Rat für
Informations
Infrastrukturen

7

HERAUSFORDERUNG DATENQUALITÄT

Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel

INHALT

ABKÜRZUNGSVERZEICHNIS

BMBF	Bundesministerium für Bildung und Forschung
CERN	Conseil européen pour la recherche nucléaire
CTS	Core Trust Seal
DESY	Deutsches Elektronen-Synchrotron
DFG	Deutsche Forschungsgemeinschaft
DIN	Deutsches Institut für Normung
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
FAIR	Findable, Accessible, Interoperable, Reusable
FDM	Forschungsdatenmanagement
GWK	Gemeinsame Wissenschaftskonferenz
ISO	International Organization for Standardization
IUPAC	International Union of Pure and Applied Chemistry
IUPAP	International Union of Pure and Applied Physics
KI	Künstliche Intelligenz
MARC	Machine-Readable Cataloging
NFDI	Nationale Forschungsdateninfrastruktur
PSI	Private Sector Information
RatSWD	Rat für Sozial- und Wirtschaftsdaten
RDA	Research Data Alliance
RfII	Rat für Informationsinfrastrukturen
SFB	Sonderforschungsbereich

KURZFASSUNG

Die qualitative Dimension wissenschaftlicher beziehungsweise wissenschaftlich erzeugter Daten berührt die ureigene Autonomie der Wissenschaft und deren fachliche Details: Forscherinnen und Forscher entscheiden selbst über die Wahl ihrer theoretischen Ansätze und die in diesem Rahmen eingesetzten Methoden und Instrumente, um Forschungsdaten zu erheben und auf dieser Basis Forschungsergebnisse zu erzielen. Die Qualität der Forschungsdaten hängt somit von forschungseigenen Maßstäben ab.

Der digitale Wandel in Wissenschaft, Wirtschaft und Gesellschaft schafft hier einen neuen Kontext. Die Qualitätsproblematik stellt sich heute aufgrund der Vielzahl mit digitaler Technologie erzeugter und erzeugbarer Daten sowie deren Transfer- und Verwendungsmöglichkeiten in veränderter Form. In der Wissenschaft betrifft dies alle Disziplinen, Fachgebiete und Forschungsformen – wobei graduelle Unterschiede bestehen, je nachdem, wie „datenintensiv" hier bereits in der Vergangenheit gearbeitet wurde. Neue Fragen stellen sich aber auch an Schnittstellen zwischen der Forschung und den bislang hiervon getrennt gedachten Infrastrukturbereichen – den sogenannten „Wissensspeichern", etwa Bibliotheken, Archiven, Sammlungen, Rechenzentren und Datenzentren.

Auch, was Zukunftsfragen der Gesellschaft angeht, findet das Thema Datenqualität heute vermehrt Aufmerksamkeit: Die mit großen digitalen Datenmengen erzeugten Simulationen der Klima- und Erdsystemforschung zum Beispiel stehen öffentlich zur Diskussion. Kommerzielle Unternehmen bieten zunehmend eigene, „hypothesenfreie" Datenauswertungen an, und die öffentlich getragene Wissenschaft gerät in einigen Bereichen in Abhängigkeit von Infrastrukturleistungen und Daten privatwirtschaftlicher Akteure. Vor diesem Hintergrund haben Wissenschaft und Wissenschaftspolitik Datenqualität als eine nicht länger zu unterschätzende Herausforderung erkannt. Sie ist zum Gegenstand des Nachdenkens über neue rahmensetzende und regulierende Maßnahmen geworden.

In diesem Positionspapier untersucht der Rat für Informationsinfrastrukturen (RfII) aktuelle Herausforderungen für Datenqualität im Wissenschaftssystem und leitet hieraus Empfehlungen ab.

In der Ausgangslage (Kap. 1) macht der RfII darauf aufmerksam, dass bisherige Konzeptualisierungen von Datenqualität nur bedingt zu den spezifischen Bedürfnissen der wissenschaftlichen Erkenntnisgewinnung passen. Denn sie entstammen häufig einem Management(-theorie)-Kontext und müssen mit Augenmaß in die Logik von Forschungsprozessen übersetzt werden. Der RfII beschreibt verschiedene Konzepte und Instrumente, die sich zur Steuerung von Datenqualität eignen. Eine Expertise, die hierzu Einzelheiten darlegt, findet sich in einem gesonderten Anhang zu diesem Positionspapier. Gemeinsam ist allen

existierenden Datenqualitätsmodellen, dass sie keinen hinreichenden Aufschluss über die Frage geben, wie Daten und ihre Dokumentation effektiv mit den Anforderungen digitaler Forschungsprozesse in unterschiedlichen Disziplinen und Feldern verknüpft werden können.

Um konkrete Problemlagen abzubilden, durchschreitet der RfII in Kapitel 2 die Stationen des wissenschaftlichen Datenlebenszyklus und skizziert, welche Herausforderungen sich der Forschung in den jeweiligen Phasen der Datentransformation stellen. Dabei zeigen sich an den verschiedenen Schnittstellen – von der Erhebung der Daten bis zu deren Weitergabe und Publikation – zahlreiche Schwierigkeiten, die Datensicherung und die damit verbundene Steigerung der Datenqualität zwischen Managementaufgabe und Forschungspraxis zu synchronisieren. Der RfII sieht hierin eine Mehrebenen-Aufgabe, bei der Forschende und Mitarbeiter der Infrastrukturen eng zusammenarbeiten müssen. Leitende Gütekriterien, gekoppelt an methodische Standards müssen hierbei aus den wissenschaftlichen Communities und Fachgemeinschaften kommen.

Die Herausforderung Datenqualität potenziert sich durch weitere Entwicklungen im Wissenschaftssystem, die ganz allgemein Fragen der Forschungsqualität betreffen. Der RfII verweist in Kapitel 3 auf die sogenannte Replikationskrise sowie auf die viel diskutierte Überlastung des Begutachtungssystems. Die quantitative Überdehnung der Publikationsanforderungen sieht er ebenso als Problem für ein intensives Engagement in Sachen Datenqualität wie – vor allem auf internationaler Ebene – mangelnde wissenschaftskonforme Rahmensetzungen und Regulierungen. Ein auch in den bereits länger datenintensiv arbeitenden Forschungsfeldern virulentes und sich heute zuspitzendes Problem ist zudem die Abhängigkeit von Messinstrumenten beziehungsweise Hard- und Software-Komponenten kommerzieller Hersteller.

Der RfII leitet aus diesen Einschätzungen in Kapitel 4 ein Bündel von Empfehlungen ab, die eine gemeinsam beziehungsweise im Dialog wahrzunehmende Qualitätsverantwortung einfordern. Adressaten dieser Empfehlungen sind:

- Datenproduzenten, Verarbeiter und diverse Nachnutzer von Forschungsdaten,
- Forschende, ihre Fachgemeinschaften und die Infrastrukturanbieter,
- Hochschulen und außeruniversitäre Forschungseinrichtungen als organisatorische Gestalter sowie
- die Wissenschaftsorganisationen, die Förderinstitutionen und die Fachministerien in Bund und Ländern, die auch für die Datenqualitätsanstrengungen der Wissenschaft den monetären und programmatischen Rahmen setzen.

Der RfII legt seinen Empfehlungen einen prozessorientierten Begriff von Datenqualität zugrunde. Dieser hat sowohl Transformationen in den Forschungsprozessen als auch die aktuell rapiden Veränderungen der technischen Möglichkeiten zur

Datenverarbeitung im Blick. Offenheit und Dynamik, aber auch eine enge Anbindung an wissenschaftliche Methoden und Forschungsformen müssen wesentliche Leitlinien einer zu entwickelnden Datenkultur sein.

Dies beinhaltet, eine angemessene, an fachlichen Standards orientierte Dokumentation von Forschungsdaten als wissenschaftliche Kernaufgabe und Bestandteil des Berufsethos zu begreifen. Die Sicherung und Steigerung von Datenqualität ist ein Grundwert der guten wissenschaftlichen Praxis. Der RfII fordert die Fachgemeinschaften auf, diesen Grundwert noch stärker als bislang in der Methodenausbildung der Disziplinen und Forschungsfelder zu berücksichtigen. Hierzu gehört auch, Forschungsprozesse und Forschungsinfrastrukturen – einschließlich Bibliotheken, Rechenzentren, etc. – deutlicher zu verbinden. Gleichzeitig benötigt die Arbeit mit Forschungsdaten ein höheres Maß an fachlicher Reputation.

Für die Steigerung der Datenqualität ist die Berücksichtigung der unterschiedlichen Schnittstellen zum Datenlebenszyklus in jedem Stadium des Forschungsprozesses unabdingbar. Hierfür bedarf es kohärenter Datenbeschreibungen und -deklarationen. Vormals implizites Wissen muss expliziert und – möglichst weitgehend – maschinenlesbar dokumentiert werden. Der RfII fordert dazu auf, angemessene technische Prüfverfahren weiterzuentwickeln. Kaum durchschaubare Hard- und Software-Eigenschaften oder die Beendigung von infrastrukturrelevanten Produktlinien durch kommerzielle Anbieter erschweren die wissenschaftlichen Bemühungen um hohe Datenqualität. Hier empfiehlt der RfII erhebliche Anstrengungen der Fachgemeinschaften und Fachgesellschaften, um anbieterseitig eine höhere Produkttransparenz einzufordern.

Eine Chance, um zugleich der Arbeit mit Forschungsdaten mehr Anerkennung zu verleihen, die Datenqualität zu steigern und Replikationsstudien attraktiver zu machen, sieht der RfII in der Ausdifferenzierung von wissenschaftseigenen Datenprodukten. Entsprechende Produktformen reichen von der Ko-Publikation von Forschungsergebnis und zugehörigem Datensatz bis hin zum Aufbau kuratierter Datensammlungen, die bereits Apps für die Weiterverwendung der Daten enthalten können. Auch die Etablierung einer eigenständigen Rezensionskultur für Forschungsdaten wäre hilfreich – nicht als Nischenprodukt, sondern an sichtbarer Stelle, also in angesehenen Fachzeitschriften.

