ARTIKEL

2.3. Mit Standards forschen und Handlungsräume schaffen

von Diekjobst, Anne; Geelhaar, Tim; Hodel, Tobias; Mähr, Moritz; Seltmann, Melanie;

Version 1.0 | Veröffentlicht 4. Juli 2024

1. Einleitung

Standards, Standardisierungen und standardisierte Workflows haben einen wesentlichen Einfluss auf Inhalt, Form und Interpretation historischen Materials. Das gilt bereits für das Material selbst, aber auch für herkömmliche Editionen und mehr noch für digitalisierte und digitale Objekte. Aus historischer Perspektive lässt sich standardisiertes Arbeiten sehr weit zurückverfolgen. Urkunden zum Beispiel folgen seit poströmischer Zeit einem stark formalisierten, also standardisierten Aufbau, was Jean Mabillon im 17. Jahrhundert nutzte, um eine Methode zur Identifizierung von Urkundenfälschungen zu entwickeln (Mabillon 1681) . Im gleichen Zug wurde eine Frühform der Textedition etabliert. Im 19. Jahrhundert wurden kritische Apparate zum notwendigen Bestandteil wissenschaftlicher Editionen. Heutzutage folgen die meisten digitalen Editionen dem XML-basierten Dokumentenformat TEI der Text Encoding Initiative.[1]

Standards, ihr Zustandekommen und ihren Einfluss zu kennen und sie selbst anzuwenden, ist für die geisteswissenschaftliche Erschließung historischen Materials unerlässlich. Denn Standards ermöglichen Verständnis, sie können es allerdings auch erschweren. Ein eindrückliches Beispiel ist die Edition der Hanserezesse, also der Ergebnismitschriften von Treffen der späteren Hanse, wie Angela Huang und Ulla Kypta herausgearbeitet haben (Huang und Kypta 2011). Ihr Editor Karl Koppmann hatte die Rezesse im 19. Jahrhundert der Zeit gemäß in ein juristisch-verfahrenstechnisches Korsett gezwängt. Er nahm an, dass es von Anfang an ein festgelegtes Prozedere aus Einladungsbriefen, Mitschriften und Abschlusserklärungen gegeben haben müsse. Indem er dieses Modell auf das historische Material übertrug, projizierte er den vermeintlichen Beginn der Hanse als Organisation gut 150 Jahre in die Vergangenheit. Die Hanseforschung brauchte selbst dann über 100 Jahre, um sich von dem so entstandenen Missverständnis wieder zu lösen. Aber auch andere Editionen des 19. und 20. Jahrhunderts haben historisches Material ebenso sehr erschlossen wie verzerrt. Standards wie moderne Orthographie, die Anpassung an das Buchformat als Publikationsform oder standardisierte Auslassungen entfremdetenden gedruckten Text immer weiter vom Ausgangsmaterial.

Der Erschließungsprozess wird durch die Digitalisierung noch erweitert, indem weitere Ebenen der Aufbereitung und damit der Standardisierung hinzukommen. Die computergestützte Arbeit erfordert nämlich die Umwandlung von Text in Daten nach formalisierten Datenmodellen, die maschinell prozessiert umgesetzt werden. Das Deutsche Textarchiv (DTA) macht diese zusätzlichen Auszeichnungsebenen sehr transparent, wenn es Texte im gut dokumentierten DTA-Basisformat aufbereitet.[2] Dieser Standard erlaubt es zusammen mit einem Cascaded Analysis Broker (CAB), jedes (historische) Eingabewort im transkribierten Text in eine eindeutige „kanonische“ Schreibweise zu transformieren und zu repräsentieren. Für die linguistische Analyse, die über eine reine Wortsuche hinausgeht, sind aber noch weitere Ebenen erforderlich, z. B. die Auszeichnung von Part-of-Speech (grob gesagt Wortarten), von Named Entities wie Personen, Orten und Sachen und vor allem Lemmatisierung. Diese standardisierten digitalen Auszeichnungsverfahren entfernen das analysierbare Material noch weiter vom Ausgangsmaterial. Bei automatisierten Auszeichnungen treten immer wieder Fehler auf, die schlimmstenfalls die Analyse und damit auch die spätere Interpretation negativ beeinflussen können.

Daher ist es für die digitale Quellenkritik von großer Bedeutung, Standards, Standardisierungen und ihre Rolle im Arbeitsprozess zu kennen, ihren Nutzen wie auch ihre Nachteile einschätzen und mit ihnen reflektiert umgehen zu können. Dieses Kapitel will dazu beitragen, diese Kompetenzen im Umgang mit digitalen Daten und Datenmodellen zu fördern und zu vertiefen. Es richtet sich an ein breites Publikum: an Studierende, Lehrende und Forschende, aber auch an Editor:innen und Entwickler:innen. Es nimmt die Perspektiven derer ein, die entweder medial aufbereitetes Material auswerten oder selbst Material aufbereiten. Allgemein wird das Kapitel die Sinnhaftigkeit von Standards ebenso wie die Notwendigkeit zur Reflexion über Standards im Arbeitsprozess aufzeigen. Konkret führt es vor, wie durch die Auswahl und den Einsatz von Standards Handlungsspielräume für den Umgang mit Daten geschaffen werden; mit anderen Worten, wie Auswertungs- und Interpretationsmöglichkeiten von der standardisierten Aufbereitung und Bereitstellung von Daten abhängen. Auf diese Weise will das Kapitel auch zwischen den Informations-, Computer- und Geisteswissenschaften vermitteln, um verschiedentliche Vorbehalte abzubauen. Denn das Stichwort Standardisierung löst bisweilen stark ablehnende Haltungen aus. Standardisierung ist jedoch notwendig und nützlich, sie führt nicht zwangsläufig zu einer inhaltlichen Reduktion des Quellenmaterials und

trägt ebenso wenig dazu bei, dass quantifizierende Methoden automatisch die hermeneutische Analyse ablösen.

Zunächst klärt das Kapitel die Fragen, was unter Standard zu verstehen ist, auf welchen verschiedenen Ebenen Standards zum Einsatz kommen, welche ethische und politischen Implikationen die Verwendung von Standards beinhaltet und welche normativen Konsequenzen sich daraus für die eigene Arbeit ergeben. Der zweite Abschnitt konzentriert sich auf digitale Standards in den Geisteswissenschaften. Dabei liegt der Schwerpunkt auf der Interdependenz von Standards, die im sog. Schichtenmodell präsentiert wird. Der dritte Abschnitt widmet sich dem Phasenmodell und damit dem Einsatz von Standards im geisteswissenschaftlichen Arbeitsprozess. Er gibt Antworten, zu welchem Zeitpunkt man sinnvollerweise über den Einsatz von Standards nachdenken sollte, welche Folgen deren Einsatz haben kann und wie man dadurch Plan- und Kalkulierbarkeit erhöht. Von zentraler Bedeutung ist hierbei die Balance zwischen vorgegebenen Standards und individuellen Lösungen/Anpassungen, die es für jedes Forschungsprojekt neu auszutarieren gilt (Hiltmann 2018). Auf diesem Weg soll das Kapitel helfen, bei der Konzeption und Durchführung von digitalen Projekten durch entscheidende Fragen zu tragfähigen Entscheidungen zu kommen. Der vierte Abschnitt thematisiert den kritischen Umgang mit Standards, indem es Faktoren fokussiert, die die Analyse beeinflussen und somit bei der Auswertung digital aufbereiteten Materials mitgedacht werden müssen, um zu verhindern, dass es zu ähnlichen interpretatorischen Fehlleistungen wie im Fall der Hanserezesse kommt. Das Schlusskapitel bündelt wesentliche Aussagen und bietet einen Ausblick auf die Möglichkeiten standardisiert erfasster Daten im *semantic web*. Inhaltlich schließt dieser Beitrag an `Kapitel 2.4` des Handbuchs an.

Was aber ausdrücklich nicht geleistet werden kann, ist eine systematische Vorstellung und Bewertung aller Standards und Verfahren. Hier wird nämlich bereits ein entscheidendes Charakteristikum von Standards sichtbar: Obwohl sie im Grundsatz vereinheitlichen sollen – man denke nur an die Normen des Deutschen Institut für Normung e.V. (DIN)[3] – gibt es immer wieder neue bzw. Anpassungen und Auslegungen bestehender Standards. Eine systematisierende Übersicht wäre also schon aufgrund der Vielzahl und Vielfalt von Standards nicht zielführend.[4].

2. Ein Standard kommt selten allein

Es scheint eine Ironie des Lebens zu sein, dass unser Alltag in fast allen Belangen reglementiert und standardisiert ist, dass es aber wiederum für Standards keine allgemeingültige Definition gibt. Gemeinhin steht Standard für Richtschnur, Norm, Maßstab oder Qualitätsniveau.[5] Einen guten Definitionsansatz bietet die deutschsprachige Wikipedia. Danach ist ein Standard eine „vereinheitlichte, weithin anerkannte und meist angewandte (…) Art und Weise, etwas zu beschreiben, herzustellen oder durchzuführen, die sich gegenüber anderen Arten und Weisen durchgesetzt hat oder zumindest als Richtschnur gilt."[6] Standards können demnach Regelwerke oder Verfahren zur Herstellung von Produkten, aber auch Maßstäbe zu deren Bewertung sein. Sie beziehen sich auf Verfahren oder auf die Gestalt und Qualität des Ergebnisses. Grundsätzlich sollen Standards Vereinheitlichung, Kohärenz und Konsistenz sicherstellen, weshalb sie unabhängig von Personen, Orten und Zeiten transparent formuliert und schriftlich fixiert sein sollten. Sie sollten möglichst selten verändert werden und dauerhaft gelten. Ihre wesentlichen Aufgaben bestehen darin, Kooperation und Interoperabilität zu ermöglichen, Koordinationskosten zu senken sowie Komplexität zu reduzieren. Daher spielen Standards insbesondere in der Wirtschaft, Produktion, Technik, Computertechnologie, Administration und im Datenmanagement eine herausragende Rolle.

Der Prozess der Festlegung von Standards wird als Standardisierung bezeichnet, der wiederum dokumentiert werden muss, um nachvollziehbar zu sein. In diesem Prozess werden nicht nur einzelne Standards festgelegt, sondern auch miteinander in Beziehung gesetzt. Denn Standards treten meist gemeinsam auf, indem sie aufeinander aufbauen, aufeinander verweisen oder sich ergänzen: TEI baut beispielsweise auf XML (Extensible Markup Language) auf, das mithilfe von XSLT (Extensible Stylesheet Language Transformation) in HTML (Hypertext Markup Language) für die Darstellung im Browser umgewandelt werden kann. Die Standardisierung setzt hier bei den kleinsten Einheiten an, z. B. der Auszeichnung von Markup durch spitze Klammern wie <head> und den eigentlichen Elementen, die mit diesen Klammern definiert werden, also "head" für Überschrift. Solche Festlegungen aggregieren dann zu sehr komplexen Computersprachen, die sich wiederum wegen ihrer Standardisierung ineinander überführen lassen.

