oder nicht. Dazu gehören eine Projektbeschreibung und Angaben zum Inhalt. Dem sollte eine Anleitung zur Mitarbeit und Lizenzangaben folgen [73]. An diesem Punkten wurde sich orientiert für die Erstellung einer allgemeinen README-Datei auf der Übersichtsseite des Repositoriums. Zusätzlich wurden Beispiele aufgezählt, in welchem Kontext die Materialien verwendet werden können. Ebenso gab es Hinweise und Links zu den Creative Commons Lizenzen sowie deren Anwendung. In der README-Datei findet sich eine Verlinkung auf die CONTRIBUTING-Datei, die eine Anleitung zur Mitarbeit enthält. Für weitere Ordner und deren Unterordner im Repositorium wurden zum Teil ebenfalls README-Dateien mit einer kurzen Beschreibung des Inhalts angelegt.

Die Ordnerstruktur greift die in Kapitel 3.5.2 ermittelten inhaltlichen Einheiten auf. Es wurde ein Ordner pro Algorithmus angelegt. Auf der nächsten Unterebene wurde die Unterteilung in ein Klassifikations- und Regressionsproblem durch weitere Ordner abgebildet. In diesen befinden sich dann auf der untersten Ebene drei Ordner entsprechend der drei Materialtypen Code, Bilder und die Erklärtexte. Der Bilderordner enthält in seiner README-Datei eine Gallery, um die Suche nach geeignetem Material zu erleichtern. Nutzende müssen so nicht erst jede SVG-Datei Einzel öffnen, um den Inhalt betrachten zu können.

Neben dem Bereitstellen von Informationen ging es auch um die Einführung eines Workflows zur

Kommunikation. Der Gedanke der Vernetzung ist ein zentrales Anliegen im OER. Darüber hinaus wurde in den fachspezifischen Empfehlungen herausgestellt, wie wichtig Feedback ist. GitHub bietet als Feature dafür Issues an [60]. In der Software-Entwicklung werden diese in der Regel für Fehlermeldungen wie zum Beispiel Bugs benutzt. Gerade im Open-Source-Bereich stellen sie aber auch eine Möglichkeit für Nutzende dar, mit den Entwickelnden in Kontakt zu treten und zum Beispiel neue Features vorzuschlagen oder um sich Hilfe zu holen. Issues können in GitHub mit Labels versehen werden, welche eine schnelle optische Einordnung erlauben [61]. Bestimmte Labels wie Bug oder help wanted sind bereits implementiert. Es können jedoch auch Eigene angelegt werden. Für die OER-Kollektion wurden drei Use Cases definiert: das Stellen einer Frage, das Einbringen einer neuen Idee und zuletzt das Geben von Feedback. Das Label question war bereits vorhanden. Für die Kollektion wurden idea und feedback neu erstellt. Um eine gewisse Einheitlichkeit bei den Issues zu gewährleisten, wurde ein Template angelegt [74]. Wird ein neues Issue erstellt, so wird in einem Text nach dem Anliegen gefragt. Zusammen mit den Labels soll damit eine Steuerung der Kommunikation bewerkstelligt werden. Ein weiterer Aspekt ist die Unterstützung der Dokumentation. Ein neu erstelltes Issue hat zunächst den Status open. Andere können darauf antworten und so kann daraus eine Diskussion entstehen. Wenn das Thema beendet ist, wird es geschlossen, bleibt allerdings weiter gespeichert. So entsteht eine Historie.

Die Dokumentation der Arbeitsschritte erfolgt in einem Git-Repositorium durch die Versionskontrolle. Jede Änderung an einer Datei wird durch Git erfasst, welche dann committed werden kann. Das heißt, dass eine neue Version gespeichert wird. Alle Versionen werden im Repositorium vorgehalten. Dadurch ist eine genaue Rückverfolgung möglich mit Angabe des Datums und der Person, die den Commit vorgenommen hat. Des Weiteren können die einzelnen Commits mit einer Commitnachricht versehen werden [75]. Die Anleitung für die sieben Regeln für großartige Git-Commitnachrichten [76], wurde hier als Vorlage verwendet. Allerdings wurde entschieden, sie zu vereinfachen, um es anwenderfreundlich zu halten. Für kleinere Änderungen reicht eine einzeilige Mitteilung aus. Sie sollte mit einem Großbuchstaben beginnen und im Imperativ geschrieben sein. Bei größeren Änderungen sollte die Mitteilung aus einem Titel mit einer kurzen Beschreibung bestehen, die beinhaltet, was getan wurde und warum. Prägnante Commitnachrichten sind wichtig, denn sie erleichtern beim Scrollen durch die Commit-Historie das Nachvollziehen der einzelnen Schritte, ohne das jeder Commit geöffnet werden muss.