Für die Forschungs- und Informationsinfrastrukturen hält der RfII dort, wo dies noch nicht geschieht oder noch im Aufbau ist, eine an wissenschaftlichen Standards orientierte Qualitätssicherung, zum Beispiel durch Evaluationen, für unabdingbar. Nur so lassen sich Infrastrukturen auch in den bislang weniger datenintensiven Forschungsbereichen dauerhaft zu Kompetenzzentren weiterentwickeln, die standardsetzend auch in die Forschung hineinwirken können. Dabei können unterschiedliche Wege beschritten werden, die auch das Lernen

von Partnereinrichtungen einschließen. Kooperationen sollten sich auch auf die laufende Verbesserung und Anpassung der technischen Infrastruktur beziehen, die nach Auffassung des RfII durchgehend höchsten Ansprüchen genügen muss, um Forschung in Deutschland international konkurrenzfähig zu halten.

Eine Grundvoraussetzung für Datenqualität sind die Fähigkeiten der in Forschung und Infrastruktur Beschäftigten. Hierzu hat der RfII 2019 mit DIGITALE KOMPETENZEN – DRINGEND GESUCHT eigenständige Empfehlungen veröffentlicht. Ein Aufbrechen der Versäulung in den Ausbildungen von Wissenschaftlern einerseits und von Infrastrukturpersonal andererseits betrachtet er als ebenso wichtig wie die generelle Steigerung der IT-Kompetenz in allen Disziplinen und eine ständige Weiterbildung quer zu den formalen wissenschaftlichen Qualifizierungszielen.

Die Förderpolitik ist heute erst rudimentär auf die Bedeutung der Datenqualität im digitalen Wandel eingestellt. Grundsätzlich müssen die Laufzeiten von Förderprojekten flexibler ausgestaltet werden können, um Datenaspekten bereits in der Phase der Antragskonzeption genügend Raum zu geben. Hiermit zusammenhängend sollte in der Bewertung zurückliegender Forschungsleistungen dem qualitativen Ertrag von Forschung (zum Beispiel gut dokumentierte Datensätze) der Vorzug vor hoher Quantität eingeräumt werden. Bis öffentliche Förderer ihre Programme entsprechend erweitert haben, sieht der RfII für Stiftungen ein großes Potenzial, mit innovativen Förderformaten zur Weiterentwicklung von Datenqualität als Katalysatoren zu wirken. Des Weiteren empfiehlt der RfII, die Erstellung innovativer Datenprodukte als eigenständigen Förderbereich zu etablieren.

Den Hochschulen und außeruniversitären Forschungseinrichtungen rät der RfII, Datenqualität als ein Kernelement in ihren Forschungsstrategien zu verankern. Hiermit sollten neue Kooperationen verknüpft werden, die Infrastrukturbereiche wie zum Beispiel die Rechenzentren, Bibliotheken sowie universitäre und außeruniversitäre Sammlungen aktiv einbeziehen. Insbesondere infrastrukturtragende außeruniversitäre Forschungseinrichtungen können künftig eine führende Rolle in der Entwicklung von Standards für das Datenmanagement spielen. Hochschulen sollten das Themenfeld Forschungsdaten überdies stärker in der Lehre verankern und entsprechende Expertise bei Berufungen berücksichtigen. Bund und Länder sieht der RfII in der Pflicht, die Wissenschaftsinstitutionen in der Weiterentwicklung der Datenqualität in Deutschland tatkräftig zu unterstützen. Der RfII sieht hier auch die durch die Gemeinsame Wissenschaftskonferenz (GWK) eingerichtete Nationale Forschungsdateninfrastruktur (NFDI) als wichtigen Akteur an. Darüber hinaus sollten Bund und Länder weiter bestrebt sein, gerade für erfolgreiche, aber prekär finanzierte Forschungs- und Informationsinfrastrukturen an Hochschulen langfristige existenzsichernde Optionen zu suchen.

Der europäische FAIR-Prozess hat sich als ein erfolgreicher Weg erwiesen, das Bewusstsein für Mindestanforderungen an wissenschaftliche Daten und deren Zugänglichkeit zu schärfen. Der RfII befürwortet eine inhaltliche Vertiefung des FAIR-Prozesses, um Integration und Transfer von Daten im europäischen und internationalen Forschungsraum noch weiter voranzutreiben. Er empfiehlt, FAIR künftig stärker mit disziplin- und forschungsfeldspezifischen Gütekriterien zu koppeln, um die Qualität und Verwendungsmöglichkeiten „FAIRer" Daten zu erhöhen. Eine FAIR ergänzende europaweite Qualitätsoffensive für Forschungsdaten würde hierbei zielführend sein. Nötig ist aber auch eine nochmalige Steigerung der Kommunikationsanstrengungen in Wissenschaft und Gesellschaft, um die im digitalen Wandel notwendige Anforderung, implizites Wissen für multiple wissenschaftliche (und ggf. weitere) Verwendungen zu explizieren, zum Grundwert einer globalen Datenkultur zu machen.

EINLEITUNG: DATENQUALITÄT ALS UNTERSCHÄTZTES THEMA

Wissenschaft basiert auf einem Qualitätsversprechen: Forschungsergebnisse werden auf der Grundlage akzeptierter methodischer Grundsätze erzielt. Die im Forschungsprozess verwendeten und erzeugten Daten genügen dabei hohen Qualitätsanforderungen, die von den wissenschaftlichen Fachgemeinschaften selbst gesetzt und kontrolliert werden. Die Gewinnung neuer wissenschaftlicher Erkenntnisse hängt in diesem Kontext eng mit der Steigerung der Qualität von Daten und datenbasierten Prozessen zusammen. Hierauf bauen auch die gesellschaftlichen Erwartungen in die Leistungsfähigkeit der Wissenschaft auf.

Wissenschaftlich qualifiziertes Wissen im digitalen „Weltenwandel“

Diese elementaren Merkmale der Produktion wissenschaftlich qualifizierten Wissens unter den Bedingungen eines „Weltenwandels“ zu verwirklichen, der durch die forcierte Digitalisierung angetrieben wird[1], stellt die Forschung vor neue Herausforderungen. Denn Datenverarbeitung ist heute in allen wissenschaftlichen Disziplinen und Forschungsfeldern mit hohem Automatisierungsgrad in enorm komplexen und vielfältigen Formen sowie mit technisch nahezu unbegrenzten Vernetzungsoptionen möglich. Forschungsprozesse haben sich teils drastisch verändert. Datenmengen wachsen. Es steigen aber auch die Abhängigkeiten, Prüfprobleme und neue Formen der Intransparenz, die Aussagen über Datenqualität äußerst voraussetzungsvoll machen – gerade auch über disziplinäre Schranken hinweg. Die Forschung ist im digitalen Weltenwandel in einer besonderen Position: Einerseits treibt sie ihn selbst massiv voran – vor allem im Bereich der Mathematik und Informatik sowie mittels datenintensiver Technologien in den Natur- und Ingenieurwissenschaften – andererseits ist sie ihm auch ausgesetzt, was Anpassungen des Methodenverständnisses und des Umgangs mit Daten in allen Disziplinen betrifft.

Neue Anforderungen an Datenqualität

Nur wenn die Qualität von Daten auch unter diesen Bedingungen wissenschaftlichen Ansprüchen gerecht wird, kann die Wissenschaft weiterhin leisten, was die Gesellschaft von ihr erwartet. Datensätze und Methodik müssen auch unter digitalen Vorzeichen transparent und von anderen Wissenschaftlerinnen und Wissenschaftlern nachvollzogen werden können. Die auf dieser Grundlage gewonnenen Ergebnisse müssen unter bestimmten jeweils zu spezifizierenden Umständen valide oder sogar replizierbar sein. Ohne gehaltvolle und gut dokumentierte Daten haben Forschungsergebnisse und sich aus ihnen entwickelnde Innovationen keinen Bestand. Hier sind auch das Vertrauen beziehungsweise der Rückhalt, den ein funktionsfähiges Wissenschaftssystem in der Gesellschaft benötigt, berührt: Beides würde bereits mittelfristig erodieren, wenn berechtigte Zweifel an der Qualität wissenschaftlicher Daten aufkommen. Dies macht die

[1] Den Begriff des „Weltenwandels“ zur Umschreibung der Zäsur, die Digitalität für die Wissenschaft heute bedeutet, entnimmt der RfII: Strohschneider (2018) – Neujahrsansprache.

Qualitätssicherung von Daten und Datenprozessen zu einer permanenten Herausforderung, sowohl für die konkreten Forschungsszenarien in allen Disziplinen als auch für das Wissenschaftssystem und seine gesellschaftliche Rolle im Ganzen.

Merkmale von Datenqualität in der Wissenschaft

Doch was genau ist eigentlich Datenqualität in einem wissenschaftlichen Kontext? Tatsächlich hat die Wissenschaft das, was der Begriff meint, lange Zeit lediglich unter der generellen Überschrift der „Methoden" diskutiert. Eine wissenschaftsgerechte Qualität von Daten lässt sich daher nicht ohne Weiteres hinreichend tiefenscharf definieren. Der RfII hat hierzu 2016 einen ersten Versuch unternommen. Demnach umfasst der Begriff Datenqualität sowohl allgemeine und unter Methodengesichtspunkten geforderte typische Eigenschaften der Daten als auch deren durch qualitätssichernde Maßnahmen gegebenenfalls zusätzlich geschaffene Eignung für eine weitere Nutzung.[2] Detaillierter ausgearbeitete Qualitätsmaßstäbe und -modelle für digitale Daten entstammen in einer historischen Betrachtung zunächst der Managementtheorie, der Wirtschaftsinformatik und Überlegungen zur industriellen Verfahrensoptimierung in Produktionskreisläufen. Solche Modelle passen auf wissenschaftliche Forschungsprozesse nur sehr begrenzt, da Wissenschaft

- methodisch-kontrolliert, aber nicht produkt-, sondern ergebnisoffen verfährt,
- Qualität auch auf dem Weg zum Ergebnis fortlaufend (auch für künftige, noch unbekannte Forschungsfragen) steigern will,
- Wissensbestände und Daten anders als ein Unternehmen breit zirkulieren lässt sowie (idealerweise: altruistisch) teilt,
- zu Zwecken der Referenzierung von Ergebnissen, der Dokumentation von Forschungslinien sowie der Zeitreihenbildung in Langzeitstudien Daten nachhaltiger archiviert als die Wirtschaft und
- im Sinne eines kumulativen Erkenntnisfortschritts Daten stetig wiederverwendet beziehungsweise in nicht vorhersehbaren zeitlichen Abständen „alte" Datenbestände zur Beantwortung neuer Fragen wieder aufgreift.