In einem Arbeitsprozess treten Standards auf so gut wie allen Ebenen auf. Ein Beispiel aus der geisteswissenschaftlichen Arbeit: Literaturrecherche und -verwaltung funktionieren aufgrund bibliographischer Standards wie MARC 21 (MAchine-Readable Cataloging Version 21) , RDA (Resource Description and Access) und RDF (Resource Description Framework).[7] Moderne Literaturdatenbanken wie Citavi, Zotero oder Endnote nutzen diese für den Datentransfer, sie unterstützen aber auch wissenschaftliches Schreiben, indem sie helfen, Referenzen automatisiert nach festgelegten Zitationsrichtlinien[8] zu setzen und Literaturverzeichnisse zu erstellen. Beim Verfassen von Texten werden wiederum standardisierte Zeichen verwendet. Microsoft Word verwendet hierzu den eigenen Standard Westeuropäisch (Windows), was besonders dann zu beachten ist, wenn man nichtlateinische Alphabete oder bestimmte Sonderzeichen nutzen will, und vor allem, wenn man Dateien speichert. Um Konvertierungsprobleme zu vermeiden, hat sich das viel umfassendere UTF-8 im Internet durchgesetzt, das fast 98 % aller Webseiten nutzen.[9] UTF-8 ist selbst eigentlich nur ein Transformationsformat für

Unicode, daher auch der Name Unicode Transformation Format (UTF). Unicode wiederum basiert auf dem Zeichensatz Universal Coded Character Set (UCS), mit dem sich 145.000 Zeichen darstellen lassen und der vom Unicode-Konsortium entwickelt und gepflegt wird. [10] Um einen verständlichen Text zu produzieren, ist selbstverständlich die Grammatik und die Orthographie der verwendeten Sprache einzuhalten. Für das Abspeichern wiederum werden andere Standards verwendet wie das Portable Document Format (PDF), welches wiederum Verlage zuweilen für die Publikation bevorzugen, die dann, versehen mit einer ISBN, auf den Markt gebracht wird. Dieses Beispiel zeigt letztlich auch, dass im Grunde alle Geisteswissenschaftler:innen mit Standards arbeiten und dies auch schon vor dem Aufkommen digitaler Methoden und der Digitalisierung von Kulturgütern.

Trotz des Anspruchs auf Einheitlichkeit und dauerhafte Gültigkeit sind Standard sehr variabel, d. h. sie werden immer wieder abgeändert, variiert, erweitert, verworfen und ersetzt. Neue Standards konkurrieren mit bestehenden, alte werden an aktuelle Bedarfe angepasst oder ganz neue Standards für neuartige Bedarfe entwickelt. Das Arbeiten mit

Standards erfordert immer wieder deren Auslegung und Anpassung, was bis zum (bewussten) Regelverstoß führen kann, je nachdem, wofür Standards eingesetzt werden. Dies richtet sich nach dem Forschungsziel, den Untersuchungsgegenständen und -methoden, den Projektstrukturen und/oder nach den Anforderungen der Fördermittelgeber. [11] Wie weit man bei der Adaptation von Standards gehen kann, um spezifische Anforderungen oder Ansprüche zu realisieren, ohne dabei den Zweck des Standards zu beeinträchtigen, gehört zu den Kompetenzen, die man nur im Umgang mit Standards erlernen kann. Schlimmstenfalls kommt es zu Inkompatibilitäten, was Arbeitsprozesse behindern, verzögern oder sogar verunmöglichen kann.

Standards können sehr wohl aufgegeben, ersetzt oder durch fehlende Anwendung in Vergessenheit geraten. Obwohl Standards nicht aufhören zu existieren, kann es wegen der fehlenden Anwendung oder auch der bewussten Abwertung (*deprecation*) dazu kommen, dass ein Standard obsolet wird. Dies kommt besonders in der Informationswissenschaft vor. Es hängt davon ab, welche Gemeinschaft welchen Nutzen aus einem Standard zieht und ihn daher aufrechterhält. Ein Beispiel ist hier die *International Standard Book Number (*ISBN), ein Standard nach DIN-Norm bzw. ISO 2018. Sie war eine Wegmarke im modernen Buchhandel und der bibliographischen Arbeit. Als fest etablierter Standard wird sie weiter benutzt, obwohl sie mittlerweile an die *European Article Number* (EAN) für den Vertrieb gekoppelt worden ist. Da mittlerweile der (Online-)Handel weit mehr als nur Bücher vertreibt, hat sich der *Document Object Identifier* (DOI) etabliert, mit dem sich verschiedenste Texte z. B. Kapitel oder Zeitschriftenaufsätze in unterschiedlichen Medien und Formaten – gerade auch Online-Publikationen – identifizieren, katalogisieren und verkaufen lassen.

Standards haben ihre eigene Geschichte und gewinnen mit der Zeit ein Eigenleben (Yates und Murphy 2019). Dies hängt sowohl von der eigenen Beharrungskraft eines etablierten Standards ab, als auch von den Communities und Infrastrukturen, die diese Standards tragen, wie das Schaubild illustriert. Die Art des Standards und sein Nutzen hängen maßgeblich von den Gemeinschaften ab, die ihn trägt.

Neben gemeinschaftlich festgelegten Standards gibt es allerdings auch Quasi-Standards, die sich in der Praxis durchsetzen oder aber durch eine dominante Position vertreten werden.

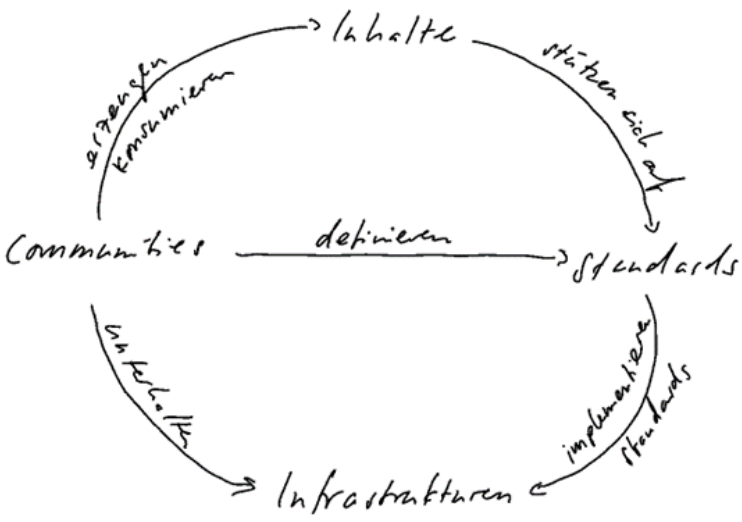

Abb 1: Visualisierung des Zusammenspiels von Standards mit Inhalt, Infrastruktur und Communities. Eigene Abbildung. Lizenziert unter einer CC-BY 4.0 Lizenz.

Dies kennt man von Microsoft, das mit seinen Dateiformaten sehr lange Zeit ein Quasi-Monopol beansprucht hat. Das Beispiel verweist zugleich auf inhärente Machtstrukturen, die sich insbesondere im kommerziellen Anwendungsfeld von Standards zeigen. In der Regel sind es aber nichtstaatliche Communities, die Standards festlegen. Am bekanntesten ist sicherlich die International Organization for Standardization (ISO), die 1946 gegründet wurde und als Verein nach schweizerischem Recht operiert. Zusammen mit der International Electrotechnical Commission (IEC) und der International Telecommunication Union (ITU) gründete sie 2001 die World Standards Cooperation (WSC), die sich für die Stärkung von konsensuell festgelegten Standards einsetzt. Die ISO führt derzeit 24.597 Standards, darunter technische wie den MP3-Standard, klassifikatorische wie Länderkürzel (de, at, ch) und verfahrensbezogene fürs Qualitätsmanagement (ISO 9000). Einen Eindruck von der Verteilung dieser Standards bietet eine Visualisierung, mit der die ISO angibt, mit welchen Standards die ISO die 17 Ziele für eine nachhaltige globale Entwicklung fördert. Mit 13.611 Standards überwiegt eindeutig das Ziel Nr. 9 Industrie, Innovation und Infrastruktur. [12]

Abb. 2: Visualisierung der ISO mit Bezug zu den globalen Entwicklungszielen. Online: https://www.iso.org/sdg/SDG09.html (zugegriffen: 15. April 2024). Ohne Angabe einer Lizenz.

Für Webtechnologien ist das World Wide WebConsortium (W3C) eine zentrale Instanz, die am 1. Oktober 1994 am MIT Laboratory for Computer Science in Cambridge (Massachusetts) gegründet worden ist. Das W3C beschreibt seine Vorgehensweise wie folgt: “W3C publishes documents that define Web technologies. These documents follow a process designed to promote consensus, fairness, public accountability, and quality. At the end of this process, W3C publishes Recommendations, which are considered Web standards.” [13] Zu den W3C Standards gehören unter anderem das Hypertext Transfer Protocol (HTTP) und der Uniform Resource Identifier (URI), aber auch Webstandards wie HTML, CSS, SVG, Ajax, Semantic Web (z. B. SPARQL) und XML-Technologien. Neben solchen umfassenden Gremien gibt es solche, die sich Teilgebieten widmen, wie das bereits erwähnte TEI-Konsortium. Internationale Vereinigungen setzen sich in der Regel aus nationalen Gremien zusammen, so ist das DIN Mitglied in der ISO. Auf nationale Ebene werden aber auch neue Vereinigungen gegründet wie z. B. die Nationale Forschungsdaten Infrastruktur, NFDI, [14] mit der “Datenbestände von Wissenschaft und Forschung für das gesamte deutsche Wissenschaftssystem systematisch erschlossen, vernetzt und nachhaltig sowie qualitativ nutzbar gemacht” werden, wie es auf der Webseite des eingetragenen Vereins heißt. [15]

Das Streben nach Standardisierung bringt Institutionalisierung und damit auch Machtstrukturen hervor. Denn diese Communities und Institutionen erfüllen wichtige Regulierungs– und Wächterfunktionen, sie können Akteure, Ideen und Verfahren ebenso anerkennen wie ausschließen. Jedoch sind Standards darauf angewiesen, dass eine möglichst große Gruppe von Menschen sie verwendet und damit ihren Status als Standard stützt. Es ist daher in den Augen der Autor:innen dieses Beitrages im höchsten Maße wünschenswert, partizipative Verfahren und Absprachen unter Fachkolleg:innen bei der Entwicklung von Standards einzusetzen, um möglichen Machtmissbrauch oder Selbstbegrenzung z. B. durch die Verwendung proprietärer Standards zu vermeiden. Dies gilt umso mehr, um mögliche rechtliche Komplikationen zu vermeiden, die sich gerade bei Data bzw. Text Mining Verfahren ergeben können. [16] Aus diesem Grunde ist der Quasi-Standard des Publizierens in Open-Access für freie Forschung so wichtig. Der politische Aspekt der Standardisierung lässt sich auch so formulieren, dass Zutrittsbarrieren selbst transparent und überprüfbar bleiben müssen. Gerade der Einsatz von Bilderkennungssoftware und von Algorithmen zur standardisierten Personenidentifikation kann schwerwiegende Konsequenzen nach sich ziehen, so hat eine voreingenommene (im Sinne von *bias*), KI-gestützte Gesichtserkennung

Afroamerikaner:innen diskriminiert (Simonite 2019; Buolamwini und Gebru 2018). Aus dieser politischen Dimension leiten sich normative, ethische Anforderungen an Standards ab. Die wesentlichen Grundsätze können als Meta-Standards verstanden werden, weil sie vielmehr Richtlinien als konkrete Handlungsanweisungen sind. Die DFG hat zuletzt am 20. April 2022 eine aktualisierte Fassung ihres Kodex „Leitlinien zur Sicherung guter wissenschaftlicher Praxis“ veröffentlicht, die insgesamt 19 Leitlinien für die wissenschaftliche Arbeit und den Umgang mit Daten enthält.[17] In der Leitlinie 13 “Herstellung von öffentlichem Zugang zu Forschungsergebnissen” verweist die DFG auf die FAIR-Prinzipien als Vorgaben, die bei der Arbeit mit Daten und deren Veröffentlichung einzuhalten sind. Diese hat das Wilkinson Laboratory am Centro de Biotecnología y Genómica de Plantas Madrid in einem Aufsatz von 2016 eingeführt (Wilkinson u. a. 2016). Das Akronym steht für “Findability, Accessibility, Interoperability, and Reusability”, also Auffindbarkeit, Verfügbarkeit, Interoperabilität und Nachnutzungsmöglichkeit von Daten, den dazugehörigen Metadaten, aber auch von Algorithmen, Tools und Workflows, die zu diesen Daten geführt haben, wie das Autorenkollektiv betont. Zweck dieser Prinzipien ist es, “to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing.” (Wilkinson u. a. 2016)

Um die “FAIRness” ihrer Daten sicherzustellen, sollten sich sowohl Forschende als auch Data Stewards an fünfzehn konkrete Regeln halten, die aus den Prinzipien heraus abgeleitet worden sind. Die Regel F1 sieht z. B. vor, dass die Daten mit persistenten Identifiern ausgestattet werden sollen, also z. B. DOI für Texte und ORCID-Angaben für Autor:innen.[18] Nach F4 “(Meta)data are registered or indexed in a searchable resource.” sollen (Meta)Daten an einem Ort gespeichert werden, an dem diese indexiert und auffindbar sind (Wilkinson u. a. 2016). Wilkinson und seine Mitautor:innen schlagen hierfür die verschiedene Datenhubs wie das Open-Access-Repository Zenodo vor,[19] in welchen Datasets automatisch eine DOI zugewiesen bekommen, mit einer ORCID verbunden sein sollen und bei einer Internetsuche aufzufinden sind. Alle Regeln sind knapp und übersichtlich in dem Hypotheses Blogbeitrag “The road to fair” von Karla Avanço aufgeführt (Avanço 2021).[20] Im Kern tragen die FAIR-Prinzipien zu einer Steigerung des Mehrwerts einer zeitgemäßen, wissenschaftlichen, digitalen Veröffentlichung bei.