Für die Mitarbeit an der Kollektion wurde eine CONTRIBUTING-Datei geschrieben. GitHub stellt hierfür in seiner Dokumentation eine Anleitung zum Erstellen einer solchen Datei mit verlinkten Beispielen zur Verfügung [77]. Zu Beginn wurde sich zuerst für das Interesse an einer Mitarbeit bedankt. Dem folgten Beschreibungen der Kommunikations-Workflows, Hinweis auf die Lizenz, eine Schritt-für-Schritt-Anleitung, wie mithilfe eines Forking-Workflows Änderungen eingereicht werden können und schließlich wie Commitnachrichten aussehen sollen.

Um dem Charakter einer Kollektion gerecht zu werden, wurde ferner ein Style-Guide hinzugefügt. Dieser liefert Informationen zu der Ausgestaltung der einzelnen Materialtypen Code, Bilder und Erklärtexte, wie in den Kapiteln 4.2.1 bis 4.2.3 beschrieben. Es wurden möglichst wenige Vorgaben gemacht, um viel Raum für die Einbringung eigener Ideen zu lassen. Zum Beispiel wurde keine Programmiersprache vorgegeben.

Der Forking-Workflow ist die Umsetzung des Qualitätssicherungsprozesse in GitHub. Wenn eine Person kein Collaborator in einem Repositorium ist, so hat sie dennoch die Möglichkeit, sich zu beteiligen. Dazu kann sie das Repositorium forken. Damit ist gemeint, dass eine Kopie des Repositoriums im eigenen User-Account erstellt wird. Diese kann dann auch wieder geklont werden, um so eine lokale Kopie auf dem eigenen Rechner zu haben. Im nächsten Schritt wird ein Themen-Branch

erstellt. Das ist eine Abzweigung vom Hauptentwicklungsstrang für ein spezielles Vorhaben. Ein Entwicklungsstrang in Git besteht aus einer Abfolge von Commits. Mithilfe von Branches kann es in einem Repositorium parallel zum Hauptentwicklungsstrang zusätzlich Nebenentwicklungsstränge geben. Nach Beendigung wird der Themen-Branch zurückgeführt in den Hauptentwicklungsstrang und es kann ein Pull Request gemacht werden. Dabei wird eine Nachricht an das ursprüngliche Repositorium versendet. Ähnlich wie bei einem Issue kann auf die Nachricht geantwortet und über die Änderungen gemeinsam diskutiert werden. Gegebenenfalls werden sie überarbeitet. Anschließend können die Änderungen übernommen werden. Das wird als mergen bezeichnet [78]. Auch dafür lässt sich ein Template anlegen [79]. Bei jedem Pull Request wird automatisch in dem sich öffnenden Text-Editor eine Checkliste angezeigt. Sie fragt ab, ob es bereits andere offene Pull Requests zu einem ähnlichen Thema gibt. Des Weiteren wird nachgefragt, ob die Änderungen den Vorgaben in der CONTRIBUTING-Datei entsprechen, wie zum Beispiel die Commitnachrichten. Außerdem wird darauf hingewiesen, einen Titel und eine kurze Beschreibung hinzuzufügen.

Auf der Übersichtsseite des Repositoriums sollte die Lizenz deutlich sichtbar sein. Das wurde zum einen bereits in der allgemeinen README-Datei berücksichtigt, allerdings bietet GitHub noch eine weitere Option an. Es werden mittlerweile eine Reihe von Lizenzen zur Verfügung gestellt [80]. Aus diesen kann eine ausgewählt werden, die automatisch dem Repositorium hinzugefügt wird. Die CC-BY 4.0 Lizenz ist aber zurzeit nicht verfügbar. Es lässt sich aber auch eine leere Vorlage auswählen, in die dann der Lizenzvertrag hinein kopiert wurde.