Alte und neue Qualitätsdiskurse in der Wissenschaft

Gütekriterien und standardisierte Verfahren der Datenbewertung sind immer schon ein Thema in der Wissenschaft gewesen – auch bevor das digitale Zeitalter in aller Munde war. Ebenso gibt es eine der Digitalisierungsthematik vorlaufende öffentliche Debatte über allgemeine Krisenphänomene in der Wissenschaft. Diese betreffen nicht direkt die Datenqualität, sondern Fragen der Leistungsbewertung und der guten wissenschaftlichen Praxis. Diskutiert wird

[2] Vgl. RfII (2016) – Leistung aus Vielfalt, Anhang A, S. A-8 f. Im Regelfall sind die Forschungsmethoden, die über Auswahl, Erhebung, Verarbeitung und Weitergabe der Daten entscheiden, wiederum abhängig von theoretischen Vorannahmen bzw. Weichenstellungen in einzelnen Disziplinen und Forschungsfeldern bzw. der unterschiedlichen „Schulen", die sich in multiparadigmatischen Wissenschaftsgebieten finden. Die Einbettung der Methodenwahl in Theorien unterscheidet insofern auch wissenschaftliche Datenarbeit von datenbasierten kommerziellen Recherchen und „Analysen".

unter anderem, ob und unter welchen Bedingungen wissenschaftliche Studien und Forschungsergebnisse nachvollziehbar, in einigen Bereichen auch wiederholbar beziehungsweise replizierbar sein müssen.[3] Auch in diesen Diskursen spielen das Vertrauen in die Wissenschaft, ihre Fähigkeit zur Selbstorganisation und geeignete politische Rahmensetzungen zur Erweiterung ihrer Leistungsfähigkeit eine wichtige Rolle. Der RfII sieht zwischen diesen Entwicklungen und dem bisher unterschätzten Thema der Datenqualität eine enge Verbindung.

Wachsendes gesellschaftliches Interesse an wissenschaftlich erzeugten Daten

Darüber hinaus ist das Interesse von Staaten, Regierungen und der Zivilgesellschaft an qualitätsgesicherten Forschungsdaten im globalen Maßstab gewachsen. Forschungsdaten rücken zum einen als Evidenzbasis auch für außerwissenschaftliche Entscheidungsprozesse in den Fokus der allgemeinen Aufmerksamkeit. Zum anderen gelten sie als „Rohstoff" für zunehmend schnellere Innovationszyklen. Bei allen Akteuren wächst so die Einsicht, wie sehr in den nächsten Jahrzehnten des Digitalzeitalters die Qualität von Daten und Datenprozessen mit der Qualität von Wissenschaft im Ganzen zusammenfallen wird. Digitalität gibt dem Thema Datenqualität eine neue Dimension, die rahmensetzendes Handeln in der Wissenschaftspolitik, aber auch eine Aufmerksamkeitssteigerung und Engagement in allen wissenschaftlichen Communities/Fachgemeinschaften und den Wissenschaftsorganisationen dringlich macht.

Güte digitaler Forschungsmethoden sichern

Handlungsbedarf ergibt sich – jenseits der qualitätsneutralen Rede einer „Nutzung" von Daten – insbesondere auch mit Blick auf die Güte digitaler Forschungsmethoden. Sie ist es, die entscheidet über

- die Validität und Anschlussfähigkeit der späteren Forschungsergebnisse in den wissenschaftlichen Fachgemeinschaften (disziplinär und interdisziplinär);
- das Gelingen des Transfers in die Anwendung in Wirtschaft und Gesellschaft – auch dort, wo kommerzielle Akteure inzwischen eigene Datensätze und Analyseverfahren anbieten sowie als Anbieter für Daten, Geräte und Analyseinstrumente für die Forschung an Gewicht gewinnen;
- das gesellschaftliche Vertrauen in den besonderen Wert und in die Nachhaltigkeit der wissenschaftlichen Wissensproduktion – auch im Vergleich zu esoterischen oder stärker interessengeleiteten Formen der Meinungsbildung, die nicht auf Daten beruhen und die nicht im Rahmen intersubjektiv überprüfbarer Standards und Verfahren gewonnen werden.

Chancen der Digitalität: Neue Forschungsfragen und -felder

Der schiere Umfang der heute potenziell zugänglichen Daten sowie die technischen Möglichkeiten, diese zu teilen und neu zu kombinieren, eröffnet der Wissenschaft die Chance, vollständig neue Forschungsfragen und Forschungsfelder zu bearbeiten. Steigende Quantität bedeutet in diesem Sinne auch höhere

[3] Vgl. hierzu detailliert Kapitel 3.1.

Vergleichbarkeit und Ausweitung des Optionenspielraums für den Einsatz wissenschaftlicher Methoden. Um mit diesen Möglichkeiten zielgerichtet umgehen zu können, sind Verständigungsprozesse erforderlich, die nicht nur disziplinäre Grenzen überschreiten, sondern auch tradierte institutionelle Barrieren überwinden. Gemeint sind Trennlinien zwischen dem Forschungsprozess einschließlich seiner Akteure im engeren Sinne und den forschungsermöglichenden Einrichtungen der Infrastruktur (zum Beispiel Bibliotheken, Sammlungen, Archive, Rechenzentren und Datenzentren) und ihrem Personal.[4] Digitalität überschreitet als Querschnittstechnologie die Grenze zwischen Forschung und Anwendung – zu beobachten ist dies etwa in der am Übergang von Laborforschung und Krankenbehandlung angesiedelten („translationalen") klinischen Forschung oder in der ingenieurwissenschaftlichen Simulationsforschung. Verknüpfungsanforderungen neuen Typs machen neue Qualitätsmaßstäbe erforderlich, sollen sie nicht einfach "irgendwie" bedient werden, sondern transparent, vergleichbar und in wissenschaftlicher Weise beherrschbar. Der Anspruch an Forschungsdaten, die über frühere Trennlinien hinweg auf qualitativ höchstem Niveau prozessiert, verstanden und weiterverarbeitet oder zeitlich nachgelagert wiederverwertet beziehungsweise nachgenutzt werden können, ist also denkbar hoch.

Schärfung des Qualitätsbewusstseins als Zukunftsaufgabe

Mit dem vorliegenden Positionspapier stellt der RfII die Frage, wie der mit neuen, digitalen Methoden gestiegene Qualitätsanspruch an heutige und künftige Daten eingelöst werden kann. Unter dem bewusst weit gefassten Stichwort „Neue Datenkultur" hat er bereits 2016 auf die wichtige Rolle von Informationsinfrastrukturen für die Qualitätssicherung von Daten hingewiesen.[5] Mit den hier vorgelegten Empfehlungen betont der RfII, dass Qualität für die Zukunft der Wissenschaft gerade angesichts der exponentiell steigenden Datenmengen im Digitalzeitalter in all ihren disziplinären Ausprägungen und Verknüpfungen essenziell ist. Dabei geht es keineswegs allein um ein „Weiter so". Denn: Neue Kulturen der offenen Zirkulation und des globalen Teilens von digitalen Daten haben nur bei *geschärftem* Qualitätsbewusstsein einen echten wissenschaftlichen Mehrwert.

Wissenschaftsweiter Qualitätsdiskurs nötig

Gefordert ist ein gemeinsamer Qualitätsdiskurs aller Akteure unter der Voraussetzung einer genauen Analyse der Veränderungen, die der digitale Wandel für methodische Forschung mit sich bringt, wie auch verbindlicher Qualitätssicherungsmaßnahmen in der Wissenschaft. Der RfII ist da von überzeugt, dass ein solcher Diskurs mittelfristig auch in konkretes Forschungshandeln, in die Praxis der Forschungseinrichtungen und Wissenschaftsorganisationen sowie in wissenschaftspolitische Rahmensetzungen einmünden wird. Handlungen und strategische Weichenstellungen also, die geeignet sein werden, das Qualitätsversprechen, das die Wissenschaft der Gesellschaft gibt, auch in Zukunft einzulösen und die digitale Transformation verantwortungsbewusst zu gestalten.

4 Vgl. hierzu auch RfII (2019) – Digitale Kompetenzen, S. 27 f., Empfehlung 4.5.

5 Vgl. RfII (2016) – Leistung aus Vielfalt, Kap. 4.6.

1 ZUR AUSGANGSLAGE

1.1 DIGITALER WANDEL UND DATENQUALITÄT – WAS HAT SICH VERÄNDERT?

Der wissenschaftliche Erkenntnisprozess ist darauf angelegt, auf Grundlage der jeweils bestmöglichen Datenbasis und angeleitet durch sich weiterentwickelnde Theorien und Methoden besonders qualifiziertes („wahres“[6]) Wissen hervorzubringen. Dieses Wissen zeichnet sich durch Validität gemäß methodologischer Maßstäbe sowie durch maximale Nachprüfbarkeit aus. In Form von Publikationen muss es sich in fachlichen Debatten behaupten und durchsetzen.

Digitale Forschung hat experimentierenden Charakter

Die Verwendung digitaler Techniken in Methoden, Organisation und Kommunikation von Forschung ändert dieses normative Leitbild der Wissensproduktion nicht grundlegend. Sehr wohl lassen sich aber in der sachlichen und zeitlichen Dimension Veränderungen in den Forschungsformen beobachten. Ebenso regen die digitale Transformation und die durch sie ermöglichten Verknüpfungen von Daten neue Fragestellungen an und lassen „alte“ Probleme in neuem Licht erscheinen. Insbesondere greift der Einsatz von auf digitaler Basis arbeitenden Medien und Werkzeugen sowie die Erhebung und Weiterverarbeitung bereits digital erzeugter Daten in viele Etappen des Forschungsprozesses sehr tiefgehend ein (vgl. auch 2.1). Vielfach hat der Einsatz neuer digitaler Werkzeuge in der Forschung experimentierenden Charakter und bringt Ergebnisse hervor, für deren Validierung Maßstäbe und Kriterien erst noch entwickelt werden müssen. Traditionelle Mechanismen der Vergewisserung von Qualitätsaspekten des Forschungshandelns gilt es heute in allen wissenschaftlichen Disziplinen neu zu reflektieren.