Je nach Herkunft der Daten bzw. des zu datifizierenden Materials sollten auch die sog. CARE-Prinzipien eingehalten werden. Diese wurden 2020 in dem Aufsatz “The CARE Principles for Indigenous Data Governance” (Carroll u. a. 2020), erschienen im Data Science Journal, erstmals ausführlich vorgestellt und begründet. Die Autor:innen kritisieren die FAIR-Prinzipien dahingehend, dass diese rein datenorientiert formuliert seien und Fragen nach Herkunft, Kontrolle, Verantwortung sowie des kollektiven Nutzens außen vor lassen. Dabei komme es bei der Datafizierung von Wissensbeständen zu einem Ungleichgewicht zwischen indigenen Rechten und Interessen einerseits und Institutionen, die koloniales Erbe verwalten, andererseits. Denn bei der Datafizierung bestehe die Gefahr, dass indigene Interessengruppen erneut übergangen würden. Es würden Prozesse der kulturellen Enteignung infolge des Kolonialismus im Bereich der Data Science fortgeschrieben, wenn den ursprünglichen Rechteinhaber:innen nicht Mitsprache und Mitwirkung gewährt werde. Dies betrifft in erster Linie die sog. GLAM-Einrichtungen, also Galerien, Bibliotheken, Archive und Museen, die solche Objekte verwahren und digitalisieren, die im Kolonialismus geraubt worden sind. Um nun die indigenen Gruppen, ihre Rechte und Interessen ernst zu nehmen, sollten die CARE-Prinzipien Anwendung finden.

In Anlehnung an FAIR sind damit vier Prinzipien gemeint, die sich in jeweils drei Regeln konkretisieren lassen.[21] Der kollektive Nutzen “*collective benefit”* sieht vor: “Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data”. Konkret heißt dies, dass Regierungen und Organisationen die Nachnutzung von Daten durch indigene Völker erleichtern sollen (C1) und dass die (Nach-)Nutzung den Austausch zwischen Bürger:innen, Regierungen und Institutionen fördern solle (C2), wie auch zum Wohle und zur Förderung indigener Gruppen zu erfolgen habe (C3). Die Kontrollmöglichkeit (*authority to control*) soll wiederum den betroffenen Gruppen größere Mitsprache bei der Verwaltung dieser Daten einräumen (A3), um so einen verantwortlichen Umgang mit den Daten sicherzustellen (*responsibility*), was sich vor allem in die Pflicht übersetzen lässt, die Würde der indigenen Gruppen zu achten (R1) und diesen die Möglichkeit zu geben, ihre eigene *Data Literacy* im Umgang mit ihrem kulturellen Erbe zu erweitern (R2). An vierter Stelle geht es allgemein um ethische Aspekte (*ethics*): Möglicher Schaden soll verhindert oder zumindest minimiert bzw. der Nutzen für indigene Völker gefördert werden (E1). Ungleichheiten in Machtstrukturen, Ressourcen und deren Verfügung zum Schaden indigener Gruppen sollen vermieden und der Nutzen für die Zukunft sichergestellt werden (E2 und E3).

Während die FAIR-Prinzipien bereits allgemeine Anerkennung gefunden haben und eine gewisse autoritative Kraft entwickelt haben, stehen die CARE-Prinzipien trotz bereits lang andauernder Diskussionen und Umsetzungen, z. B. seitens des Smithsonian-Museums, noch vor ihrer Durchsetzung. So sinnvoll wie ethisch nachvollziehbar diese Prinzipien auch sind, scheint aber noch nicht klar zu sein, wie sich diese Richtlinien in Vorgaben übersetzen lassen, die durch den Computer verarbeitet werden können, was für Standards in den DH eine herausragende Rolle spielt.

3. Digitale Standards in den Geisteswissenschaften

Der vorangegangene Abschnitt hat gezeigt, dass wir auf verschiedensten Ebenen bereits mit Standards arbeiten, die noch nicht einmal immer digital und spezifisch geisteswissenschaftlich sein müssen. Nun gilt es, auf digitale Standards in den Geisteswissenschaften zu fokussieren. Wie auch bei den allgemeinen Standards sind hier technische Rahmenbedingungen, disziplinäre Eigenentwicklungen, Interdependenzen und Herausforderungen an die Kompatibilität zu berücksichtigen. Die Einsatzfelder von Standards reichen von grundlegenden technischen Speicherformen (z. B. Dateiformaten wie “txt”, “csv”, oder “xml”) bis zu hochgradig abstrahierten Konzepten zur Beschreibung von inhaltlichen oder materiellen Phänomenen (Linked Open Vocabularies). Nur wenige Standards sind dabei bisher von den Geisteswissenschaften selbst entwickelt worden. Sie gehen auf Anforderungen und Bedürfnisse zurück, die schon in der prädigitalen Zeit in den jeweiligen Disziplinen entwickelt wurden, worauf wir als erstes eingehen werden. Die Präsentation eines Schichtenmodells für Standards wird anschließend veranschaulichen, wie Standards sowohl vertikal als auch horizontal miteinander zusammenhängen oder sogar voneinander abhängen. Diese Übersicht soll vor allem dazu anregen, schon bei der Projektkonzeption den Einsatz von Standards einzuplanen. Dies wird nämlich umso wichtiger, wenn wir uns der Rolle von Standards in der computergestützten Datenverarbeitung und -analyse zuwenden. In diesem Zusammenhang vertreten wir im letzten Teil die folgenden zwei Thesen: 1) Um maschinelle Lernverfahren effizient und kritisch zu nutzen, sind Standardisierungen auf einige Schichten zwingend notwendig. 2) Nur über die Anwendung von Standards können Verknüpfungen erstellt werden, die tatsächlich weiterführend sind und Gleiches mit Gleichem verbinden, wie das insbesondere fürs *semantic web* propagiert wird.

3.1 Fachliche Standards

Zunächst einmal ist es bemerkenswert, dass seit der Etablierung des Begriffs “Digital Humanities” vor rund zwanzig Jahren nur wenige fachspezifische Standards in den Geisteswissenschaften (weiter-)entwickelt wurden. Zu nennen wäre das an die TEI angelehnte Auszeichnungsformat für Urkunden, die sog. Charters Encoding Initiative mit ihrem Standard CEI,[22] das die Formalia des Urkundenaufbaus nachbildet. Aus dem Bereich der Editionswissenschaft ist auf das bereits erwähnte DTA-Basisformat (DTABf) zu verweisen und aus dem Museumswesen kommt das CIDOC Conceptual Reference Model (CRM)[23] als Ontologie zur standardisierten Beschreibung von Artefakten des kulturellen Erbes. Hinzu kommen noch die so genannten Normdaten, die seit 2012 von der Deutschen Nationalbibliothek in der Gemeinsamen Normdatei (GND)[24] geführt werden. Es handelt sich um die Zusammenlegung bereits existierenden Verzeichnisse wie der Personennamendatei (PND), der Gemeinsame Körperschaftsdatei (GKD), der Schlagwortnormdatei (SWD) und der Einheitssachtitel-Datei des Deutschen Musikarchivs (DMA-EST). Die GND wird von der GND-Kooperative getragen, einem kooperativen “Zusammenschluss von Organisationen und Institutionen der Kultur und Wissenschaft der Bundesrepublik Deutschland, Österreichs und der Schweiz (DACH-Raum) mit der Zielsetzung, den Einsatz einheitlicher Standards für die Erschließung, Schnittstellen und Formate in Bibliotheken sicher zu stellen und die spartenübergreifende Harmonisierung der Erschließung und Datenvernetzung zu fördern.”[25] Technisch beruht die GND wiederum auf dem bibliothekswissenschaftlichen Standard MARC 21.

Diese Standards rekurrieren auf Praktiken des Katalogisierens, Edierens und des Beschreibens von Sammlungen, die in den vergangenen Jahrzehnten und Jahrhunderten in den Geisteswissenschaften entwickelt, eingeübt und perfektioniert worden sind. Man kann auch davon sprechen, dass unterschiedliche Disziplinen mit ihrer jeweiligen Fachsprache und ihren epistemologischen Grundannahmen eigene Standards besitzen, die dann durch Formalisierung in digitale Standards überführt werden, d.h. dass sie mithilfe eines Codesystem eindeutig beschrieben und referenziert werden können. Die neuen Standards führen daher analog existierende Praktiken fort, um den disziplinenspezifischen Anforderungen entsprechend nutzbringend zu sein. Die Beherrschung von disziplinär gebundenen Standards ist mittlerweile Ausweis der eigenen Qualifikation und der Zugehörigkeit zu einer wissenschaftlichen Community. Im Bereich der Bibliothek und auch in der Bibliothekswissenschaft zeigt sich dies etwa an einem vertieften Verständnis des bibliographischen Standards MARC 21. Hieran wird auch sichtbar, dass die Arbeit mit Standards keineswegs eine Simplifizierung der fachlichen Komplexität darstellt, weil der Standard selbst die Komplexität der Datenerhebung widerspiegelt und daher auch ein vertieftes Anwenderverständnis erfordert.

Gleichzeitig kann der disziplinäre Fokus die Anschlussfähigkeit an andere, ebenfalls disziplinenspezifische, Standards erschweren oder gar unterbinden. Daher ist zu überlegen, inwiefern Standards disziplinenspezifisch eingesetzt werden sollen, wenn dies mit Kompatibilitätsproblemen insbesondere bei interdisziplinärer Forschung einhergeht.

Da bei geisteswissenschaftlichen Forschungsprozessen die Aufbereitung sowohl von Primärdaten (also den eigentlichen Gegenständen wie Bild, Text oder Objekt) als auch von Metadaten (also Daten über die Primärdaten z. B. beteiligten Personen) relevant ist, müssen alle Datengenerierungs- und Auswertungsprozesse digital gedacht werden. Dementsprechend können wir Metadaten in einem breiten und nicht ausschließlich informatischen Sinn verstehen, sondern bewusst auch aus fachwissenschaftlicher,[26] bibliographischer[27] und informationstechnologischer Perspektive.[28] Zudem wollen wir projektinterne bzw. spezifische Formen der Datenaufbereitung in Betracht ziehen und Normdaten, genauso wie Linked-Open-Data-Vokabulare oder umfangreiche Ontologien sowie Schnittstellendefinitionen als

Form der Standardisierung begreifen.

Über die Aushandlungsprozesse in der Findung von Standards hinaus besteht die Möglichkeit, Übergänge von einem Standard in einen anderen zu bauen. «MARC 21 Feld 100»[29] lässt sich etwa auf «Dublin Core Creator»[30] *matchen* (im Sinne von "gleichsetzen") und somit lässt sich eine Konkordanz der zwei unterschiedlichen Formen der Informationsbeschreibung erreichen. Solche *crosswalks* sind für zukünftige Auswertungen zentral, da sich gerade im GLAM-Bereich unterschiedliche Standards etabliert haben, von denen in absehbarer Zukunft sicherlich nicht abgerückt wird.[31] Mittels solcher Crosswalks lassen sich aus unterschiedlichen Quellen und in diversen Kontexten erarbeitete Daten effizient verbinden. Gerade Linked Data Anwendungen unterstützen die Vorgehensweise und ermöglichen die Weiterverarbeitung von Datenstämmen unterschiedlicher Herkunft ohne dass man Kopien ziehen oder eine eigene Infrastruktur aufbauen muss.