Alle Dokumentdateien wurden in Markdown und in englischer Sprache geschrieben. Um die Auffindbarkeit des Repositoriums auf GitHub zu erhöhen, wurde es mit Schlagwörtern versehen. Es wurden die folgenden ausgewählt: machine-learning, creative-commons, supervised-learning, teaching-materials.

4.3.2 Anlegen einer Organisation

Um die Pflege und Weiterentwicklung der OER-Kollektion langfristig zu ermöglichen, wurde beschlossen, sie in einem öffentlichen Repositorium über GitHub zu betreiben. Da nur schwer abzuschätzen ist, wie viele Personen sich zukünftig an der Mitarbeit beteiligen werden, wurde vorsorglich eine Organisation angelegt. Das ist einer von drei Kontotypen auf GitHub.

Es gibt den persönlichen Account. Damit können User Repositorien anlegen, in denen sie der Owner sind. Sie können andere User einladen und zum Collaborator machen. Allerdings können diese zwei Rollen für große Projekte nicht immer ausreichend sein. Daher gibt es als zweiten Kontotyp den der Organisation [62]. In ihr gibt es neben der Rolle des Owners noch eine Reihe weiterer Rollen wie zum Beispiel Member oder Billing Manager. Jede Rolle ist mit unterschiedlichen Rechten verbunden [81]. Um nicht alle Personen, die Teil einer Organisation sind, Einzel verwalten zu müssen, können Teams angelegt werden. Diesen können bestimmte Rechte wie Read oder Write auf einzelnen Repositorien in der Organisation gegeben werden [82]. Hat jemand zum Beispiel ein Read-Recht auf ein Repositorium, so kann diese Person daran mitarbeiten, indem sie das Repositorium forkt. Der Forking-Workflow wurde in Kapitel 4.3.1 beschrieben. Dadurch kann in einer Organisation ein interner Qualitätssicherungsprozess umgesetzt werden.

Dieser Kontotyp erlaubt eine hohe Flexibilität für zukünftige Formen der Zusammenarbeit sowohl intern in der ZB MED als auch darüber hinaus mit externen Personen. Nachdem die Organisation angelegt wurde, wurde sich der Name Machine-Learning-OER-Collection überlegt. Abgestimmt darauf wurde das erste Repositorium in ihr Machine-Learning-OER-Basics benannt. Zukünftig können weitere Repositorien angelegt werden mit einem bestimmten Schwerpunkt.

5 Fazit und Ausblick

Das Thema der Masterarbeit war die Entwicklung eines neuen Lösungsansatzes für den aktuellen Mangel an gut adaptierbaren OER-Materialen für ML unter Einbeziehung von Konzepten und Best-Practices vorwiegend aus dem Open Science Bereich.

Die bisherige Herangehensweise war darauf ausgerichtet, Kurse und Vorlesungen als OER-Material bereitzustellen. Einen Nachteil stellt hierbei das Herauslösen einzelner Elemente aus einer bestehenden Lerneinheit dar, welches oft durch technische und rechtliche Hürden erschwert wird. Zudem muss bei der Nachnutzung eines Kurses der Inhalt auf die jeweilige Zielgruppe angepasst werden. Deshalb wurden einzelne Elemente erstellt, die als Objekte in einer Kollektion zur Verfügung gestellt wurden.

Dabei wurde eine Reihe von Erfahrungen gemacht, die im Folgenden als Lessons learned zusammengefasst werden.