Sogenannte digitale Werkzeuge verändern Methoden

Digitale Werkzeuge (insbesondere *Computing*, also Modellierung und maschinelles Rechnen aber auch neue Datenerhebungs-, Darstellungs- und Analysemethoden in der Wissenschaft), haben in den vergangenen Jahren die Methodenbasis ganzer Forschungsfelder verändert. Zu denken ist hier an Beschleunigertechnik, Teleskopie, digitale Fernerkundung (Remote Sensing), Tomografie, Geophysik, Molekularbiologie sowie digitale Bild- und Textanalyse. Zahlreiche Felder in den Natur-, Lebens- und Technikwissenschaften haben sich auf eine Welt nahezu vollständig digitalisierter Forschungsgegenstände eingestellt. In anderen Forschungsfeldern haben sich zum Teil schon seit längerem digitale Teildisziplinen

[6] „Wahrheit“ steht hier in Anführungszeichen, denn es handelt sich um ein relatives Konzept: Im wissenschaftlichen Sinne „wahres“ Wissen schließt Verbesserbarkeit nicht aus; es wartet, im Gegenteil auf Falsifizierung durch jeweils neues „wahres“ Wissen (Wissenschaft produziert keine Dogmen). Allerdings muss neues Wissen, wiederum *wissenschaftlich* (gewonnen) sein, um für die Wissenschaft „besseres“ und also „wahrheitsfähiges“ Wissen zu sein. Wissenschaft lebt in dieser Hinsicht von selbstbezüglichen Kriterien. Anders gesagt: Weder politische Relevanz oder weltanschauliche Erwünschtheit, noch wirtschaftliche Bedeutung verleihen einem Forschungsergebnis seine „wissenschaftliche“ Qualität.

herausgebildet, wie zum Beispiel *Computational Physics*, *Computational Social Sciences*, *Digital Humanities* oder die Bio-, Geo- und Archäoinformatik.

Heterogenität und Verknüpfbarkeit von Daten als Herausforderung

Aber auch hier ergeben sich durch die schiere Menge der erzeugten Daten nicht nur neue Optionen. Herausfordernd sind hier die durch die Heterogenität der Daten und den raschen Wandel im Bereich von Programmiersprachen sowie Software entstehenden neuen Fragen, zum Beispiel sowohl bezüglich der Dokumentation und Provenienz von Daten sowie ihrer Qualität für die Nutzung in interdisziplinären Zusammenhängen als auch hinsichtlich der physischen Stabilität von Datenträgern über die Zeit. Für die Natur-, Lebens- und Technikwissenschaften wie für die Sozialwissenschaften ist die Einrichtung ihrer Expertensysteme auf die flexible Bereitstellung von Schnittstellen größtenteils Neuland. Hinzu kommen die wachsenden Anforderungen und Bedarfe nach einem Kontakt zur „Umwelt" des Wissenschaftssystems: Digital verfügbare wissenschaftliche Daten werden zunehmend auch für kommerzielle, politische und zivilgesellschaftliche Verwendungen interessant (umgekehrt bedient sich Wissenschaft ja auch außerwissenschaftlicher Daten).

Zugleich verschärft die umfassende Digitalisierung der Forschungsprozesse wissenschaftliche Qualitätsfragen, die vorher zwar zum Teil auch virulent waren, sich nun aber durch die exponentielle Vermehrung und Verfügbarkeit von Daten auf einem neuen Niveau stellen. Die folgende Tabelle 1 veranschaulicht die hierdurch geschaffenen Herausforderungen:

Tabelle 1: Durch Digitalität sich verschärfende allgemeine Herausforderungen für die Qualitätssicherung von Daten.

■ Unverhältnismäßig große Auswirkungen kleiner Unaufmerksamkeiten, Fehler und Versäumnisse
■ Entscheidungen über die Brauchbarkeit verrauschter Massendaten
■ Dekontextualisierte Nutzungen einzelner Datensequenzen
■ Unklare bzw. nicht erkennbare Provenienz von Daten (besonders bei durch Algorithmen generierten Angeboten und Selektionen)
■ Intransparente Rechenvorgänge
■ Fehlleitende Algorithmen (etwa aufgrund von Skalierungsproblemen)
■ Mangelbehaftete Trainingssets bei der Programmierung maschinellen Lernens (KI)
■ Erschwerte oder unmögliche Verifikation/Validierung des Gebrauchswertes übergroßer Datenmengen
■ Arbeitsteiligkeit entlang von nur noch schwach integrierten Prozessketten (sog. „Pipelines") bei der Arbeit mit Szenarien oder in der Simulation
■ Wachsende Abhängigkeit der Wissensarbeit von proprietärer Software
■ Fehlende Archivierbarkeit des digitalen Artefakts
■ Darstellungsprobleme für die Ergebnisdimension komplexer Rechnungen und Datenkondensierung (etwa durch „Visualisierung")
■ Datenschutz und andere rechtliche Probleme
■ Niedrigschwellige Manipulationsmöglichkeiten
■ Hacking und Cyber-Spionage
■ Gezielte Datensabotage
■ … und Weiteres mehr.

Quelle: eigene Darstellung.

Wissenschaftlicher Mehrwert durch interdisziplinären Datentransfer

Wo es heute um einen Mehrwert des Digitalen geht, gilt insbesondere der Transfer von Daten über disziplinäre Grenzen hinweg als wichtig. Das ist keine im engen Sinne technische Herausforderung, sondern entscheidend ist die innere Ordnung von Datensammlungen. Damit digitale Forschungsdaten in interdisziplinären Forschungsprozessen nachnutzbar sind, muss implizites Wissen expliziert werden, welches das Maß ihrer Nutzbarkeit in (möglichen) anderen Zusammenhängen bestimmt. Das heißt: Bedingungen, unter denen Daten entstanden sind, sowie ihr jeweils aktueller Zustand in einem Prozess von Nutzung und Veränderung werden für weitere Forschungen erkennbar ausgewiesen. Auch für jede Form der maschinellen Datenverarbeitung ist diese umfassende Leistung, die nachfolgend *Explikation* genannt wird, die notwendige Grundlage.

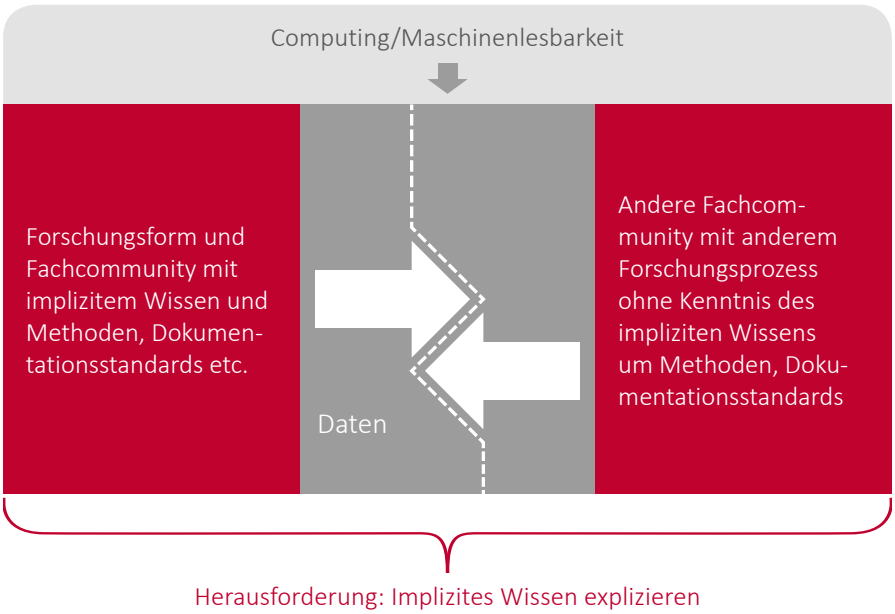

Abbildung 1: Digitalität steigert Explikationsbedarf von implizitem Wissen.
Quelle: eigene Darstellung.

„Methodischer" Erkenntnisgewinn beinhaltet die Notwendigkeit, implizites Wissen zu explizieren

Das Explizieren von dem, was Forschungsdaten an methodisch relevantem Wissen enthalten,[7] ist kein grundlegend neuer methodischer Imperativ in der Welt der Wissenschaft. Es gehört schon lange zur guten wissenschaftlichen Praxis, Inhaltsdaten zusammen mit den Informationen zu ihrer Entstehung zu speichern – Metadaten, die Inhalte zusammenfassen (Stichworte, Schlagworte) gehören hier dazu. Der Einsatz digitaler Verfahren erzeugt neue Bedarfe des Erklärens von Daten und ihrer Entstehungs- und Verwendungszusammenhänge. So sind beispielsweise das (unvermeidliche) „Rauschen" in großen Messdatenströmen und die (jeweils

[7] Der RfII verwendet hier (wie bisher auch) einen weit gefassten Forschungsdatenbegriff. Dieser umfasst sowohl analoge als auch digitale Sammlungen, denn die Verknüpfung von analogen und digitalen Daten spielt für wissenschaftliche Methoden vielfach eine wichtige Rolle. Vgl. hierzu die Begriffsklärung im Anhang A.2.

verbleibende) „Unschärfe“ digitaler Datenauswertungsverfahren auszuweisen. Ebenso müssen die Programme, Transformationsschritte und -parameter, an welche die Daten gebunden sind, expliziert werden – einschließlich der Informationen über Software-Versionen, Hardware-Generationen und gegebenenfalls Labor- oder Feldbedingungen, mit deren Hilfe und in deren Kontext die Daten entstehen. Die für heutige Forschung zwingend gebotene Notwendigkeit, Daten digital verknüpfbar auszugestalten, verlangt zudem mit Blick auf die dazu nötige „Maschinenlesbarkeit“ auch eine *stark standardisierte Explikation*. Anders als in der analogen Welt, wo im Umgang mit Forschungsdaten durch Menschen leicht improvisiert werden kann, braucht es hier maschinensprachlich fixierte Strukturen, die eine situativ unterschiedliche Handhabung möglichst ausschließen. Das Maß der damit verbundenen Festlegungen beziehungsweise die damit verbundenen Routinen können auf die Forschungspraxis zurückwirken.