Selbstredend muss ein Bewusstsein über den damit verbundenen Verlust an Informationen und an Genauigkeit vorhanden sein. In MARC 21 muss ein «Record», also eine bibliothekarisch zu erfassende Entität, vorliegen, während sich Dublin Core auf jegliche Objekte in der Welt beziehen kann. Es wird jedoch diverse Fragerichtungen und -perspektiven geben, die auf diese feine Unterscheidung verzichten können und nach einer "Urheberschaft" fragen. Dank der Kombination der unterschiedlichen Datenformen wird sich somit ein aussagekräftiges Bild ergeben. Informationsreduktion muss nicht in jedem Fall die Aussagekraft einer Auswertung reduzieren.

3.2 Schichtenmodell: Interdependenzen von Standards

Sehr technisch gesprochen bildet jeder Computer und jeder informatische Standard eine Kombination von geschichteten Standards, die von der Hardware über die Maschinensprache zu Programmiersprachen und Graphical User Interfaces gehen und diese verbinden. Wir können also Standards und die Koppelung von Standards aus den unterschiedlichen Schichten verstehen. Gleichzeitig können wir unterschiedliche Standards als leitend für einzelne oder mehrere Schichten denken, die horizontal und vertikal angedockt werden können. Die Notwendigkeit zur Nutzung unterschiedlicher Standards wird somit offensichtlich. Text- und Bilddaten sind an unterschiedlichen Stellen und mit verschiedenen Aussagezielen von Standardisierung betroffen. Einerseits kann die Beschreibung als Objekt (im Sinne von inhaltlichen Metadaten) auf Standards beruhen, andererseits kann auch die Art und Weise der Zurverfügungstellung oder Adressierung standardisiert erfolgen (etwa durch das International Image Interoperability Format, kurz IIIF oder die Vergabe von Document Object Identifier, DOIs). Schließlich ist auch die Übertragung über Hypertexttypeprotokolle (HTTP) nur Dank Standardisierung möglich. Wie die Schichtung verstanden werden muss, kann anhand eines Beispiels zur Organisation und Verarbeitung von Textdaten demonstriert werden. Es gibt unterschiedliche horizontale und vertikale Schichten, die auf unterschiedliche Standards aufbauen können, sich teilweise jedoch bedingen bzw. voneinander abhängig sind. Zu den horizontalen Schichten gehört erstens die **Textcodierung**. Bei dieser zentralen und omnipräsenten horizontalen Schicht für die Standardisierung von Textdaten steht der genutzte Zeichensatz im Zentrum, der als Standard verstanden wird. Ob Unicode oder in die Jahre gekommene ASCII-Codierungen genutzt werden, erweitert oder beschränkt die Nutzung von Zeichensätzen. Alle diese Standards müssen dennoch als Einschränkungen verstanden werden, etwa wenn Zeichenformen beschrieben werden sollen, die keine Codepunkte sind. Die Medieval Unicode Font Initiative (MUFI) versucht dem beispielsweise für mediävistische Zeichenrepräsentationen Abhilfe zu verschaffen.[32]

Eine zweite horizontale Schicht, zentral für Textdaten, ist das genutzte **Dateiformat**. Im "Rich Text Format" (RTF) gespeicherte Daten unterscheiden sich von Textdateien (TXT) und ebenso von im eXtensible Markup Language (XML) Format verfassten Dateien. Obwohl es sich bei allen Dateien um offene Standards handelt, bedeutet die Entscheidung für ein Format eine Einschränkung der visuellen und inhaltlichen Aussagekraft. TXT-Dateien speichern keine Textformatierungen wie z. B. Fettdruck, was das RTF-Format in begrenztem Umfang tut. Beide Formate verbrauchen daher aber auch weniger Speicherplatz und sind leicht auslesbar und somit schneller zu verarbeiten. XML-Dateien können alle möglichen weiteren Informationen enthalten, wodurch diese Dateien größer werden und langsamer verarbeitet werden. Außerdem müssen XML-Dateien erstens wohlgeformt sein, d. h. sie müssen den grundlegenden syntaktischen Regeln und Strukturrichtlinien für XML entsprechen, und sie müssen valide sein, d.h. sie müssen den semantischen Regeln und Strukturvorgaben des jeweils verwendeten XML-Schemas entsprechen.

Die vertikalen Schichten in Textdaten sind z. B. **lexikalische**, **syntaktische** und **semantische** Schichten, die sich auf denselben Text beziehen. Die genutzten Standards können sich indes stark unterscheiden. Während die lexikalische Schicht ein Inside-Outside-Beginning (IOB) *tagging* nahelegt, damit mit automatischen Systemen gearbeitet wird, kann für die Semantik auf RDF-basierende Ontologien als Standard abgestellt werden. Noch komplexer wird die Angelegenheit, wenn zusätzlich morphologische Phänomene und damit Wortteile analysiert werden sollen, da dadurch die Grundeinheit des Wortes aufgebrochen werden muss, was wiederum typischerweise andere Standards verlangt, damit sinnvoll gearbeitet werden kann.

Auch wenn gewisse Ansätze wie die Text Encoding Initiative versuchen, eine Vielzahl der Standards in einem einzigen zu vereinen, so

müssen wir anerkennen, dass immer mehrere Standards ineinandergreifen.

Um Interoperabilität zu erreichen, ist es notwendig, kompatible Standards zu bemühen und bei der Nutzung von Standards ihre Stellung im beschriebenen Schichtenmodell zu verorten. Schon bei der Wahl von Dateiformaten muss daher geklärt werden, welchen Standards gefolgt werden soll. Wichtig ist, dass nicht nur von einem Standard aus gedacht wird, sondern die gesamte Datenumgebung und -infrastruktur als Teil von unterschiedlichen Standardisierungsprozessen miteinbezogen werden muss.

Besonders der Aspekt der Interoperabilität macht die Standardisierung zu einem vordringlichen Problem. Denn die Festlegung eines neuen Standards, z. B. für die Datenauszeichnung, kann die Datenweiterverarbeitung durch andere Programme entweder deutlich vereinfachen oder aber erschweren. Im Text Mining und der Korpuslinguistik ist es üblich, dass die Primärdaten getrennt von den Metadaten im plaintext, also txt, als Input weiterverarbeitet werden. Es kann bereits zu Schwierigkeiten kommen, wenn die Primärdaten als Ausgangsformat in TEI vorliegen. Noch komplizierter wird es, wenn in TEI noch projektspezifische Sonderzeichen definiert werden, um eine besondere bildgetreue Repräsentation des Ausgangstextes zu erreichen. Natural Language Processing Verfahren, z. B. mittels der Software *SpaCy*,können dann nicht verwendet werden, weil Wörter mit solchen spezifischen Sonderzeichen nicht mehr auf ihr Lemma zurückgeführt werden können. Dadurch kann es zu unbrauchbaren Lemmatisierungsergebnissen und Wortstatistiken kommen. Das herausragende Editionsvorhaben “Schule von Salamanca” am Max-Planck-Institut für Rechtsgeschichte hat größte Anstrengungen unternommen, das Schriftbild der zu edierende Werke buchstabengetreu abzubilden und in XML auszuzeichnen.[33] Die Ergebnisse sind aber nicht unmittelbar mit Standard NLP-Tools zur Lemmatisierung und wortstatischen Auswertung kompatibel, was einen erhöhten Aufwand darstellt, um mittels Transformationsskripten weitere Ebene der Texterschließung zu ermöglichen und auf Methoden des *natural language processing* zurückzugreifen.

3.3 Forschen mit Standards

Nicht nur in der Anwendung, sondern auch in der Entwicklung von Forschungsmethoden, ist der Rückgriff auf Standards sinnvoll, bisweilen gar zwingend: Um maschinelle Lernverfahren effizient und kritisch zu nutzen, sind denn auch Standardisierungen auf einigen Schichten zwingend notwendig.

Das Argument hier ist entsprechend stark methodengetrieben und plädiert für die Nutzung geisteswissenschaftlich breit abgestützter Standards. Nur so kann die aktuelle und zukünftige Nutzung maschineller Lernverfahren garantiert werden. Insbesondere *machine learning*, also Verfahren, die supervisierte oder unsupervisierte Methoden der Datenaufbereitung umfassen, sind auf große Massen von Daten und gleichzeitig qualitativ hochwertig plus uniform erschlossene Daten als Trainings-, Validierungs- und Testdaten angewiesen. Maschinelle Lernverfahren brauchen je nach Anwendung große bis unglaublich große Datenmengen.[34] Daher ist es für einzelne Personen oder Forschungsgruppen nur selten möglich, selbst und von Grund auf Datenkorpora zusammenzustellen, die umfangreich genug sind.

Als Beispiel können wir uns an maschinellen Lernverfahren orientieren, die zum Training von Taggern zur Erkennung von Named Entities (Eigennamen) eingesetzt werden. Nur wenn ein standardisiertes Verständnis besteht, was wie ausgezeichnet wird, lassen sich unterschiedliche Datenbestände zu umfangreichen Datenstämmen kombinieren, die für Trainingsprozesse genutzt werden können. Das bedeutet jedoch auch, dass Fragen nach der Art und Weise des Verständnisses von Named Entities uniform beantwortet werden.

Nur dank der Orientierung an Standards ist es darauffolgend möglich, Daten nachzunutzen und in die angesprochenen Verfahren als Trainingsdaten einzuspeisen. Der Austausch von Daten zur Erweiterung von Datensätzen und zum Training umfangreicher, aber dennoch genauer oder zumindest standardkonformer Modelle, ist demnach das Gebot der Stunde, um neue technologische Ansätze reflektiert zu nutzen.[35] Um zukünftig den Einsatz maschineller Lernformen in den Geisteswissenschaften zu erweitern, ist die Nutzung von passgenauen Standards von zentraler Bedeutung. Aber auch um unterschiedliche Forschungsansätze zu verbinden, braucht es Standards.

Die Nutzung gemeinsamer Standards schafft typischerweise die Möglichkeit, Probleme und insbesondere auch Phänomene über den eigenen Fachbereich hinaus zu diskutieren und zu thematisieren. So können unterschiedliche Blickwinkel auf Materialien ausgedrückt werden, ohne dass auf Komplexität verzichtet werden muss. Die Standardisierung wird demnach zum Ort der Aushandlung.

Nur über die Anwendung und Aushandlung von Standards können Verknüpfungen zwischen standardisierten Systemen erstellt werden, die tatsächlich weiterführend sind und Gleiches mit Gleichem verbinden, wie es *semantic web* und somit *linked open data* Verfahren propagieren.

In dieser Vernetzung ist auch schon die Standardisierung hin zu Normdaten angelegt, die selbst Resultat von

Standardisierungsprozessen sind. Nur auf diesem Weg kann stabil und langfristig auf Konzepte, Personen, Geographika oder Werktitel verweisen werden.

Diese Überlegungen zur Standardisierung bilden die Grundlagen, um überhaupt in einem nächsten Arbeitsschritt die praktischen Umsetzungen zu diskutieren und Fragen der Bildung von Communities anzudenken.

4. Standards im Arbeitsprozess

Das zweite Modell, das wir uns für das Verständnis von Standards ansehen müssen, ist das Phasenmodell. Projekte und insbesondere das agile Projektmanagement (Kuster u. a. 2022; Preußig 2020) verlangen, dass die Ein- und Beschränkung auf Standards nicht von Beginn an als strikt gegeben verstanden wird, sondern im Rahmen eines Projekts verfeinert und ausdifferenziert werden kann. Das Phasenmodell fragt denn auch, wann welche Standards festgelegt, und vor allem, wann welche Standardisierungsentscheidungen abschließend getroffen werden müssen. Standards tragen in vielfältiger Weise dazu bei, die Zusammenarbeit der Forschenden zu vereinfachen, zu systematisieren und die Resultate ihrer Arbeit einem breiten Nutzerkreis zugänglich zu machen. Standardisierte Forschungsdaten können wesentlich einfacher nachgenutzt werden, was ihre Sichtbarkeit und Reichweite potenziell erhöht. So werden der Austausch und die Nachnutzung von Forschungsdaten und -ergebnissen erleichtert und die Transparenz der Forschung insgesamt erhöht. Standardisierte Forschungsdaten schaffen jedoch auch Zwänge und suggerieren, dass es *die eine* richtige Vorgehensweise gibt. Standards sind nicht neutral, sie haben wissens- und wissenschaftspolitische Macht und die Reflexion von Standards ist an verschiedenen Punkten der Projektarbeit unerlässlich (vgl. Kapitel 1). Im Folgenden werden daher ihre Potenziale und die damit verbundenen Fallstricke anhand der einzelnen Phasen vorgestellt.