- Es sollte so früh wie möglich mit der Festlegung der Nutzungsbedingungen und daraus folgend der Auswahl einer Lizenz für die Materialien begonnen werden. Dabei kann sich an der Zielgruppe orientiert werden. Als Faustregel gilt, je weniger restriktiv die gewählte Lizenz ist, umso größer wird der Personenkreis, der sie nachnutzen kann.
- Die Lizenz sollte gut sichtbar platziert werden.
- Die Metadaten sollten mit dem jeweiligen Objekt verknüpft sein. Nur so kann gewährleistet werden, dass die Objekte und ihre Metadaten zusammenbleiben, auch wenn sie herausgenommen und verteilt werden.
- Die Metadaten sollten maschinenlesbar sein.
- Es sollten in der Planungsphase Maßnahmen zur Nachnutzbarkeit getroffen werden. Die FAIR-Prinzipien können hier zur Orientierung genommen werden.
- Dem schließt sich die Empfehlung an, die Materialien, wenn möglich, über ein Repositorium zu veröffentlichen, welches die FAIR-Prinzipien unterstützt. Als ein generisches Repositorium, in dem auch Bildungsmaterialien publiziert werden dürfen, bietet sich zurzeit Zenodo an. Allerdings sollte berücksichtigt werden, wer die Zielgruppe ist und wo diese sucht.
- In der Planungsphase sollte sich über fachspezifische Besonderheiten informiert werden.
- Das schließt auch die didaktische Vermittlung ein. Besonderheiten des jeweiligen Fachgebiets, wie zum Beispiel im Machine Learning das Arbeiten mit externen Datenquellen, müssen frühzeitig bedacht und eingeplant werden.
- Bei der Erstellung der Materialien sollte die technische Erweiterbarkeit einbezogen werden. Zum Beispiel lassen sich Texte in Markdown in verschiedenen Formaten ausgeben.
- Best-Practices, die Anwendung in Open Source Projekten finden, lassen sich übertragen. README-Dateien stellen einen einfachen Weg da, Informationen zu vermitteln. Eine CONTRIBUTING-Datei definiert die Kriterien zur Mitarbeit. Kombiniert mit einem Style-Guide werden die Anforderungen an einen einheitlichen Qualitätsstandard der Objekte in einer Kollektion abgedeckt.
- Mit der Auswahl von geeignete Tools und Diensten zur Bereitstellung der Materialien und zur kollaborativen Zusammenarbeit sollte frühzeitig begonnen werden. Idealerweise sollten sie Features zum Projektmanagement anbieten und den Aufbau einer Community unterstützen.

Der neue Lösungsansatz hat sich bei der Planung und Umsetzung zum einem von den Erfahrungen und den daraus gewonnen Erkenntnissen im Open Science Bereich inspirieren lassen, zum anderen erwies sich die Transformation des Sammlungsbegriffes in Folge der Digitalisierung als hilfreich. Das hat sicherlich auch damit zu tun, dass der Umgang mit digitalen Informationen Vorgehensweisen hervorbringt, die unabhängig vom jeweiligen Entstehungskontext sind. Die FAIR-Prinzipien [35], die sich generell auf digitale Objekte übertragen lassen, liefern hierfür ein gutes Beispiel. Dennoch muss angemerkt werden, dass die Einarbeitung zeitintensiv ist. In der Masterarbeit wurde es dadurch bemerkbar, dass es zu zwei Recherchephasen kam, obwohl ursprünglich nur eine einkalkuliert war. Deshalb wurde die Publikation der Kollektion auf Zenodo nicht wie geplant umgesetzt.

Grundsätzlich ist die Erstellung von OER-Materialien durch die Maßnahmen zur Nachnutzbarkeit mit einem höheren Aufwand verbunden. Diese Investition zahlt sich erst langfristig aus. Außerdem kommt bei einer Kollektion die Pflege und Weiterentwicklung hinzu, wie im Kapitel 3.4 Merkmale einer digitalen Sammlung erläutert wurde.

Um diesen Zeitaufwand zu reduzieren, gäbe es die Möglichkeit, mit dem integrierten Tool GitHub Actions Workflows zu automatisieren [83]. Des Weiteren erleichtert die Integration, die zwischen GitHub und Zenodo besteht, den Publikationsworkflow [84].

Als ein Punkt in den Leassons learned wurde die technische Erweiterbarkeit angegeben. Ein konkretes Beispiel liefert GitHub Pages [85], mit dem sich Webseiten erzeugen lassen. Die bereits erstellen Erklärtexte lassen sich als Markdown-Dateien in diese einbinden. Somit wird ein Mehraufwand vermieden.

Diese drei Beispiele zeigen, dass durch die Masterarbeit erst der Grundstein für die OER-Kollektion gelegt wurde. Dadurch das Maßnahmen zur Nachnutzbarkeit einbezogen wurden, bietet sie viel Flexibilität bei der Weiterentwicklung und der Umsetzung neuer Ideen an.

Literatur

[1] Verein openscienceASAP. *Open Science - Was ist Open Science?* 7. Apr. 2022. URL: http://openscienceasap.org/open-science/.