Explikation ist voraussetzungsvoll

Wie voraussetzungsreich das geforderte Explizieren der Daten und ihrer Genese ist, zeigt der Umstand, dass das Spektrum wissenschaftlich erzeugter Daten sehr heterogen ist: von beliebig reproduzierbaren Datensets, zum Beispiel in der Genomanalyse, bis hin zu einmaligen, nicht reproduzierbaren Beobachtungsdaten, wie sie in der Astrophysik erzeugt werden. Auch im Bereich der Biodiversitätsforschung oder bei archäologischen Grabungen sind Probennahmen und Dokumentationsprozesse oftmals nicht wiederholbar. In solchen Fällen sind digitale Daten Unikate und eine ähnlich ersatzlose Referenz wie alles „Analoge“, sofern eine physische Bewahrung als Sammlungsgegenstand nicht möglich ist oder die physische Repräsentation des Gegenstands durch einen Verlust der Sammlung (Krieg, Naturkatastrophen) verloren geht. Oft sind es aber auch die hohen Investitionskosten, die eine erneute Erhebung von Forschungsdaten limitieren oder unmöglich machen. Die Forderung nach einer angemessenen Explikation von Forschungsdaten geht hier Hand in Hand mit dem Anspruch einer optimalen wissenschaftlichen Verwertung, denn die Nutzung (existierender) digitaler Ressourcen ist nicht nur zeitlich effizient, sie sichert im gegebenen Rahmen auch die Vergleichbarkeit von Forschungsprozessen und -ergebnissen.

Datenqualität nicht nur sichern, sondern steigern

Mit den angerissenen Themen drängt sich die Frage auf, ob und wie sich Qualität in digitalen Forschungsprozessen zum einen sichern, und wie sie sich zum zweiten steigern lässt, um damit zum Treiber einer ganz neuen Dynamik in der Wissenschaft und den auf wissenschaftliche Daten angewiesenen Innovationssystemen zu werden. Damit einher geht die Frage: Wie lässt sich Datenqualität für die fortgeschrittene digitale Arbeit in der Wissenschaft „organisieren“? Wer übernimmt welche Verantwortung und Aufgaben, wie lässt sich der experimentelle Charakter des Einsatzes digitaler Werkzeuge mit existierenden Qualitätsdiskursen verknüpfen? Und: Reichen die bislang eingeübten Verfahren wissenschaftlicher Qualitätskontrolle und der (Selbst-)Vergewisserung über Gütekriterien noch aus mit Blick auf die Steigerungsdynamiken, die eine forcierte Digitalisierung ermöglicht? Welche wären neu zu entwickeln?

Aus Sicht des RfII ist es unumgänglich, darüber zu diskutieren, dass im Spannungsfeld zwischen der Qualität von Daten und derjenigen von Methoden Neuerungen nötig sind. Ebenso sollte gefragt werden, wie Datenqualität terminologisch verbindlich beschrieben werden kann und welche Empfehlungen für ihre Generierung und Sicherung sich daraus ableiten lassen.

1.2 DATENQUALITÄTSKONZEPTE – HERANGEHENSWEISEN UND FORMEN DER AUSGESTALTUNG

Intrinsische Motivation – wissenschaftliches Ethos

Ein zentraler Orientierungspunkt für Datenqualitätskonzepte in der Wissenschaft ist die intrinsische Motivation der Forscherinnen und Forscher: „Wissenschaft als Beruf(ung)". Dieses professionelle Ethos schließt die persönliche Verantwortung für die Einhaltung intersubjektiv überprüfbarer Qualität von Forschung ein – einschließlich der Methoden und Verfahren zur Gewinnung eines Forschungsergebnisses. Ein heutiges Stichwort hierzu lautet wissenschaftliche Integrität. Sie erfordert es, die Regeln der „guten wissenschaftlichen Praxis" einzuhalten.[8] Darüber hinaus vertraut die Wissenschaft auf ihre eigene Methodenkultur. Was wissenschaftliche Gütekriterien im Gesamtsystem sind, wird überwiegend im Rahmen einer langen und gründlichen (Fach-)Sozialisation als Wissenschaftler an akademischen Einrichtungen gelehrt und gelernt. In diesem Zusammenhang bildet sich nicht zuletzt auch der Habitus der Forscherin beziehungsweise des Forschers heraus. Fachliche Expertise und Reputation wird der Forscherpersönlichkeit individuell zugeschrieben. Sie gründet vielfach auf internalisiertem „schweigendem Wissen" (*tacit knowledge*), das im Normalfall – einmal erworben und fachgerecht gehandhabt – kaum der weiteren Explikation bedarf.

Neue externe Anreize

Im 20. Jahrhundert tritt – getrieben von ehrgeizigen gesellschafts- und innovationspolitischen Zielen sowie der internationalen Konkurrenz von Staaten, Wirtschaftssystemen und Ideologien – neben die intrinsische Motivation die externe strukturelle Steuerung von Wissenschaft. Wissensproduktion wird nun aktiv, vorwiegend durch Anreize vorangetrieben und durch große Forschungsprogramme kanalisiert. Flankierend haben institutionelle Qualitätssicherungsmechanismen in die öffentlich finanzierte Wissenschaft Einzug gehalten. In diesem Zusammenhang wird nicht allein von einem am Wahrheitsbegriff orientierten und von Neugier (*curiosity*) getriebenen Erkenntnisprozess, von „Methoden", Verfahren der „Prüfung" und von „wissenschaftlichen Leistungen" gesprochen. Es tritt auch der Terminus „Qualität" hinzu – einschließlich der damit verknüpften Qualitätsmaße und Qualitätssicherungsmaßnahmen in Forschung *und* Lehre.

[8] Hierzu hat die DFG jüngst neue Leitlinien vorgelegt: DFG (2019) – Leitlinien zur Sicherung guter wissenschaftlicher Praxis.

Dies geschieht zunächst in Analogie zur Rede über Fertigungsprozesse beziehungsweise die Güte von Produkten und Verfahren (s. Anhang A.1, Kap. 2.1).

Ein kohärenter fächerübergreifender Diskurs über wissenschaftliche „Datenqualität" setzt erst mit der Zunahme digitaler Angebote seit den 1990er Jahren ein. Vor allem ist seither die *Sicherung* von Qualität ein Thema. Weniger hingegen wird das für den Wissenschaftsprozess besonders wichtige Thema der *Steigerung* einer Qualität von (digitalen) Daten oder die Rolle von Qualität in einem umfassenderen Transformationsprozess von Wissenschaft und Gesellschaft insgesamt behandelt.

Fünf Leitideen der Qualitätssteuerung für Forschungsdaten

Grob lassen sich in der Diskussion um die Qualität von Forschungsdaten fünf Konzepte beziehungsweise Leitideen zur (Selbst-)Steuerung des wissenschaftlichen Handelns unterscheiden, die im Folgenden näher ausgeführt werden:

1. Normsetzungen und Standardisierungen (zum Beispiel ISO-Normen, fachliche Standards), einschließlich der Normierung von Qualitätsmanagement: Der hierbei vorherrschende Steuerungsmodus ist primär *juridischer* Art;
2. Datenvalidierung und organisations- beziehungsweise prozessbezogene Operationalisierungen von Datenqualität (zum Beispiel durch Zertifikate oder Gütesiegel): Die hier vorherrschende Idee von Steuerung ist primär *organisatorischer, auf Anreize und Aufwertung setzender* Art;
3. Leitlinien und Policies als Grundregeln für den Umgang mit Forschungsdaten (inkl. zugehöriger Datenmanagementpläne): Hier soll die Internalisierung von Regeln eher durch *vertragsartige beziehungsweise verständigungsorientierte* Instrumente gelingen;
4. Idealtypische und schematische Beschreibungen der zu optimierenden Prozesse (zum Beispiel Datenlebenszyklen): Hier wird ein primär *verfahrenstechnisches* Ideal zur Herstellung von Datenqualität verfolgt;
5. Definition und Setzung pragmatischer Faustregeln (allen voran die Formel „fit for purpose") sowie Festlegung genereller Prinzipien (aktuell zum Beispiel die FAIR Data Principles): In diesem Modell wird vor allem auf eine *pragmatische* und primär *prozedurale* Steuerung von Qualitätsentwicklungen gesetzt.

Herkunft von Datenqualitätskonzepten – primär aus Wirtschaft und Verwaltung

Umfassende Datenqualitäts- und Datenqualitätsmanagementkonzepte entstammen bis heute primär dem wirtschaftlichen Bereich und sind allgemein im Bereich der öffentlichen Verwaltung – und hierbei auch in der öffentlich finanzierten Wissenschaft – adaptiert worden. Parallel hierzu hat es in der Forschung – insbesondere in der geräteintensiven Großforschung und beim Betrieb großer Datenbanken – eigene und überaus erfolgreiche Anstrengungen zur Entwicklung von Qualitätsstandards und qualitätssichernder Verfahren gegeben (vgl. 1.2.3). Domänenübergreifende „eigene", das heißt auf die Möglichkeiten und Herausforderungen von Digitalität für Forschungsdaten explizit zugeschnittene Qualitätsdiskurse hat die Wissenschaft jedoch bislang nur in Ansätzen hervorgebracht.

Ebenso ist es eine immer wieder offene Frage, welcher politischen Rahmensetzung die Wissenschaft bedarf, ohne dass die notwendige Autonomie ihrer Selbststeuerung, das heißt die intrinsische Motivation von Forscherinnen und Forschern zur Weiterentwicklung von Qualitätsstandards, gehemmt oder beschädigt wird.

1.2.1 NORMIERUNG UND DAS SETZEN VON STANDARDS

„De jure"- und „de facto"-Standards

In der Technologieentwicklung werden klassisch durch Normierung (Mindest-) Standards und damit Qualitätsmaße geschaffen: „Gut" ist, was dem Standard entspricht und damit in der Breite funktionsgerecht verwendbar ist. Der Bedarf an einer flächendeckenden, Kompatibilität und Qualität garantierenden Normierung entstammt ursprünglich der Welt der maschinellen Bauteile, frühe, koordinierte Normungsanstrengungen sind eine Errungenschaft des industriellen Maschinenbaus. Zügig durchgesetzt hat sich aber auch die in gleicher Weise durchgeführte Normierung für Prozesse – nämlich verbindlich ausgehandelte und niedergelegte unmissverständliche Definitionen und detaillierte Durchführungsvorschriften. Für die Datenqualität in der Wissenschaft sind beispielsweise Standardisierungen im Bereich der Kommunikations- und Informationstechnologie (Dateiformate und Datenträger, Datenübertragung, Webtechnologien, Schnittstellen) wie auch im Bereich der Dokumentation und Erschließung wissenschaftlicher Information (Vokabulare, Katalogisierung, Suchdienste) relevant. In beiden Bereichen kommen technische „de jure"-Standards, wie die DIN-Norm zur Anwendung, ebenso wie eine Vielzahl sogenannter „de facto"-Standards, die sich über Anwendungen und Akzeptanz verbreiten und durchsetzen.