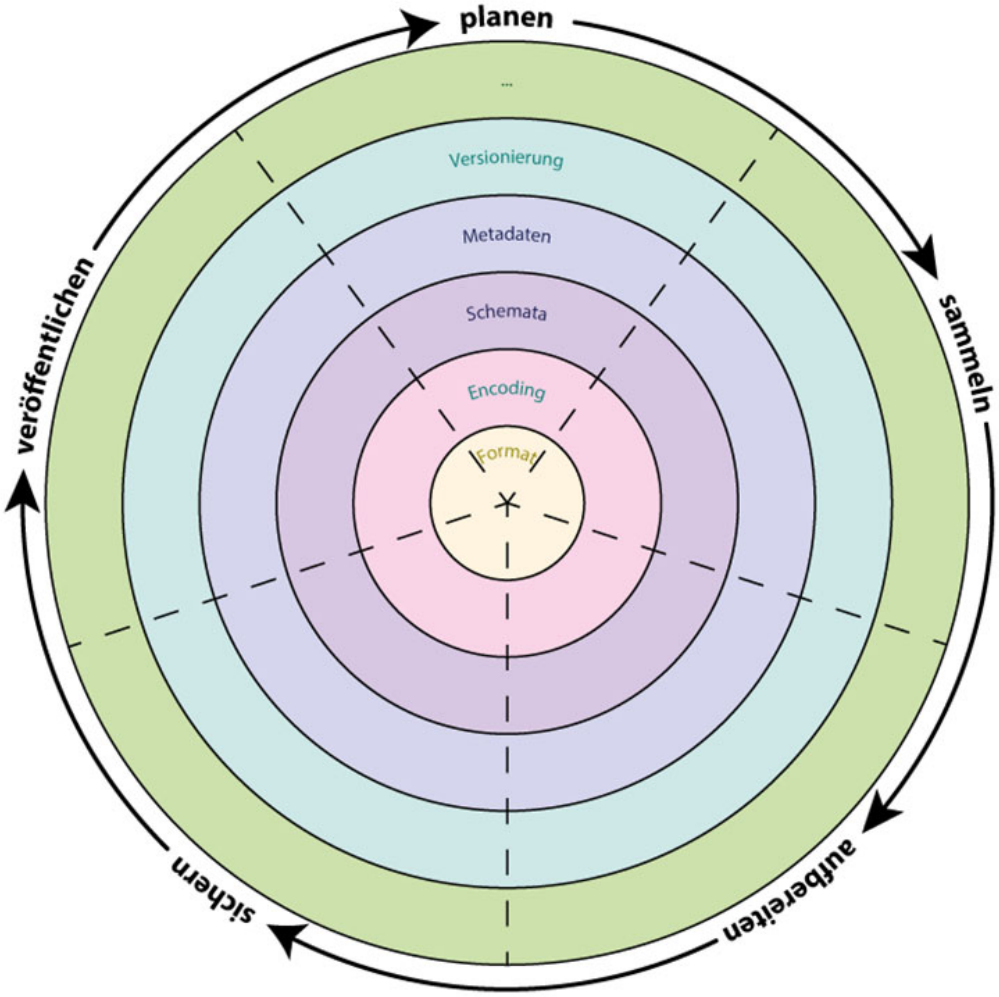

Abb. 3: Data Life Cycle mit Phasen des Datenlebenszyklus als Kreissektor sowie verschiedenen Ebenen von Standards als farblich unterschiedlichen Kreisringen. Eigene Abbildung. Lizenziert unter einer CC-BY 4.0 Lizenz.

Eine typische Form des Phasenmodells entspricht dem Datenlebenszyklus, der in Abb. 3 durch die Kreissektoren visualisiert wird.

Während aller Phasen des Datenlebenszyklus müssen Standards immer auf verschiedenen, hier farblich hervorgehobenen, Ebenen betrachtet werden. Standards existieren auf den Ebenen Format, Encoding, Schemata, Metadaten sowie Versionierung und ggfs. noch weiteren Ebenen.

4.1 Planung: Weichen stellen und Struktur schaffen

Bereits vor Beginn eines Forschungsprojektes ist darüber nachzudenken, welche Standards in der eigenen Disziplin auf sämtlichen relevanten Ebenen existieren und welche davon für das konkrete Projekt verwendet werden sollen bzw. für welche Fälle es keine oder nicht ausreichende Standards gibt. Diese Überlegungen sind Teil dessen, was in einen Datenmanagementplan (kurz DMP)[36] fließt, der von immer mehr Fördergebern gefordert ist. Aber auch, wenn kein solcher DMP gefordert wird, hat es viele Vorteile, einen solchen zu erstellen. Zum einen unterstützt er die Projektstrukturierung und die sinnvolle Strukturierung und Publikation der genutzten Daten, zum anderen hilft er während der weiteren Projektphasen immer wieder, Erwartungen und Ziele abzugleichen. Durch den DMP wird folglich nicht nur potenziellen Geldgebern Rechenschaft darüber abgelegt, wie mit Daten umgegangen wird, sondern auch sich selbst.

Sichtung, Auswahl und Anwendung passender Standards kosten zweifelsohne einige Zeit. Die Kosten für die Aufbereitung von Daten sowie für die Nutzung vorhandener Infrastrukturen können für Projekte jedoch teilweise mitbeantragt werden, sofern die Daten für die Publikation aufbereitet werden.[37] Bei der Wahl eines geeigneten Standards steht man oft vor dem Problem, sich zwischen (zu) vielen Optionen entscheiden zu müssen. Entscheidet man sich bspw. für einen alten, aber weit verbreiteten Standard oder für einen neuen, passgenaueren Standard? Überwiegen die Vorteile der großen Verbreitung und der hohen Nutzerakzeptanz des alten Standards seine Nachteile? Oder werden beim Einsatz eines alten Standards, der die technischen Anforderungen der aktuellen Computertechnologie nicht erfüllen kann, Potenziale verschenkt? Eine Möglichkeit der Bewältigung dieser Frage ist zu überprüfen, welcher aktuelle Standard die Anforderungen des Forschungsprojekts am besten abdeckt. Dies kann z. B. dadurch geschehen, dass man ähnliche Projekte sucht und daraufhin untersucht, welche Standards angewendet wurden. Außerdem sind die Konsultation von Handreichungen oder auch die Rücksprache mit den verschiedenen Vertretern der NFDI-Konsortien ein Weg sich zu informieren. Anschließend sollte die Frage gestellt werden, inwiefern Anpassungen des eigenen Workflows vorgenommen werden müssen. Sind diese Anpassungen nicht sehr aufwändig, sollten sie vorgenommen werden. Bei aufwändigen Anpassungen muss zwischen Aufwand und Nutzen abgewogen werden. Wenn dennoch mehrere Standards unterstützt werden sollen, gibt es die Möglichkeit, Mappings zu anderen Standards[38] oder direkt verschiedene Austauschformate zur Verfügung zu stellen.

Zu den wesentlichen Vorzügen von Standards und standardisierten Verfahren gehört die Fähigkeit, Komplexität bewältigen oder sogar reduzieren zu können. Sie tun das, indem sie gemeinsame Protokolle etablieren, Verfahren vereinheitlichen und Vorgaben für die Präsentation der Daten machen. Die Fähigkeit der Komplexitätsbewältigung erkaufen sich die Standards, indem sie Freiheit und Flexibilität der einzelnen Forschenden in einem gewissen Maße einschränken. Die bewältigbare Komplexität und die Freiheit stehen in einem inversen Zusammenhang.

4.2 Sammeln: Zusammenarbeit vereinfachen und vereinheitlichen

Bei der Sammlung von Forschungsdaten sollten die ausgewählten bzw. vereinbarten Standards direkt eingesetzt werden. Es mag zwar so scheinen, dass die Sammlung damit aufwändiger ist und mehr Expertise voraussetzt, jedoch werden so spätere Anpassungen reduziert und die Verfälschung der Daten durch etwa mangelhafte Dokumentation bei der Sammlung wird vermieden. Werden Daten oder aber die entsprechende Dokumentation verändert, wird eine Versionierung unerlässlich. Über Projektabsprachen und der Orientierung an klar definierten Vorgehensweisen muss sichergestellt werden, dass alle Beteiligten über geltende Standards informiert sind und diese einheitlich umsetzen, was für die Qualitätssicherung wie auch die Arbeitseffizienz bedeutsam ist (Artstein 2017). Bei Annotationsvorgängen wird dies beispielsweise durch Inter Annotator Agreements sichergestellt.[39] Dadurch ist es unerheblich, ob gleichzeitig oder zeitlich versetzt gearbeitet wird oder ob sich die Beteiligten untereinander kennen oder nicht. Standards erlauben es, dass unterschiedliche Projekte Infrastrukturen gemeinsam nutzen und Daten nachnutzen können. Solche Synergieeffekte helfen, Kosten zu senken. Es reduziert auf längere Sicht die Einarbeitungszeit, weil Mitarbeitende die Programme und Standards bereits kennen und für die Infrastrukturen und (Meta-)Daten bereits Dokumentationen vorliegen, die bei der Einarbeitung helfen.[40] Außerdem werden Arbeitsprozesse durch gemeinsame Standards entlastet und vereinfacht, weil zentrale Entscheidungen abgenommen und nicht wieder neu getroffen werden müssen. Im Idealfall können durch die Verwendung offener Standards andere Forschende zu einem späteren Zeitpunkt die Daten mit einem überschaubaren Aufwand für ihre eigenen Forschungsziele weiterverwenden.[41] Standards, die eine aktive Community haben, erhöhen nicht nur die aktive Weiterentwicklung des Standards, sondern auch durch gegenseitige Verweise die Sichtbarkeit der einzelnen Projekte. Hierbei ist zu beachten, dass die Standards einer Community auch dann noch in gewissem Maße weitergetragen werden müssen, wenn sie sich verändern. So sollte nicht für jedes “Problem” eine neue Lösung gesucht werden, sondern zuerst untersucht werden, inwiefern bereits vorhandene Regeln verwendet und angepasst werden können.[42]

4.3 Aufbereiten: Beschreiben, Analysieren, Interpretieren

Was für die Sammlung von Forschungsdaten gilt, gilt für deren Aufbereitung noch viel mehr. Je mehr Menschen sich einen Forschungsgegenstand oder ein Erkenntnisinteresse teilen, desto wichtiger wird eine geteilte Sprache, um gemeinsames Handeln und Verstehen sicherzustellen. Dabei verständigen wir uns darauf, wie wir uns auf die Welt und ihre Zeugnisse beziehen wollen. Wir verwenden eindeutige Bezeichnungen, zählen oder beschreiben in klar definierten Beschreibungssprachen. Solche Sprachen setzen sich aus festgelegten Einheiten zusammen, die wiederum regelbasiert aufeinander bezogen werden, und umfassen alle Ebenen und sowohl die beschriebene Entität als auch die Entität, mit der beschrieben wird. Daher ist es beispielsweise wichtig, Normdaten und standardisierte Vokabulare zu verwenden. Beachtet werden sollte bei der Wahl der Sprache jedoch, dass diese weitreichende Konsequenzen hat. Die Expressivität, also die Frage, was alles ausgedrückt und beschrieben werden kann, hängt genauso davon ab wie Fragen der Nachnutzbarkeit. Wenn Vokabulare und Normdateien proprietär sind, binden sie ein Projekt an gewisse Software oder technische Infrastrukturen. Darum werden besondere Ansprüche in Hinblick auf das Prinzip der Interoperabilität an sie gestellt. Dank Vokabularen und vor allem Normdaten ist eine Anschlussfähigkeit an Daten gegeben, die im Rahmen anderer Projekte erarbeitet wurden.