[2] Deutsche UNESCO-Kommission e. V. *Bildung - Open Educational Resources.* 7. Apr. 2022. URL: https://www.unesco.de/bildung/open-educational-resources.

[3] John Hilton III u. a. „The Four R's of Openness and ALMS Analysis: Frameworks for Open Educational Resources". In: *Open Learning* 25.1 (2010), S. 37–44.

[4] David Wiley. *The Access Compromise and the 5th R.* 7. Apr. 2022. URL: https://opencontent.org/blog/archives/3221.

[5] Jöran Muuß-Merholz. *Zur Definition von „Open" in „Open Educational Resources" – die 5 R-Freiheiten nach David Wiley auf Deutsch als die 5 V-Freiheiten.* 7. Apr. 2022. URL: https://open-educational-resources.de/5rs-auf-deutsch/.

[6] Paul Klimpel und Till Kreutzer. *Was ist CC?* 7. Apr. 2022. URL: https://de.creativecommons.net/was-ist-cc/.

[7] iRights e.V. *Creative Commons - Wie funktionieren freie Lizenzen?* 7. Apr. 2022. URL: https://irights.info/dossier/creative-commons.

[8] David Wiley. *Defining the Open in Open Content and Open Educational Resources.* 7. Apr. 2022. URL: https://opencontent.org/definition/.

[9] Deutsche UNESCO-Kommission e. V. *Agenda Bildung 2030 - Bildung und die Sustainable Development Goals.* 7. Apr. 2022. URL: https://www.unesco.de/bildung/agenda-bildung-2030/bildung-und-die-sdgs.

[10] UNESCO. *UNESCO-Empfehlung zu Open Educational Resources (OER).* 2019.

[11] Informationsstelle OER. *geförderte Projekte der OERinfo-Förderrichtlinie.* 7. Apr. 2022. URL: https://open-educational-resources.de/ueber-oerinfo/gefoerderte-projekte-der-oerinfo-foerderrichtlinie/.

[12] Wissenschaftsrat. *Zum Wandel in den Wissenschaften durch datenintensive Forschung - Positionspapier.* 2020.

[13] Tom Caswell u. a. „Open Content and Open Educational Resources: Enabling universal education". In: *International Review of Research in Open and Distributed Learning* 9.1 (2008). DOI: 10.19173/irrodl.v9i1.469.

[14] DFG. *Appell zur Nutzung offener Lizenzen in der Wissenschaft.* 9. Apr. 2022. URL: https://www.dfg.de/foerderung/info_wissenschaft/2014/info_wissenschaft_14_68/.

[15] Open Education Global. *History of OEG.* 7. Apr. 2022. URL: https://www.oeglobal.org/about-us/history-of-oeg/.

[16] Open Education Global. *What We Do.* 7. Apr. 2022. URL: https://www.oeglobal.org/about-us/what-we-do/.

[17] David Wiley. *What's the Difference Between OCWs and MOOCs? Managing Expectations.* 7. Apr. 2022. URL: https://opencontent.org/blog/archives/2909.

[18] Ludwig Bothmann u. a. „Developing Open Source Educational Resources for Machine Learning and Data Science". In: *CoRR* abs/2107.14330 (2021). arXiv: 2107.14330.

[19] ZB MED. *Forschung im Programmbereich Data Science and Services.* 7. Apr. 2022. URL: https://www.zbmed.de/forschen/forschung-bei-zb-med/forschung-data-science-and-services/.

[20] Rabea Müller u. a. „A lesson for teaching fundamental Machine Learning concepts and skills to molecular biologists“. In: *Proceedings of the Second Teaching Machine Learning and Artificial Intelligence Workshop*. Hrsg. von Katherine M. Kinnaird, Peter Steinbach und Oliver Guhr. Bd. 170. Proceedings of Machine Learning Research. PMLR, 2022, S. 68–72.

[21] EMBL-EBI Training. *Virtual course - Systems biology: From large datasets to biological insight*. 2021. URL: https://www.ebi.ac.uk/training/events/systems-biology-large-datasets-biological-insight/#vf-tabs__section--tab1.

[22] Konrad Foerstner u. a. *foerstner-lab/2021-06-21-Supervised_Machine_Learning_as_part_of_an_EBI_Systems_Biology_course: v0.1.0*. 2021. URL: https://zenodo.org/record/5218745#.YinDSonMK70.