Ordnungssysteme und Standards in der wissenschaftlichen Praxis

Normierung durch DIN und ISO

Die Wissenschaft bewegt sich in digitalen beziehungsweise digital unterstützten Forschungsprozessen nahezu überall innerhalb der Normen des *Deutschen Instituts für Normung*[9] (DIN-Normen) und international zum Beispiel derer der drei *europäischen Kommissionen für Normung* oder der *International Organization for Standardization* (ISO). Überwiegend sind dies Normierungen, die im Ursprung Regelungsbedarfen in der Industrie und im Dienstleistungssektor entsprechen.

Wissenschaftliche Standardisierungsorganisationen

In der wissenschaftlichen Dokumentation wie auch in den Fächern und Disziplinen finden sich darüber hinaus wissenschaftseigene, teils sehr starke Standardisierungsorganisationen, die in ähnlicher Weise die Setzung und Pflege

[9] Europäisches Komitee für Normung (CEN), Europäisches Komitee für elektrotechnische Normung (CENELEC), Europäisches Institut für Telekommunikationsnormen (ETSI).

von Standards betreiben – zum Beispiel durch kontrollierte Vokabulare (Thesauri), Referenzmodelle und Normdateien, Taxonomien und Nomenklaturen sowie weitere Systematiken (s. Anhang A.1, Kap. 2.2). Auch diese werden manchmal zur Anerkennung als DIN- oder ISO-Norm vorgeschlagen, teils aus pragmatischen Gründen, um den Prozess der Weiterentwicklung zu stabilisieren, teils wegen der erhofften höheren Schlagkraft und Akzeptanz „im System".

Metadatenstandards

Zu den prominentesten Vorgaben für interoperable Informationssysteme in den Wissenschaften zählen Normen und Standards für Metadaten im Bibliothekswesen. Die Entwicklung der Metadatenstandards beschleunigte sich in den 1990er Jahren, folgte aber häufig keiner international und über Domänengrenzen hinweg verbindlichen Leitlinie. Es gab keine Begrenzung für den Typ oder die Menge der Ressourcen, die durch Metadaten beschrieben werden sollten, wie auch keine Begrenzung für die Anzahl der sich überschneidenden Metadatenstandards für jede Art von Ressourcen beziehungsweise Subjektdomänen. Breit und fächerübergreifend akzeptiert ist heute zum Beispiel der Dublin Core Standard als Basis für die Beschreibung jeglicher Art von Dokumenten, sowie das Dateiformat MARC für den Austausch bibliografischer Daten zwischen verschiedenen Einrichtungen. Im deutschsprachigen Raum organisiert der Standardisierungsausschuss an der Deutschen Nationalbibliothek den Einsatz einheitlicher Standards für die Erschließung, Formate und Schnittstellen in Bibliotheken und entscheidet auf fachlicher Ebene über Grundsatzfragen.

Rahmen für digitale Forschungsprozesse speist sich aus unterschiedlichen Quellen

In den verschiedenen wissenschaftlichen Kulturen entsteht also – in unterschiedlicher Geschwindigkeit – ein allmählich wachsendes Rahmenwerk für digitale Forschungsprozesse, das teils auf gewachsenen Normierungsprozessen aufsetzt, teils durch Gedächtniseinrichtungen (wissenschaftliche Archive, Bibliotheken und Sammlungen) vorangetrieben wird (für eine Übersicht s. Anhang A.1, Kap. 2.2). Die involvierten Akteure schaffen eine Grundlage, um Daten zukünftig in bereits normierter Weise zu erstellen beziehungsweise in normierte Systeme zu migrieren, oder bieten auch Übersetzungsregeln zwischen gewachsenen Wissensorganisationssystemen an. So hat sich seit den 1990er Jahren eine Vielfalt entwickelt, die durchaus der Breite an methodischen Zugriffen, Gegenständen und Forschungsformen entspricht.

Umsetzung noch wenig kohärent

Demgegenüber fehlt es in der Forschungspraxis in vielen Bereichen an der Umsetzung, gerade was die Anwendung von Standards in der digitalen wissenschaftlichen Dokumentation betrifft. Punktuell mangelt es an Rückkopplungen zwischen den Akteuren (zum Beispiel zwischen den Standardisierungsausschüssen von Infrastrukturträgern, internationalen Expertenkomitees und den wissenschaftlichen Fachgemeinschaften) oder auch an hinreichend verbindlichen Entscheidungsprozessen. Akteure wie die global agierende Research Data Alliance (RDA) versuchen, dieses Defizit zu beheben, indem sie sich für verbindliche

Standards im Bereich des Forschungsdatenmanagements einsetzen.[10] Einfluss auf die wissenschaftliche Dokumentation haben darüber hinaus einschlägige Standards, die die Auffindbarkeit und Lesbarkeit von Informationen im World Wide Web ermöglichen.[11]

Standardisierung von „Datenqualität"

Orientierung am „Total Data Quality Management"

Orientierungspunkte lassen sich den Forschungen und der normativen Modellbildung zum allgemeinen Qualitätsmanagement aus der Managementtheorie und der Wirtschaftsinformatik entnehmen. Bis heute prägend sind hierbei die Arbeiten von Wang und Strong und Ansätze des Total Data Quality Management, die auch die spätere Ausarbeitung der FAIR-Prinzipien beeinflusst haben (s. 1.2.5). Datenqualität wird hier bedarfsorientiert, also aus der Perspektive der Datennutzung beziehungsweise des „Datenkonsums" heraus definiert, und entlang von vier Merkmalen unterschieden:[12]

- *inhaltsbezogene* Datenqualität (intrinsic): Daten besitzen eine Qualität aus sich heraus, indem sie beispielsweise fehlerfrei, glaubwürdig und objektiv sind.
- *kontextbezogene* Datenqualität (contextual): Qualität von Daten ergibt sich aus deren Eignung für einen kontextspezifischen Zweck, aber auch beispielsweise durch ihre Relevanz, Aktualität, ihren Mehrwert durch Verknüpfung.
- *darstellungsbezogene* Datenqualität (representational): Qualität von Daten entsteht, indem sie im Hinblick auf Darstellungsformate konzise und konsistent und im Hinblick auf ihre Bedeutung interpretierbar und leicht verständlich sind.
- *zugangsbezogene* Datenqualität (accessibility): Qualität gewinnen Daten, sofern sie zugänglich und bearbeitbar sind und der Zugang zu ihnen sicher ist und auch künftig sicher bleibt. Über den Lebenszyklus von Daten hinweg soll eine kontinuierliche Qualitätsdefinition, Qualitätsmessung und Qualitätsanalyse durchgeführt werden.[13]

Stärker operationalisiert finden sich solche Anforderungen in der 2009 erstmals veröffentlichten ISO-Norm 8000 „Datenqualität und Stammdatenqualität", die aus dem Bereich des eCommerce stammt, oder in der Norm „Measurement of

10 Die RDA wurde 2013 „bottom up" aus der Wissenschaft heraus als Expertennetzwerk gegründet und wird finanziell von zahlreichen staatlichen und staatsnahen Akteuren unterstützt. Sie verfolgt das Ziel, einen offenen Austausch und eine Wiederverwertung von Daten über Technologien, Disziplinen und Ländergrenzen hinweg zu ermöglichen.

11 Zum Beispiel die Beschreibungssprache xml (Extensible Markup Language) oder die Vokabulare von schema.org (https://schema.org/, zuletzt geprüft am: 30.08.2019). Zu den umfangreichen Standards und Tools vgl. Website des World Wide Web Consortium (W3C) – https://www.w3.org/standards/, zuletzt geprüft am: 30.08.2019.

12 Vgl. Wang/Strong (1996) – What Data Quality Means to Data Consumers, S. 9 und 18 f.

13 Wang (1998) – Total Data Quality Management.

Data Quality (ISO/IEC 25024), die Teil eines Normenpakets zur Softwarequalität ist. Letztere stellt vor allem auf die Qualitätsmerkmale Provenance, Accuracy und Completeness ab (s. a. Anhang A.1, Kap. 2.3).

Verbreitung in der Wissenschaft

In der Wissenschaft hat diese Normierung von Datenqualität nur bedingt Resonanz gefunden: Am ehesten finden sich hieraus abgeleitete Ansätze in der Qualitätssicherung von Kohortenstudien in der Medizin. Formale Prüfverfahren für Datenqualität finden sich traditionell auch in den Ingenieurwissenschaften. Auch im Kontext von Großforschungseinrichtungen wie dem CERN, in der Forschung mit Satellitendaten oder im Umfeld der großen Proteindatenbanken haben sich entsprechende Prüfverfahren entwickelt, die in die jeweiligen Communities/Fachgemeinschaften ausgestrahlt haben. In datenintensiv arbeitenden Wirtschaftsunternehmen haben sich Ansätze für eine explizite *Data Governance* entwickelt, um Verantwortlichkeiten und formalisierte Prozesse für den Umgang mit Daten festzulegen. Solche Konzepte sind vor allem dort zu finden, wo Daten die Geschäftsgrundlage bilden, wie zum Beispiel in der Finanz- und Kreditwirtschaft. In der Wissenschaft haben sich im Rahmen großer Längsschnittstudien formale Verfahren für das Datenmanagement herausgebildet, durch die unter anderem umfangreiche Datenschutzauflagen erfüllt werden. Eine andere Spielart von *Data Governance* findet sich zum Beispiel unter dem Schlagwort *Good Clinical Data Management* in der medizinischen Forschung (vgl. Anhang A.1, Kap. 2.3).

Zusammengenommen geben diese Ansätze eine Richtung vor, in der sich spätere prozedurale Regelwerke aus der Wissenschaft heraus und mit unmittelbarem wissenschaftlichem Bezug – wie zum Beispiel die FAIR-Prinzipien – weiterbewegen könnten.