4.4 Nachnutzung: Publizieren & Langzeitsichern

Um schließlich eine Nachnutzung von Forschungsdaten zu ermöglichen, müssen auch bei der Publikation und Archivierung Standards eingehalten werden. Hierfür sollten insbesondere die FAIR-Prinzipien herangezogen werden, wie im ersten Unterkapitel dargelegt. Sind in den vorherigen Phasen konsequent Standards verwendet worden, bietet dies die beste Voraussetzung für die Nachnutzung von Forschungsdaten. Denn damit Forschungsdaten nachgenutzt werden können, müssen sie auffindbar sein. Obwohl die erste Anforderung der FAIR-Prinzipien (*findable*) wie eine Selbstverständlichkeit klingt, ist ihre Umsetzung alles andere als einfach. Sowohl die Forschungsdaten als auch ihre Metadaten müssen mit einer global eindeutigen und dauerhaften Kennung (z. B. DOI) versehen werden. Die Daten müssen genau beschrieben werden. Das umfasst nicht nur jedes (Meta-)Datenfeld, sondern auch die Genese der Daten. Nur so können die Daten von Menschen nachvollzogen und von Maschinen automatisch indexiert und durchsuchbar gemacht werden. (Wilkinson u. a. 2016) Die zweite Anforderung der FAIR-Prinzipien (*accessible*) schreibt vor, dass Daten und Metadaten anhand ihrer eindeutigen Kennung unter Verwendung von standardisierten Kommunikationsprotokollen abgerufen werden können. Dieses Protokoll soll offen, kostenlos und universell implementierbar sein sowie bei Bedarf Authentifizierungs- sowie Autorisierungsverfahren ermöglichen. Zudem soll auf die Metadaten zugegriffen werden können, selbst wenn die Daten (aus juristischen Gründen) nicht oder nicht mehr verfügbar sind.

Standardisierte Schnittstellen und Protokolle ermöglichen den Austausch von Informationen zwischen menschlichen und nicht-menschlichen Akteuren (drittes FAIR-Prinzip *interoperable)*. Während bei der Zusammenarbeit von Menschen eine gewisse Unschärfe und Variabilität der Daten zwar lästig, aber verkraftbar ist, versagen Maschinen in so einem Fall den Dienst. Standards sind also unabdingbar, wenn Menschen mit Maschinen oder Maschinen mit Maschinen kommunizieren sollen, wenn Daten in eine Datenbank eingegeben oder Daten aus einer Datenbank in eine andere verschoben werden sollen. Umgekehrt erleichtern Standards die automatische Anreicherung von Daten. Wer eine digitale Literaturverwaltung verwendet und dabei bibliographische Daten aus dem Netz mittels ISBN, EAN, Dublin Core oder DOI abruft, greift auf Standards zurück, um standardisierte, normierte Daten aus anderen Datenbanken in die eigene zu überführen.

Forschungsdaten und Metadaten sollen so gesichert werden, dass sie über einen möglichst langen Zeitraum von möglichst vielen Forschenden nachgenutzt werden können (viertes FAIR-Prinzip *reusable*). Das setzt zum einen voraus, dass das Archiv bzw. das Repositorium eine langfristige Aufbewahrung sowohl technisch (Integrität) als auch organisatorisch (Finanzierung) sicherstellen kann. Zum anderen müssen die Daten in möglichst voraussetzungsarmen Formaten und gut dokumentierten Standards aufbewahrt werden. Dies erlaubt es, dass die Daten auch von anderen Communities genutzt werden können.

5. Kritischer Umgang mit Standards

In den vorangegangenen Unterkapiteln ist bereits wiederholt angeklungen, dass der Einsatz von Standards das Arbeiten nicht nur erleichtert, sondern auch einschränken oder erschweren kann. Da sich hieraus viele Vorbehalte gegen ein Arbeiten mit Standards speisen und es sehr wohl Schattenseiten gibt, wollen wir noch einige ausgewählte Aspekte diskutieren, um so auch zum kritischen, selbstreflexiven Umgang mit Standards anzuregen. Im Folgenden werden wir allgemeine Beobachtungen zum sogenannten “Gatekeeping” und der Rolle von Autoritäten, sodann zu Asynchronizität im Projektverlauf, zur Revidier- und Erweiterbarkeit von Standards und zum Ausbruch aus Standards anstellen.

Mit “Gatekeeping” wird die bedeutende Funktion der Zugangskontrolle bezeichnet. Übertragen auf Standards kann man von Gatekeeping

sprechen, wenn Einzelne, Forschungsverbünde, Organisationen oder private Unternehmen darüber wachen, wer welchen Standard zu welchen Bedingungen wie lange nutzen darf. Im Falle privater Unternehmen wird die Nutzbarkeit durch die Geschäftsbedingungen festgelegt, denen man sich zu unterwerfen hat, wenn man deren Dienste nutzen will. Was zunächst trivial klingt, wird dann jedoch zu einer Herausforderung, wenn sich die Nutzungsbedingungen während der Projektlaufzeit ändern, indem ein Unternehmen von einmaligen Kosten für ein Softwarelizenz auf ein jährliche Abonnementsystem umstellt. Möglichkeiten der Mitsprache bei der Weiterentwicklung oder gar beim Ende eines Services sind für die Nutzenden in der Regel sehr eingeschränkt. Um sich vor möglichen Kostenfallen zu schützen und Abhängigkeiten vorzubeugen, werden daher in der Forschung Open-Source-Lösungen bevorzugt. Weitere wichtige Gatekeeper sind die im ersten Unterkapitel genannten Organisationen wie die International Organization for Standardization oder das W3-Konsortium. Indem sie Standards entwickeln (lassen), anerkennen bzw. nicht zulassen und vor allem deren Konsistenz bewahren, nehmen sie für sich eine hohe Autorität in Anspruch. Dies ist zweifelsohne für den Erhalt von Standards notwendig, erschwert aber Mitsprache und Gestaltungsfreiheit. In der Forschung kommt es ebenfalls zum Gatekeeping und zum impliziten wie expliziten Machtgefälle. Wir haben bereits auf die Verankerung von Standards in Forschungsgemeinschaften hingewiesen. Wissenschaftliche Gemeinschaften tragen zu der Genese, der Adaption, dem Erhalt und der Nutzbarkeit von Standards bei, indem sie eine Forschungspraxis etablieren. Welche Forschung wie betrieben wird, hängt von einer Vielzahl unterschiedlicher Faktoren ab, von der Infrastruktur einer Forschungseinrichtung bis zum Budget eines Projektes. Organisationen und Communities wie z. B. das TEI-Konsortium, die Standards entwickeln und pflegen, beeinflussen die Forschung durch ihre Standardsetzung auch in methodischer Hinsicht. Die Wahl eines oder mehrerer Standards verläuft entsprechend nicht nur an inhaltlichen und methodischen Linien, sondern entscheidet sich auch an (forschungs-)politischen Überlegungen, weil die Anwendung anderer technischer Standards beispielsweise erst einmal wieder erfordert, dass diese zumindest von einem Teil der Forschungsgemeinschaft anerkannt werden.

Eine erste Anregung zu einer kritischen Haltung möchten wir mit Blick auf die Frage nach der Anerkennung von Autoritäten geben. Welche Personen, Organisationen und Institutionen bestimmen den Diskurs um eine Thematik, Methode oder Auswahl zu verwendender Standards? Wie in jedem wissenschaftlichen Feld ist es wichtig, die Forschungslandschaft zu überblicken. Damit einher geht die Frage nach der Zugänglichkeit. Welche Möglichkeiten bestehen, sich an der Erstellung und Pflege von wissenschaftlichen Standards zu beteiligen? Gibt es Chancen, Teil dieser Gemeinschaften zu werden, die maßgeblich Standards setzen? Welche Faktoren begrenzen die Zugänglichkeit? Schließlich ist nach der Transparenz zu fragen. Innerhalb der Communities, vor allem dann, wenn sie organisationsförmig werden, muss transparent gemacht werden, wie Entscheidungen getroffen werden, wie eine transparente und nachvollziehbare Kommunikation nach außen gestaltet wird und wie divergierende Meinungen und Urteile in Einklang gebracht werden. Mit der Wahl eines Standards wird man oft, explizit oder implizit, auch Teil einer Community mit einer bestimmten Werthaltung. Es lohnt sich deshalb, ethische Gesichtspunkte in die Entscheidung für oder gegen einen Standard einzubeziehen. Auf der Ebene der Projektarbeit kann die Asynchronität schnell zu einer Herausforderung werden. So ist es bereits eine zeitaufwändige Aufgabe, Standards zu finden, die zu Fragestellung und Daten passen – Zeit, die oft im Vorfeld eines Projekts investiert werden muss. In einem optimierten Prozess der linearen Bearbeitung wäre die Wahl eines Standards zu Beginn eines Projektes abgeschlossen und die Durchführung der damit einhergehenden Methodik würde den weiteren Verlauf bestimmen. Tatsächlich aber ist bei einem agilen, adaptiven Projektmanagement in verschiedenen Phasen und auf unterschiedlichen Ebenen immer wieder über die Angemessenheit der anzuwendenden Standards nachzudenken, wie das vorherige Unterkapitel illustriert. Dies kann auch dadurch notwendig werden, wenn sich im Laufe der Arbeit herausstellt, dass andere Software oder andere Standards verwendet werden müssen. Da die Einarbeitung in Standards und ggf. deren Modifikation Zeit braucht, sind solche Standardwechsel während der Projektlaufzeit ein gewagtes Unterfangen, weil dadurch zusätzlicher Zeitaufwand entsteht, der nicht mehr im Einklang mit dem Projektplan steht (Geelhaar 2023). Besonders extern finanzierte Projekte leiden dann unter der Asynchronität von Finanzierung und Umsetzung. Im weiteren Verlauf tragen Zeit- und Geldmangel oft dazu bei, dass Standards nicht oder nicht in der angemessenen Gründlichkeit umgesetzt werden. Das wiederum erschwert das Projektreporting und die Nachnutzbarkeit. Eine umfassende und gründliche Dokumentation der Forschungspraxis ist als Vorbild für neue Projekte unerlässlich und muss ein fester Bestandteil von laufenden Projekten sein.