[23] Konrad Foerstner u. a. *Supervised Machine Learning Methods - A short introduction*. 2021. URL: https://github.com/foerstner-lab/2021-06-21-Supervised_Machine_Learning_as_part_of_an_EBI_Systems_Biology_course.

[24] forschungsdaten.org. *Repositorium*. 7. Apr. 2022. URL: https://www.forschungsdaten.org/index.php/Repositorium.

[25] GitHub Docs. *GitHub Docs - About repositories*. 7. Apr. 2022. URL: https://docs.github.com/en/repositories/creating-and-managing-repositories/about-repositories.

[26] Informationsstelle OER. *Die OER.de-Karte*. 7. Apr. 2022. URL: https://open-educational-resources.de/karte/.

[27] Informationsstelle OER. *OERSI – Die Suche nach OER für die Hochschullehre*. 9. Apr. 2022. URL: https://open-educational-resources.de/ueber-oerinfo/impressum-2/.

[28] OER Commons. *OER Commons & Open Education*. 9. Apr. 2022. URL: https://www.oercommons.org/about.

[29] MERLOT. *How We Got Started*. 9. Apr. 2022. URL: https://info.merlot.org/merlothelp/topic.htm#t=How_We_Got_Started.htm.

[30] OER Commons. *Explore. Create. Collaborate*. 9. Apr. 2022. URL: https://www.oercommons.org/.

[31] MERLOT. *SmartSearch*. 9. Apr. 2022. URL: https://www.merlot.org/merlot/.

[32] iRights e. V. *Open Educational Resources - Offen für Kommerz? Bildungsmaterialien und das Problem nicht-kommerzieller Lizenzen*. 7. Apr. 2022. URL: https://irights.info/artikel/oer-creative-commons-noncommercial/28879.

[33] Andreas Degkwitz. „Digitale Sammlungen - Vision eines Neubeginns“. In: *Degkwitz, Andreas: Bibliothek. 38 2014* (2014).

[34] Thomas Stäcker. „Die Sammlung ist tot, es lebe die Sammlung! - Die Digitale Sammlung als Paradigma moderner Bibliotheksarbeit“. In: *BIBLIOTHEK – Forschung und Praxis* (2019). DOI: http://dx.doi.org/10.18452/19773.

[35] Mark D. Wilkinson u. a. „The FAIR Guiding Principles for scientific data management and stewardship“. In: *Scientific Data* 3.1 (2016), S. 160018. DOI: 10.1038/sdata.2016.18.

[36] Christof Schöch. „Aufbau von Datensammlungen“. In: *Digital Humanities: Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: J.B. Metzler, 2017, S. 223–233. DOI: 10.1007/978-3-476-05446-3_16.

[37] Andreas Christian Müller, Sarah Guido und Kristian Rother. *Einführung in Machine Learning mit Python: Praxiswissen Data Science*. 1. Auflage. Heidelberg: O'Reilly, 2017.

[38] TIB BLOG. *Die FAIR Data Prinzipien für Forschungsdaten*. 7. Apr. 2022. URL: https://blogs.tib.eu/wp/tib/2017/09/12/die-fair-data-prinzipien-fuer-forschungsdaten/.

[39] EU Horizon Europe. *Data Management Plan Template - Version 1.0 05 May 2021*. 7. Apr. 2022.

[40] GitHub Docs. *Referencing and citing content*. 8. Apr. 2022. URL: https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content.

[41] Zenodo. *Principles*. 8. Apr. 2022. URL: https://about.zenodo.org/principles/.

[42] DataCite. *Welcome to DataCite*. 8. Apr. 2022. URL: https://datacite.org/.

[43] Zenodo. *Terms of Use v1.2*. 8. Apr. 2022. URL: https://about.zenodo.org/terms/.

[44] Zenodo. *Zenodo - Startseite*. 8. Apr. 2022. URL: https://zenodo.org/.

[45] Zenodo. *Frequently Asked Questions*. 8. Apr. 2022. URL: https://help.zenodo.org/.

[46] GitHub. *GitHub - Choose the plan that's right for you*. 8. Apr. 2022. URL: https://github.com/pricing.

[47] Zenodo. *OAI-PMH*. 8. Apr. 2022. URL: https://developers.zenodo.org/#oai-pmh.