1.2.2 VALIDIERUNG UND ZERTIFIZIERUNG

Leistungsziel von Zertifizierung: Vertrauen in Prozessqualität

Zertifizierungsverfahren richten sich auf die organisatorische oder institutionelle Umsetzung von Qualitätsnormen oder aber Standards. Sie schreiben Verantwortung fest. Allgemein handelt es sich um Konformitätsbewertungen, die zum einen Vertrauen in die Prozessqualität der Erzeugung, Verarbeitung und Speicherung von Daten schaffen. Zum anderen bieten Zertifikate – zum Beispiel in Form von Gütesiegeln – eine Orientierungshilfe beim Zugang zu Repositorien und erhöhen die Bereitschaft Forschender zur Datenübergabe. Gleiches gilt auch für Daten von nicht wissenschaftlichen Einrichtungen, wie zum Beispiel statistischen Ämtern oder Sozialversicherungsträgern, die ihre Daten für die wissenschaftliche Nutzung zur Verfügung stellen. Verfügen sie über eine Zertifizierung oder alternativ eine Akkreditierung als Forschungsdatenzentrum, schafft dies einen deutlichen Vertrauensvorschuss bezüglich der Validität der angebotenen Daten für wissenschaftliche Nutzungszwecke.

Für den Bereich digitaler Informationsinfrastrukturen sind als Beispiele für erfolgreiche, sichtbare und allgemein akzeptierte Zertifizierungen das internationale Core Trust Seal für „vertrauenswürdige Repositorien“ (unter dem Dach der Research Data Alliance) oder das Akkreditierungsverfahren für Forschungsdatenzentren beim deutschen Rat für Sozial- und Wirtschaftsdaten (RatSWD) zu nennen.

Beispiel: Core Trust Seal

Das Core Trust Seal (CTS) ist ein peer-review-gestütztes Selbstevaluierungsverfahren, in dessen Rahmen Einrichtungen ihre Konzepte und Leitlinien für die Datenarchivierung entlang eines 16-Punkte-Katalogs bewerten (vgl. Anhang A.1, Kap. 2.4). Zum Erwerb des Siegels muss ein Betreiber unter anderem darlegen:

- welche Maßnahmen zur Sicherung von Integrität und Authentizität der Daten sowie zum langfristigen Erhalt der Interpretierbarkeit der Daten ergriffen werden;
- welche Metadatenstandards eingesetzt werden (differenziert nach beschreibenden, strukturellen und technischen Metadaten) und
- in welchem Umfang Daten durch das Datenarchiv kuratiert werden.

Damit besteht ein klarer Bezug zur Qualität der Inhaltsdaten und der Metadaten, auch wenn Datenqualität nicht direkt Gegenstand der Zertifizierung ist.

Beispiel: Akkreditierung durch den RatSWD

Die Akkreditierung sozial- und wirtschaftswissenschaftlicher Datenzentren durch den RatSWD stellt mehr auf die Zugänglichkeit der Ressourcen ab, wobei auch hier erhoben wird, ob die „Datenprüfung (auf Qualität und Güte der weitergegebenen Daten)“ zur Aufgabe des Forschungsdatenzentrums gehört und welche Verfahren gegebenenfalls eingesetzt werden.[14] Dies ist insofern ein Meilenstein in der Steigerung von Datenqualität in den Wirtschafts- und Sozialwissenschaften, als sich „Zugänglichkeit“ und wissenschaftliche „(Nach-) Nutzbarkeit“ zunächst primär auf Daten bezog, die außerhalb der Wissenschaft und abseits wissenschaftlicher Zweckbestimmungen zum Beispiel von statistischen Ämtern, Sozialversicherungsträgern und öffentlichen Einrichtungen zur Regulierung des Arbeitsmarktes erhoben wurden. Langfristig hat die Akkreditierung durch den RatSWD zu vergleichbaren Qualitätsstandards und dadurch erleichtertem Transfer zwischen genuin in der Wissenschaft und außerwissenschaftlich erhobenen Daten und Datensätzen geführt. Einige der akkreditierten Forschungsdatenzentren haben sich zusätzlich als vertrauenswürdige Repositorien zertifizieren lassen.

[14] RatSWD (2018) – Tätigkeitsbericht der Forschungsdatenzentren 2017, S. 16.

Tabelle 2: Zertifikate von Datenrepositorien, geordnet nach Häufigkeit.

	Name des Zertifikats	Häufigkeit [15]	Anbieter
1.	**Core Trust Seal (CTS)** seit 2017	62	Core Trust Seal Board, Zusammenlegung von WDS und DSA unter dem Dach der Research Data Alliance (international)
2.	**World Data System Certificate** Vergabe bis 2017, jetzt Core Trust Seal (s. Nr. 1)	55	ICSU World Data System (international)
3.	**RatSWD Akkreditierung**	32	Rat für Sozial- und Wirtschaftsdaten (DE)
4.	**Data Seal of Approval (DSA)** Vergabe bis 2017, jetzt Core Trust Seal (s. Nr. 1)	31	DANS (NL) bzw. Data Seal of Approval Board & General Assembly (international)
5.	**CLARIN Certificate**	27	CLARIN ERIC (EU), eine Forschungsinfrastruktur im ESFRI-Programm
6.	**DINI Zertifikat „Open-Access-Repositorien und Publikationsdienste“**	6	DINI – Deutsche Initiative Netzwerkinformation (DE)
7.	**Nestor-Siegel** DIN 31644	1	Nestor Kompetenznetzwerk Langzeitarchivierung (DE)
8.	**Trustworthy Repositories Audit & Certification (TRAC)** ISO 16363	1	Ursprünglich Consultative Committee for Space Data Systems, aktuell: ISO/TC 20/SC 13 Space data and information transfer systems (technical committee)

Quelle: eigene Darstellung basierend auf einer Auswertung der Datenbank unter re3data.org, Stand 08.08.2019.

Weltweit erst wenige zertifizierte bzw. akkreditierte Infrastrukturen

Die Zahl der in vergleichbarer Weise zertifizierten oder akkreditierten Forschungs- und Informationsinfrastrukturen ist weltweit insgesamt bislang überschaubar: Von über 2300 registrierten Datenrepositorien in der Datenbank re3data.org ist nur ein Bruchteil zertifiziert (s. Tabelle 2). Dies spiegelt den noch geringen Institutionalisierungs- und Professionalisierungsgrad wider, der vom RfII bereits 2016 diagnostiziert wurde.[16] Den meist projektfinanzierten Diensten fehlt es schlicht an Personal, um die Prozesse zu etablieren und zu beschreiben, die für eine Zertifizierung der Qualitätssicherungsmaßnahmen erforderlich sind. Hinzu kommt, dass beispielsweise eine CTS-Zertifizierung die dauerhafte institutionelle Verantwortungsübernahme einer Universität oder Universitätsbibliothek für ein Repositorium voraussetzt.

Geht es um Konformität der Daten selbst, so existieren eine Reihe pragmatischer Ansätze beziehungsweise Werkzeuge für (automatisiert durchführbare) Kompatibilitätstests (vgl. Anhang A.1, Kap. 2.4). Ihr Zweck ist es, Nutzerinnen

[15] Datenrepositorien können mehrere Siegel haben, daher ergibt die Summe der einzelnen Zeilen nicht die Gesamtzahl zertifizierter Repositorien.

[16] Vgl. RfII (2016) – Leistung aus Vielfalt, Kap. 2.5.

und Nutzern von Diensten schnell und unkompliziert eine Selbstprüfung zu ermöglichen, zum Beispiel beim Hochladen von Daten. Die automatisierte Validierung von Daten und Software wird punktuell auch in der wissenschaftlichen Qualitätssicherung eingesetzt, zum Beispiel bei Datenarchiven oder der Begutachtung von Veröffentlichungen.

1.2.3 FORSCHUNGSDATEN-POLICIES UND DATENMANAGEMENTPLÄNE

Grundregeln für den Umgang mit Forschungsdaten

Die konkrete Arbeitsebene in der Forschungspraxis wird durch Normsetzungen, Definitionen von Standards oder Zertifikate häufig nicht erreicht. Hier greifen eher Maßnahmen des Forschungsdatenmanagements (FDM), etwa sogenannte Datenmanagementpläne als operationalisierbare Weiterentwicklungen von Leitlinien und Maßgaben (Policies), die als projekt- oder einrichtungsspezifische Grundregeln für den Umgang mit Forschungsdaten dienen.

Für den Bereich derartiger Selbstverpflichtungen und Leitlinien liegen die Anfänge in den 1990er Jahren. Neben eigenen Selbstverpflichtungen international agierender Forschungskonsortien (zum Beispiel das Humane Genomprojekt) fanden entsprechende Vorgaben auch Eingang in die Anforderungskataloge für Drittmittelanträge von Forschungsförderorganisationen oder auch – ausgehend vom angelsächsischen Raum – in die Universitäten (s. Anhang A.1, Kap. 2.5). Auch auf der staatlichen Ebene gibt es heute Initiativen, um den Umgang mit Forschungsdaten zu regeln oder zu verbessern, häufig in Verbindung mit e-Science, Digitalisierungs- oder Open-Access-Strategien.

Open Access bringt FDM-Policies auf den Weg

Die Zunahme von FDM-Policies in den vergangenen Jahren, hängt zum einen damit zusammen, dass die nationalen Forschungsförderer relativ früh das Open-Access-/Open-Science-Paradigma aufgegriffen haben.[17] Die Vorlage einer Forschungsdaten-Policy oder von Datenmanagementplänen wird schrittweise zur Pflicht gemacht, so zum Beispiel in Projekten, die im Rahmen des europäischen Forschungsrahmenprogramms Horizont 2020 gefördert werden sollen, teils auch in Förderprogrammen der DFG und des BMBF. Dies hat dazu geführt, dass die Bedeutung von Vorgaben für das Forschungsdatenmanagement und hieraus abgeleiteter Datenmanagementpläne zumindest im Bereich datenintensiver und in hohem Maße mit Drittmitteln forschender Forschungsinstitute und Universitätseinrichtungen stark zugenommen hat. Für die organisationsinterne Umsetzung von Policies und Plänen werden überdies vielfach umfangreiche

[17] Zu Open Science als Treiber für die Entstehung nationaler und subnationaler Regelungen zu Forschungsdaten und Informationsinfrastrukturen vgl. RfII (2017) – Fachbericht Länderanalysen, S. 9 f. In Europa haben zum Beispiel der Schweizerische Nationalfonds und der Norwegische Forschungsrat Anforderungen an Forschungsdatenmanagementpläne obligatorisch gemacht.

Handreichungen, Checklisten und auch digitale Tools bereitgestellt sowie Anlaufstellen für die FDM-Beratung eingerichtet.