Hinsichtlich der Beschaffenheit von Standards ergeben sich Herausforderungen durch deren Revidier- und Erweiterbarkeit. Standards sind nicht unabänderlich, sie können revidiert, angepasst oder durch neue Standards ersetzt werden. Dies kann durch die Nutzenden der Standards selbst geschehen, wie dies im Falle des TEI-Konsortiums der Fall ist. Obschon eine gewählte Gruppe als Autorität über die Qualität der Standards wacht, bleibt es allen offen, an der Bearbeitung und Erweiterung von Standards zu arbeiten. Wo Standards durch eine kritische Masse an Nutzenden geschaffen, gestaltet und verwaltet werden, können diese Standards aus einem begrenzten Arbeitsumfeld in einen breiteren Diskurs überführt werden. Wenn dies durch große wissenschaftliche Institutionen erfolgt, gewinnt der jeweilige Standard an Rückhalt. Denn nur mit einer großen Community oder gewichtigen Institutionen, wie z. B. der Deutschen Nationalbibliothek mit ihrem GND-Normdaten-Standard, lassen sich Standards zumindest mittelfristig absichern.[43] So wird deutlich, dass Standards selbst Gegenstand wissenschaftlicher Debatten werden, statt sie nur zu begleiten und zu formen. Letztlich stellt sich noch

die Frage nach dem Ausbruch aus dem Standard. Mit einem Wandel und einer Fortentwicklung von Forschungsfragen müssen auch Standards weiterentwickelt werden. Stellt sich nach intensiven Vorüberlegungen heraus, dass die vorhandenen Standards nicht ausreichend sind, um die eigenen Daten aufzubereiten und zu bearbeiten, dann muss über Alternativen nachgedacht werden und ein kreativer Schaffensprozess einsetzen. Dabei muss beachtet werden, dass individuelle Lösungen nicht ohne Struktur vorgenommen werden sollten und weiterhin Konventionen und andere Standards beibehalten werden sollten. Es kann beispielsweise sein, dass der Metadatenwertestandard nicht ausreichend für die zu beschreibende Ressource ist und dafür neues Vokabular definiert werden muss.[44] Oder ein Projekt hat Spezifizierungen vorzunehmen, die auch mit dem sonst verwendeten Metadatenstrukturstandard nicht mehr abbildbar sind, sodass aus diesem ausgebrochen werden muss, weil beispielsweise viel feingliedriger und genauer analysiert werden muss, als es der Standard vorgibt. Oft lassen sich Teile von vorhandenen Standards weiternutzen und nur spezifische Elemente oder Werte müssen hinzugefügt werden. Hier sollte darauf geachtet werden, dass das neue Vokabular einheitlich ist und ebenso verwendet wird. In vielen Fällen können die Anpassungen zudem zurück in den Standard eingebracht werden. In den meisten Organisationen gibt es einen Austausch mit der Community darüber, inwieweit der jeweilige Standard ausreicht und welche Erweiterungen benötigt werden. Sofern es sich nicht um Einzelfälle handelt, werden Aufnahmen in den Standard in der Regel im jeweiligen Board diskutiert. Nichtsdestotrotz können Standards auf anderen Ebenen auch mit den neuen Werten beibehalten werden. So können in diesem Fall vielleicht der Metadatenstrukturstandard[45], der Metadateninhaltstandard[46] sowie das Domänenmodell[47] beibehalten werden. Nur eine umfassende Kenntnis der Standards und ihrer zugrundeliegenden Logik erlaubt es, Standards neu zu setzen.

Ob es überhaupt sinnvoll ist, projektspezifische Anforderungen und Lösungen in einen generellen Standard zu überführen, muss im Austausch mit der Community entschieden werden. Die Sorge, dass Standards überfrachtet und unübersichtlich werden, sollte nicht unbeachtet bleiben. Manche Lösungen können projektintern entwickelt werden und später anderen Projekten zur Verfügung gestellt werden. Hierbei ist es umso wichtiger, sein Vorgehen umfassend zu dokumentieren und die Dokumentation und Begründung zur Verfügung zu stellen.[48] Dadurch können sowohl Nutzende die entsprechenden Daten richtig einschätzen und weiterverarbeiten. Aber auch andere Projekte können sich in ihren Individuallösungen an anderen orientieren und somit Best Practices oder De-Facto-Standards etablieren.

6. Konsequenzen für eine digitale Quellenkritik

Das Ziel dieses Kapitels war es, die Sinnhaftigkeit von Standards ebenso wie die Notwendigkeit zur Reflexion über Standards im Arbeitsprozess aufzuzeigen. Konkret ging es darum, Leser:innen darin zu unterstützen, Standards, Standardisierungen und ihre Rolle im Arbeitsprozess zu kennen, ihren Nutzen wie auch ihre Nachteile einschätzen und mit ihnen reflektiert umgehen zu können. Zu diesem Zweck haben wir in einem ersten Schritt ausgeführt, dass wir bereits mehr oder weniger bewusst tagtäglich mit sehr vielen Standards arbeiten und dass Standards kein Selbstzweck sind, meist aufeinander aufbauen und/oder voneinander abhängen. Im zweiten Schritt haben wir auf den Einsatz von Standards in den Geisteswissenschaften fokussiert, um zu zeigen, dass digitale Standards aus den Fachwissenschaften heraus entwickelt werden, was einerseits die Anwendbarkeit und den Nutzen dieser Standards sichert, andererseits aber auch schnell zu Inkompatibilitäten führen kann. Daran anknüpfend haben wir dargelegt, dass solche Inkompatibilitäten insbesondere die computergestützte Analyse in der Forschung behindern, weshalb es so wichtig ist, auf passende Standards zu achten, wenn man z. B. maschinelles Lernen einsetzen möchte. Außerdem sollte klar geworden sein, dass nur standardisierte Daten maschinell sinnvoll lesbar sind und für die Verknüpfung von Datenbeständen in unterschiedlichen Systemen genutzt werden können. Normdaten als ein zentraler Standard kommt daher für die künftige Forschung eine hohe Relevanz zu.

Unserem Fokus auf den wissenschaftlichen Arbeitsprozess entsprechend haben wir anschließend mithilfe des Phasenmodells beschrieben, dass Standards bei der Planung, Sammlung, Aufbereitung, Analyse und Nachnutzung zum Einsatz kommen und es deswegen wichtig ist, sich mittels eines Datenmanagementplans vorab schon Einsatzszenarien von Standards zu überlegen, um die Pfadabhängigkeiten im Blick zu behalten, die sich durch die Wahl bestimmter Standards und Arbeitstechniken ergeben. Dem Prinzip des agilen Projektmanagements entsprechend kann es sehr wohl nötig werden, während der Projektlaufzeit auf andere Standards umzustellen, was mit hohen Arbeitskosten einhergeht, die jedoch reduziert werden können, um so konsistenter frühere Standards angewendet worden sind, weil nur dann die Überführung von einem Standard auf den nächsten computergestützt geschehen kann.

Dieser Aspekt verweist darauf, dass Standards nicht in Stein gemeißelt sind, sondern auf Entscheidungen beruhen, welche wiederum Dokumentation und Transparenz erfordern. Die Offenlegung dieser Entscheidungsprozesse ist daher ein wesentlicher Teil des (selbst-)kritischen Umgangs mit Standards in der Forschung. Dies ist für Produzenten historischen Datenmaterials oder von Analysesoftware ebenso wichtig wie für deren Rezipienten. Ebenso wichtig ist das Verständnis von bzw. Bewusstsein für den Nutzen und die Grenzen eines Standards, wie auch für die möglichen Abhängigkeiten, die man als Forscher:in eingeht, wenn man sich für bestimmte Lösungen entscheidet oder aber individuelle Lösungen anstrebt.

Aus Sicht der Geschichtswissenschaft als Rezipientin digitaler Daten erhöhen Standards die Nachvollziehbarkeit von getroffenen Entscheidungen, die sich in Datenstrukturen wiederfinden. Die Nutzung bei der Produktion von Daten hilft damit bei der späteren Einordnung und reflektierten Nutzung derselben Daten. Ein Standard wird somit zum Garanten für die Korrektheit der spezifischen Repräsentationsformen und -schemata, die der Standard abdeckt.

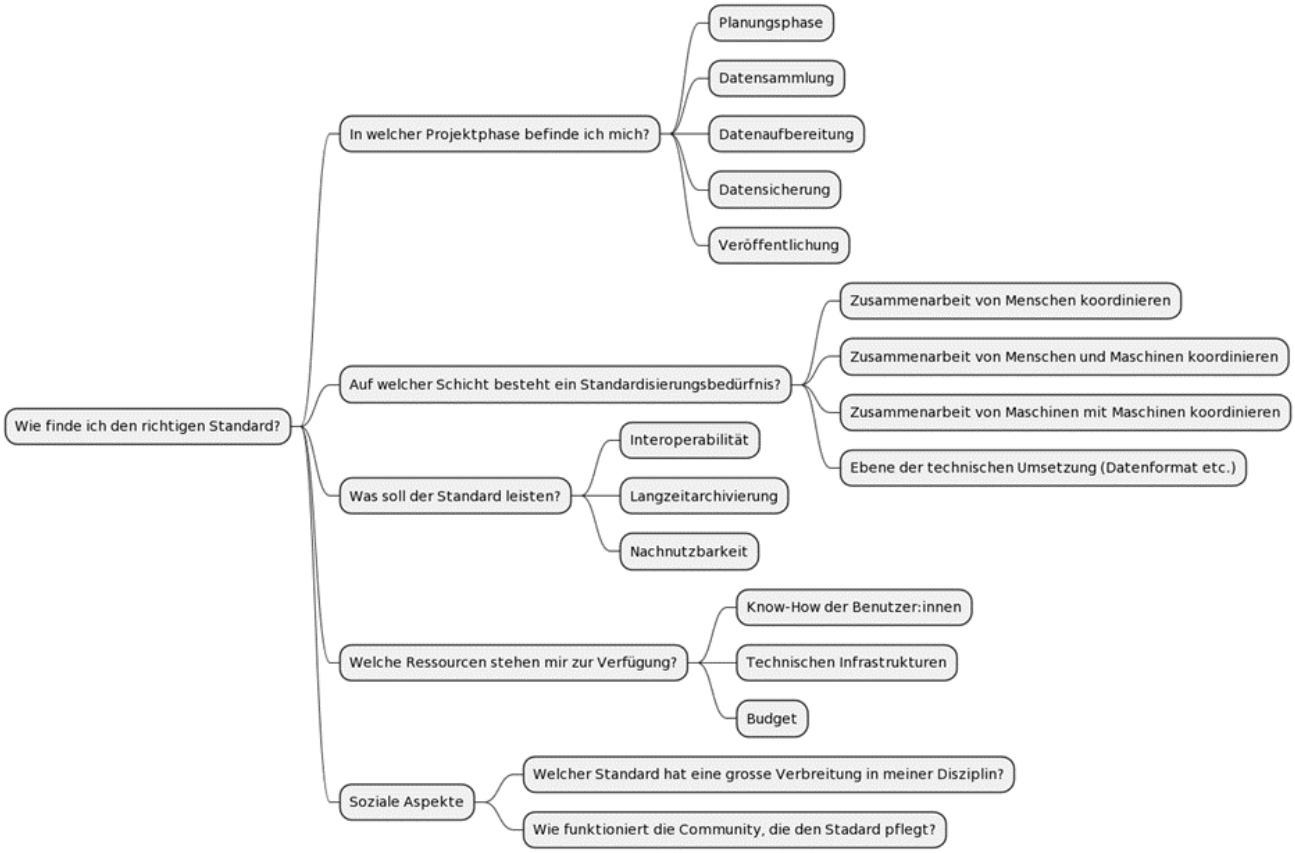

Abb. 4: Visualisierung von Merkmalen zum Umgang mit Standards. Eigene Abbildung. Lizenziert unter einer CC-BY 4.0 Lizenz.

Abschließend geben wir dennoch zu bedenken, dass die Nutzung von Standards nicht nur Vorteile bringt, sondern auch den Blick auf bearbeitete Quellen oder Dokumente trüben kann, da ein Standard über die Quelle gelegt wurde. Die kritische Kommentierung eingesetzter Standards und deren ständige Weiterentwicklung wird somit in absehbarer Zukunft eine Aufgabe der digital arbeitenden Wissenschaftsbereiche bleiben.

Literatur

AG Digitales Publizieren. 2021. “Digitales Publizieren in den Geisteswissenschaften: Begriffe, Standards, Empfehlungen.” HTML,XML,PDF. https://doi.org/10.17175/WP_2021_001.

Avanço, Karla 2021. “Understanding the FAIR Principles.” In The road to FAIR (Blog). https://roadtofair.hypotheses.org/47 (zugegriffen: 4. Juli 2024).

Artstein, Ron. 2017. “Inter-Annotator Agreement.” In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 297–313. Dordrecht: Springer. http://dx.doi.org/10.1007/978-94-024-0881-2.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons u. a. 2020. “The CARE Principles for Indigenous Data Governance.” *Data Science Journal* 19 (1): 43. https://doi.org/10.5334/dsj-2020-043.

Gilliland, Anne J. 2016. “Setting the Stage.” In *Introduction to Metadata*, herausgegeben von Murtha Baca. Los Angeles: Getty Research Institute. https://www.getty.edu/publications/intrometadata/setting-the-stage/ (zugegriffen: 4. Juli 2024).