[48] Zenodo. *REST API*. 8. Apr. 2022. URL: https://developers.zenodo.org/#rest-api.

[49] GitHub. *Opting into or out of the GitHub Archive*. 8. Apr. 2022. URL: https://docs.github.com/en/enterprise-cloud@latest/get-started/privacy-on-github/opting-into-or-out-of-the-github-archive-program-for-your-public-repository.

[50] Creative Commons. *Mehr über die Lizenzen*. 8. Apr. 2022. URL: https://creativecommons.org/licenses/?lang=de.

[51] Creative Commons. *CC BY 4.0*. 8. Apr. 2022. URL: https://creativecommons.org/licenses/by/4.0/deed.de.

[52] Wissenschaftliche Dienste des Deutschen Bundestages. *Künstliche Intelligenz und Machine Learning - Eine urheberrechtliche Betrachtung*. 2018.

[53] Project Jupyter. *Jupyter*. 8. Apr. 2022. URL: https://jupyter.org/.

[54] Wikipedia. *Project Jupyter*. Hrsg. von Wikipedia – Die freie Enzyklopädie. 8. Apr. 2022. URL: https://de.wikipedia.org/wiki/Project_Jupyter.

[55] Inkscape. *Inkscape - Startseite*. 8. Apr. 2022. URL: https://inkscape.org/de/.

[56] Wikipedia. *Vektorgrafik*. Hrsg. von Wikipedia – Die freie Enzyklopädie. 8. Apr. 2022. URL: https://de.wikipedia.org/wiki/Vektorgrafik.

[57] Atom. *A hackable text editor for the 21st Century*. 8. Apr. 2022. URL: https://atom.io/.

[58] GitHub Docs. *Need help?* 8. Apr. 2022. URL: https://docs.github.com/en.

[59] Wikipedia. *GitHub*. Hrsg. von Wikipedia – Die freie Enzyklopädie. 8. Apr. 2022. URL: https://de.wikipedia.org/wiki/GitHub.

[60] GitHub Docs. *About issues*. 8. Apr. 2022. URL: https://docs.github.com/en/issues/tracking-your-work-with-issues/about-issues.

[61] GitHub Docs. *Managing labels*. 8. Apr. 2022. URL: https://docs.github.com/en/issues/using-labels-and-milestones-to-track-work/managing-labels.

[62] GitHub Docs. *Access permissions on GitHub*. 8. Apr. 2022. URL: https://docs.github.com/en/get-started/learning-about-github/access-permissions-on-github.

[63] Wikibooks-Bearbeiter. *Inkscape/ Ebenen*. Hrsg. von Die freie Bibliothek Wikibooks. 8. Apr. 2022. URL: https://de.wikibooks.org/wiki/Inkscape/_Ebenen.

[64] Creative Commons. *Downloads*. 8. Apr. 2022. URL: https://creativecommons.org/about/downloads/.

[65] Wikipedia. *Literate programming*. Hrsg. von Wikipedia – Die freie Enzyklopädie. 8. Apr. 2022. URL: https://de.wikipedia.org/wiki/Literate_programming.

[66] UC Irvine Machine Learning Repository. *Welcome to the UC Irvine Machine Learning Repository*. 8. Apr. 2022. URL: https://archive-beta.ics.uci.edu/.

[67] OpenML. *OpenML - A worldwide machine learning lab*. 8. Apr. 2022. URL: https://www.openml.org/.

[68] kaggle. *kaggle - Startseite*. 8. Apr. 2022. URL: https://www.kaggle.com/.

[69] scikit-learn developers. *7.3. Generated datasets*. 8. Apr. 2022. URL: https://scikit-learn.org/stable/datasets/sample_generators.html.

[70] Matt Cone Project. *Markdown Guide - Code*. 8. Apr. 2022. URL: https://www.markdownguide.org/basic-syntax/#code.

[71] Scott Chacon und Ben Straub. *Pro Git book - 2.1 Git Basics - Getting a Git Repository*. 8. Apr. 2022. URL: https://git-scm.com/book/en/v2/Git-Basics-Getting-a-Git-Repository.

[72] DateiWiki. *.readme Dateierweiterung*. 8. Apr. 2022. URL: https://datei.wiki/extension/readme.

[73] Danny Guo. *Make a README*. 8. Apr. 2022. URL: https://www.makeareadme.com/.

[74] GitHub Docs. *Configuring issue templates for your repository*. 8. Apr. 2022. URL: https://docs.github.com/es/communities/using-templates-to-encourage-useful-issues-and-pull-requests/configuring-issue-templates-for-your-repository.

[75] GitHub Guides. *Git Commit*. 8. Apr. 2022. URL: https://github.com/git-guides/git-commit.

[76] cbeams. *How to Write a Git Commit Message*. 8. Apr. 2022. URL: https://cbea.ms/git-commit/.

[77] GitHub Docs. *Setting guidelines for repository contributors*. 8. Apr. 2022. URL: https://docs.github.com/en/communities/setting-up-your-project-for-healthy-contributions/setting-guidelines-for-repository-contributors.

[78] Scott Chacon und Ben Straub. *Pro Git book - 6.2 GitHub - Mitwirken an einem Projekt*. 8. Apr. 2022. URL: https://git-scm.com/book/de/v2/GitHub-Mitwirken-an-einem-Projekt.

[79] GitHub Docs. *Creating a pull request template for your repository*. 8. Apr. 2022. URL: https://docs.github.com/en/enterprise-server@3.1/communities/using-templates-to-encourage-useful-issues-and-pull-requests/creating-a-pull-request-template-for-your-repository.

[80] GitHub Docs. *Licensing a repository*. 8. Apr. 2022. URL: https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/licensing-a-repository.

[81] GitHub Docs. *Roles in an organization*. 8. Apr. 2022. URL: https://docs.github.com/en/organizations/managing-peoples-access-to-your-organization-with-roles/roles-in-an-organization.

[82] Scott Chacon und Ben Straub. *Pro Git book - 6.4 GitHub - Verwalten einer Organisation*. 8. Apr. 2022. URL: https://git-scm.com/book/de/v2/GitHub-Verwalten-einer-Organisation.

[83] GitHub Docs. *GitHub Actions*. 8. Apr. 2022. URL: https://docs.github.com/en/actions.

[84] arshul und ivotron. *Github Actions for Zenodo*. 8. Apr. 2022. URL: https://github.com/ivotron/zenodo.

[85] GitHub Docs. *Creating a GitHub Pages site*. 8. Apr. 2022. URL: https://docs.github.com/en/pages/getting-started-with-github-pages/creating-a-github-pages-site.

Bildquellen

Das Logo der TH Köln ist von Marius Barzynski, Anna Fitz, Benedikt Schmitz und Andreas Wrede. Es ist unter einer CC-BY 4.0 Lizenz lizenziert und abrufbar unter https://de.wikipedia.org/wiki/Datei:TH-K%C3%B6ln-logo-03.png.

A Liste der OER-Verzeichnisse

The Virtual Linguistics Campus:

`https://oer-vlc.de/`

DuEPublico:

`https://duepublico.uni-due.de/`

TIB AV-Portal:

`https://av.tib.eu/`

HUU Mediathek:

`https://mediathek.hhu.de/`

Digitale Sammlungen der Bayrischen Staatsbibliothek:

`https://www.digitale-sammlungen.de/de/`

rpi-virtuell:

`https://rpi-virtuell.de/`

OpenLearnWare

`https://openlearnware.de/`

Freiburger Multimedia Object Repository

`https://freimore.uni-freiburg.de/`

OERSI

`https://oersi.de/resources/`

OER Commons

`https://www.oercommons.org/`

MERLOT

`https://www.merlot.org/merlot/`

B Abschnitt 2 aus der DMP-Vorlage (HE):V1.0

Auszug aus der 1. Version der Vorlage für Datenmanagementpläne vom 05.05.2021 des EU-Forschungsprogrammes Horizon Europe abrufbar unter dem Link Data Management Plan Template:

„2. FAIR data

2.1. Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your

discipline, please outline what type of metadata will be created and how.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be harvested and indexed?

2.2. Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized access protocol?

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

How will the identity of the person accessing the data be ascertained?

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

2.3. Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will

you follow community-endorsed interoperability best practices? Which ones?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)?

2.4. Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Will the provenance of the data be thoroughly documented using the appropriate standards?

Describe all relevant data quality assurance processes. Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.“