Hiltmann, Torsten. 2018. “Forschungsdaten in Der (Digitalen) Geschichtswissenschaft. Warum Sie Wichtig Sind Und Wir Gemeinsame Standards Brauchen.” *Hypotheses. Digitale Geschichtswissenschaft Das Blog Der AG Digitale Geschichtswissenschaft Im VHD.* https://

digigw.hypotheses.org/2622 (zugegriffen: 4. Juli 2024).

Huang, Angela, and Ulla Kypta. 2011. “Ein neues Haus auf altem Fundament.: Neue Trends in der Hanseforschung und die Nutzbarkeit der Rezesseditionen.” *Hansische Geschichtsblätter* 129: 213–30. https://doi.org/10.21248/hgbll.2011.56.

Kuster, Jürg, Christian Bachmann, Mike Hubmann, Robert Lippmann, and Patrick Schneider. 2022. *Handbuch Projektmanagement: agil – klassisch – hybrid*. 5. vollständig überarbeitete und erweiterte Auflage. Berlin [Heidelberg]: Springer Gabler.

Mabillon, Jean. 1681. *De Re Diplomatica*. Vol. libri sex. Paris.

Preußig, Jörg. 2020. *Agiles Projektmanagement: Agilität und Scrum im klassischen Projektumfeld*. 2. Auflage. Freiburg München Stuttgart: Haufe Group.

Romein, C. Annemieke, Tobias Hodel, Femke Gordijn, Joris J. van Zundert, Alix Chagué, Milan van Lange, Helle Strandgaard Jensen u. a. 2022. “Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions. Starting the Conversation on How We Could Get It Done,” November. https://doi.org/10.5281/zenodo.7267245.

Simonite, Tom. 2019. “The Best Algorithms Struggle to Recognize Black Faces Equally.” *Wired*, July 22, 2019. https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/ (zugegriffen: 4. Juli 2024).

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg u. a. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Yates, JoAnne, and Craig Murphy. 2019. *Engineering Rules: Global Standard Setting since 1880*. Hagley Library Studies in Business, Technology, and Politics. Baltimore: Johns Hopkins University Press.

Endnoten

1 Der vorliegende Text wird von den Autor:innen zu gleichen Teilen verantwortet. Konzeption, Methodologie und Verschriftlichung wurden gemeinsam unternommen.

2 Zum DTA-Basisformat siehe online: https://www.deutschestextarchiv.de/doku (zugegriffen: 15. April 2024).

3 Siehe online: https://www.din.de/de/ueber-normen-und-standards/basiswissen. (zugegriffen: 15. April 2024)

4 Einen begrenzten Versuch zur Einordnung und Bewertung von Standards findet sich indes im Working Paper “Digitales Publizieren in den Geisteswissenschaften” (AG Digitales Publizieren 2021)

5 Vgl. DWDS s. v. „Standard”, online: https://www.dwds.de/wb/Standard (zugegriffen: 15. April 2024).

6 Vgl. Wikipedia, s. v. „Standard“, online: https://de.wikipedia.org/wiki/Standard. (zugegriffen: 15. April 2024).

7 Über die verschiedenen Standards im Bibliothekswesen informiert die Arbeitsstelle für Standardisierung (AfS) an der Deutschen Nationalbibliothek auf ihrer Webseite, online: https://www.dnb.de/DE/Professionell/Standardisierung/Standards/standards_node.html (zugegriffen: 15. April 2024).

8 Einen Eindruck von der Vielzahl an Zitationsstilen gibt die Literaturverwaltungssoftware Zotero, die in ihrem Zotero Style Repository 10.365 Stile aufführt, siehe online: https://www.zotero.org/styles/ (zugegriffen: 15. April 2024).

9 Vgl. die Usage statistics of character encodings for websites, siehe online: https://w3techs.com/technologies/overview/character_encoding (zugegriffen: 15. April 2024).

10 Siehe online: https://home.unicode.org/about-unicode/ (zugegriffen: 15. April 2024).

11 Gemeint sind damit z. B. die Deutsche Forschungsgemeinschaft (DFG), das deutsche Bundesministerium für Bildung und Forschung (BMBF), der Schweizer Nationalfonds (SNF), der österreichische Wissenschaftsfonds (FWF) oder die österreichischen Forschungsförderungsgesellschaft (FFG).

12 Siehe online: https://www.iso.org/standards.html (zugegriffen: 15. April 2024).

13 Siehe online: https://www.w3.org/standards/faq#std (zugegriffen: 15. April 2024).

14 Siehe online: https://www.nfdi.de/ (zugegriffen: 15. April 2024).

15 Siehe online: https://www.nfdi.de/verein/ (zugegriffen: 15. April 2024).

16 Die Gesetzgebung hat die rechtlichen Rahmenbedingungen für die Nutzung von Daten und Datenbanken zum Zweck des Data und Text Mining deutlich verbessert. Siehe für Deutschland UrHG § 60d und für die Schweiz URG 24d.

17 Siehe online: https://doi.org/10.5281/zenodo.3923601. Zusätzliche Erläuterungen und Preisbeispiele präsentiert die DFG mittlerweile in einem eigens dafür eingerichteten Portal (zugegriffen: 15. April 2024).

18 Diese Identifier selbst sind wiederum Standards zur eindeutigen Adressierung der jeweiligen Einheiten. ORCID steht für die Open Researcher Contributor Identification Initiative, die 2010 gegründet worden ist, um einen de-facto-Standard zur Autorenidentifikation festzulegen. Die zu diesem Zweck vergebene ORCID-Nummer entspricht ISO 27729 und ist damit ein International Standard Name Identifier (ISNI), der sich aus 16 Zeichen in vier Viererketten zusammensetzt. Siehe https://orcid.org/ (zugegriffen: 15. April 2024).

19 Siehe online: https://zenodo.org/ (zugegriffen: 15. April 2024). Zenodo ist ein Repositorium für wissenschaftliche Daten in verschiedenen Formen, das in Europa angesiedelt ist, von der Europäischen Kommission in Form des OpenAIRE Programms finanziert und am CERN gehostet wird.

20 Siehe ebenfalls https://www.go-fair.org/fair-principles/ und https://www.forschungsdaten.org/index.php/FAIR_data_principles (zugegriffen: 15. April 2024).

21 Für eine ausführliche Darstellung der CARE-Prinzipien siehe https://www.gida-global.org/care (zugegriffen: 15. April 2024).

22 Siehe https://www.cei.lmu.de/ (zugegriffen: 15. April 2024). Inwieweit dieser Standard aber noch angewendet wird, lässt sich nicht ermitteln.

23 Siehe https://cidoc-crm.org/ (zugegriffen: 15. April 2024).

24 Siehe https://gnd.network/Webs/gnd/DE/Home/home_node.html (zugegriffen: 15. April 2024).

25 Siehe https://www.dnb.de/DE/Professionell/Standardisierung/Standardisierungsausschuss/standardisierungsausschuss_node.html (zugegriffen: 15. April 2024).

26 So z. B. in Form von TEI-Header XML-Repräsentationen bei digitalen Editionen.

27 Hier wären Dublin Core oder MARC 21 zu nennen. Dublin Core ist ein Metadatenschema für elektronische Ressourcen, das von der Dublin Core Metadata Initiative (DCMI) gepflegt wird, siehe https://www.dublincore.org (zugegriffen: 15. April 2024).

28 Etwa wie Wissensontologien beschrieben werden.

29 Das Feld wird genutzt zur Bezeichnung des “Main Entry-Personal Name”, siehe https://www.loc.gov/marc/bibliographic/bd100.html (zugegriffen: 15. April 2024).

30 Siehe https://www.dublincore.org/specifications/dublin-core/dcmi-terms/elements11/creator/ (zugegriffen: 15. April 2024).

31 Dies ist sichtbar an den Bemühungen eigene Ontologien zu etablieren, etwa Records in Context in der Archivwelt und FRBRoo in den Bibliotheken oder CIDOC-CRM in Museen.

32 Siehe https://mufi.info/q.php?p=mufi (zugegriffen: 15. April 2024).

33 Siehe online: https://www.salamanca.school/de/index.html (zugegriffen: 15. April 2024).

34 Als wenige repräsentatives Beispiel können die aktuellen Large Language Models eingebracht werden. Sogar die offenen und verhältnismäßig kleinen Systeme, wie etwa BLOOM, basieren auf Milliarden von Token, im Falle von Bloom 366 Milliarden Token. Siehe online: https://huggingface.co/bigscience/bloom (zugegriffen: 15. April 2024).

35 Siehe dazu für die Erkennung von Handschriften den Ansatz von HTRunited: https://htr-united.github.io/ (zugegriffen: 15. April 2024). Zur Frage des Datenaustausches siehe auch Romein u. a. 2022.

36 Für weitere Informationen zu Datenmanagementplänen siehe für die Schweiz: https://www.snf.ch/de/FAiWVH4WvpKvohw9/thema/forschungspolitische-positionen; (zugegriffen: 15. April 2024). Für Deutschland: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_de.pdf (zugegriffen: 15. April 2024). Seitens der Max Planck Digital Library: https://rdm.mpdl.mpg.de/before-research/data-management-plans/ (zugegriffen: 15. April 2024).

37 Informationen hierzu bietet die DFG unter https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/beantragbare_mittel/index.html (zugegriffen: 15. April 2024).

38 Unter einem Mapping versteht man den Abgleich verschiedener Standards. Einem konkreten Wert eines Standards wird dabei die Entsprechung in einem anderen Standard zugeordnet. Dies kann entweder manuell in Tabellenform geschehen oder automatisiert mit Transformationsskripten. Schließlich bietet sich die Möglichkeit, Ontologien, die nach OWL oder ähnlichen Systematiken in Linked Open Data modeliert wurden, über sogenannte Crosswalks zu verbinden.

39 https://fortext.net/ueber-fortext/glossar/inter-annotator-agreement-iaa (zugegriffen: 15. April 2024).

40 Hilfreich in diesem Zusammenhang sind auch Dokumentationen, die durch Dritte erstellt werden. So bietet das Darmstädter Projekt

forText Tutorials und YouTube-Videos für verschiedene Softwarelösungen und Arbeitstechniken wie Annotation oder Netzwerkanalyse an. Siehe https://fortext.net/ (zugegriffen: 15. April 2024).

41 Vergleiche auch unten die Phase Nachnutzung.

42 Beispiele hierfür können in der Mailingliste der TEI gefunden werden.

43 GND = Gemeinsame Normdatendatei, siehe https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html (zugegriffen: 15. April 2024).

44 Der Metadatenwertestandard beschreibt, welche Werte wie in einem Element erscheinen dürfen, Vgl. Gilliland 2016.

45 Der Metadatenstrukturstandard beschreibt, was die Elemente sind, mit denen eine Ressource beschrieben werden darf, siehe auch Gilliland 2016.

46 Der Metadateninhaltstandard beschreibt, welche Regeln befolgt werden sollen, um Metadaten zu erzeugen, siehe auch Gilliland 2016.

47 Das Domänenmodell für Metadaten beschreibt, was die grundsätzliche abstrakte Perspektive auf die Ressourcenbeschreibung in der Domäne ist.

48 Als Beispiel kann hier das interne Wiki der Sammlung Schweizerischer Rechtsquellen angeführt werden, in dem die genaue Nutzung der TEI-Richtlinien festgelegt und vor allem Erweiterungen, etwa die Nutzung eigener Attribute, dokumentiert wird: https://www.ssrq-sds-fds.ch/wiki/ (zugegriffen: 15. April 2024).

Zitierweise

Diekjobst, Anne / Geelhaar, Tim / Hodel, Tobias / Mähr, Moritz / Seltmann, Melanie (2024): 2.3. Mit Standards forschen und Handlungsräume schaffen. In: Living Handbook "Digitale Quellenkritik". Version 1.0. hrsg. v. Deicke, Aline; Geiger, Jonathan D.; Lemaire, Marina; Schmunk, Stefan. https://doi.org/10.5281/zenodo.12656767

Metadaten

Autor:innen	Diekjobst, Anne ; Geelhaar, Tim ; Hodel, Tobias ; Mähr, Moritz ; Seltmann, Melanie ;
Language	Deutsch
DOI	10.5281/zenodo.12656767
Creative Commons Lizenztyp	Attribution CC BY (4.0)