Schwierigkeiten in der operativen Umsetzung

Die konkrete Umsetzung der Datenmanagementpläne und anderer Maßgaben für das Forschungsdatenmanagement – zum Beispiel die Zugänglichmachung beziehungsweise Veröffentlichung der Daten – liegt in der Regel beim jeweiligen Forschungsprojekt und hier häufig in den Händen des wissenschaftlichen Nachwuchses beziehungsweise des Drittmittelpersonals, dass sich in sehr kurzer Zeit und nebenher mit dem Forschungsdatenmanagement vertraut machen muss. Die Implementierung der dort festgelegten Regeln erscheint – abgesehen von Beispielen für fachspezifische Leitlinien und Selbstverpflichtungen – primär als (förder-)politisch getrieben und extern auferlegte zusätzliche Last eines der wissenschaftlichen Arbeit nicht zuträglichen Wissenschaftsmanagements. In der intrinsischen Motivation vieler Forscherinnen und Forscher finden FDM-Policies, die für die Projektdurchführung konkrete FDM-Pflichten enthalten, also bislang wenig Widerhall.

Unterschiedliche Regelungstiefen

An den bislang veröffentlichten FDM-Leitlinien fällt hinsichtlich der Gegenstände eine große Unterschiedlichkeit in der Regelungstiefe auf. So definieren nicht alle Maßgaben die grundlegenden Begriffe, klären Fragen des Besitzes der Daten oder äußern sich zu Kosten. Genauer sind hingegen die Aussagen zu den Datenmanagementplänen, zu Open Access und zu ethischen Fragestellungen. Zunehmend kommen Regelungen hinzu, Primärdaten nach einer Embargozeit für die Öffentlichkeit zugänglich zu machen, was das Open-Access-Paradigma modifiziert. Vertreter des Open-Access-Gedankens bemängeln, dass Forschungsdaten-Leitlinien nur die Offenlegung der Daten betreffen, nicht aber einen fairen Umgang mit den Daten nach der Veröffentlichung. Die Angst vor einer unredlichen Nutzung durch Konkurrenten (*data parasitism*) hindere viele Forschende daran, ihre Daten offenzulegen.[18]

1.2.4 OPTIMIERUNG VON PROZESSKETTEN: DER DATENLEBENSZYKLUS

Auf die Einsicht, dass Forschung auf Forschung aufbaut, dass Wissenschaft also auf selbstgeschaffenes Wissen permanent zurückgreift sowie dieses dadurch auch erneuert, haben die Wissenschaftstheorie und die Informationswissenschaft mit Kreislaufmodellen reagiert. Um den Umgang mit Forschungsdaten – auch unter Qualitätsaspekten – darzustellen, wurden ab Mitte der 2000er Jahre Modelle sogenannter Datenlebenszyklen entwickelt (zu Qualitätsherausforderungen entlang des Datenlebenszyklus siehe detailliert Kapitel 2).

[18] Zum Beispiel Amann et al. (2019) – Toward Unrestricted Use of Data.

Pluralität von Konzepten und Schwerpunkten

Der Datenlebenszyklus ist ein Konzept, das den zyklischen Charakter der Arbeit mit Daten aller Art (einschließlich Informationen) in ihren verschiedenen Stadien im Prozess der wissenschaftlichen Bearbeitung und Nutzung beschreibt. Als wesentliche Schritte dieses Zyklus gelten die Datengenerierung (zum Beispiel Messungen), die Datenaufbereitung, die Datenauswertung/-analyse, die Speicherung bis hin zur Langzeitarchivierung sowie die Verfügbarmachung durch Veröffentlichung bis hin zur Nachnutzung in weiteren oder neuen Forschungskontexten, die sich auch durch die Lehre ergeben können.[19] Es haben sich zahlreiche Varianten von Datenlebenszyklus-Modellen herausgebildet, die sich hinsichtlich Detaillierungsgrad, Fachspezifik oder operativer Zielstellung unterscheiden (vgl. Anhang A.1, Kap. 2.6). Entsprechend werden die Schwerpunkte unterschiedlich gesetzt: auf langfristige Archivierung, die Phase der Datensammlung und -auswertung bis hin zu Nachnutzungsszenarien oder auch der Verdeutlichung von Zuständigkeiten professioneller Akteure und die Arbeitsteilung zwischen ihnen („data creator, data scientist, data manager and data librarian“).[20]

Gebrauchswert der Zyklusmodelle

Datenlebenszyklus-Modelle veranschaulichen, dass durch die Datennutzung und Nachnutzung jeweils neue Ergebnisse in Form von Forschungsdaten generiert werden. Das Datenmanagement entlang eines Lebenszyklus muss entsprechend dafür sorgen, dass Forschungsergebnisse über alle Stadien hinweg reproduzierbar sind. Des Weiteren sind an mehreren Schnittstellen des Modells Entscheidungen darüber zu treffen, welche Daten aufbewahrt, als Datensatz eigenständig publiziert oder in eine Publikation eingehen werden, und wie lange sie verfügbar zu halten sind (s. Kap. 2). Diese Entscheidungen werden aktuell zum Beispiel von Forscherteams oder Einzelforschern, welche die von ihnen selbst generierten Daten verwalten, nach unterschiedlichen Maßstäben getroffen.

Operationalisierung des Zyklus im Kontext von Forschungsformen

Eine der jeweiligen Forschungsform[21] gemäße Operationalisierung nach zumindest groben Relevanzkriterien durch die Fachgemeinschaften wäre in diesem Zusammenhang hilfreich, liegt aber bislang kaum vor. Beispielsweise nutzen hermeneutisch-interpretierende Forschungsformen bis heute und sicher auch künftig in großem Umfang nicht-digitale Medien wie Schriften, Bilder oder auch natürliche Objekte als Quellen der Erkenntnisproduktion. Deren Behandlung im Datenlebenszyklus erfordert deutlich andere Anstrengungen als der Umgang mit Messdaten aus experimentierenden Forschungsformen oder Umfragedaten, die im Rahmen beobachtender Forschungsformen entstehen. Aber auch in den Natur- und Ingenieurwissenschaften spielen Formen der Verschränkung von digitalen Messdaten und physische Sammlungen (Gewebeproben, Bohrkerne etc.) eine Rolle.

---

19 Siehe hierzu RfII (2016) – Leistung aus Vielfalt, Anhang A, S. A-7.

20 Swan/Brown (2008) – Skills, Role and Career Structure of Data Scientists, S. 1.

21 Der Wissenschaftsrat unterscheidet insgesamt sechs Forschungsformen. Neben den hier genannten zählen dazu noch Simulationen, begrifflich-theoretische und gestaltende Forschungsformen. Vgl. WR (2012) – Empfehlungen zu Informationsinfrastrukturen, S. 35–38.

1.2.5 DER NUTZUNGSZWECK UND DIE ORIENTIERUNG AN PRINZIPIEN

Faustformel des „fit for purpose“

Eine pragmatische Kurzdefinition für Datenqualität ist die Formel „fit for purpose“, also eine Orientierung am Zweck oder der Absicht der Verwendung. Dieser Leitgedanke entstammt ursprünglich der Qualitätssicherung in industriellen Fertigungsprozessen und konstituierte zumindest im angelsächsischen Raum auch einen rechtlich verbrieften Anspruch des Kunden auf die Nutzbarkeit eines spezifischen Produkts, das von einem Hersteller erworben wird (vgl. Anhang A.1, Kap. 2.7).

Adaption von „fit for purpose“ in der Wissenschaft

Die Formel „fit for purpose“ oder „fit for use“ ist auch in vielen Feldern der Wissenschaft in Gebrauch. Datenqualität wird auf diese Weise umfassend, aber bestimmungsoffen definiert als Gesamtheit von Eigenschaften und Merkmalen von Daten hinsichtlich ihrer Eignung, einen bestimmten Zweck zu erfüllen. Attraktiv an der pauschalen Orientierung am „Zweck“ (bzw. der Absicht des Verwenders) ist, dass der Kontext (guter) Wissenschaft oder auch die Frage nach Methoden, Standards etc. zwar mitgedacht werden kann, aber nicht näher spezifiziert werden muss. Die Kurzformel, Datenqualität ergebe sich durch Aufbereitung für eine jeweilige Zweckbestimmung, suggeriert zum einen, das Konzept lasse sich leicht und konkret stets passend operationalisieren. Zum anderen ist es als relationales Konzept tatsächlich maximal flexibel, da die „fitness for purpose“ durch die Nutzung und ganz unbestimmte Nutzerbedarfe entsteht, also nahezu beliebig variieren kann.

Die FAIR-Prinzipien und ihre Operationalisierung

Der Gedanke einer Nutzungsorientierung bei zugleich offen gehaltenen Zwecken liegt auch Programmen zugrunde, die pragmatische Prinzipien für den Umgang mit Forschungsdaten im Sinne einer Selbstverpflichtung seitens wissenschaftlicher Einrichtungen, aber auch von Forschungsprojekten fordern. Ein prominentes Beispiel dafür sind die 2014 erarbeiteten FAIR Data Principles.

Die im Akronym zusammengefassten vier Prinzipien von FAIR (Findable, Accessible, Interoperable, Reusable) geben pragmatische Grundsätze an, die nachhaltig nachnutzbare Forschungsdaten erfüllen müssen.[22] Sie haben insbesondere die Verbesserung der Maschinenlesbarkeit zum Ziel.[23] Vom Paradigma der „Offenheit“ in der gängigen Definition[24] werden die Prinzipien klar abgegrenzt: Die FAIR-Prinzipien lassen sich ebenso auf Daten anwenden, die aus rechtlichen oder anderen notwendigen Gründen zugangsbeschränkt sind.[25]

---

[22] https://www.force11.org/group/fairgroup/fairprinciples (zuletzt geprüft am: 30.08.2019).

[23] Vgl. Wilkinson et al. (2016) – The FAIR Guiding Principles.

[24] „Open data and content can be freely used, modified, and shared by anyone for any purpose“, https://opendefinition.org/ (zuletzt geprüft am: 30.08.2019).

[25] Hodson et al. (2018) – Fair Data Action Plan. Interim Recommendations, S. 15 f.

Die Operationalisierung der FAIR-Prinzipien durch verschiedene Infrastruktur-Akteure hebt in hohem Maße auf die Auffindbarkeit von Forschungsdaten ab und auf das Ziel, Nutzung zu ermöglichen – also auf die „quality of services“ (, welche die Daten liefern). Allein hierin wird bereits ein deutlicher Gewinn für die Wissenschaft gesehen: FAIRe Daten werden über Disziplinen- und Domänengrenzen hinweg deutlich leichter auffindbar sein als in der Vergangenheit, so der Ansatz. Über die wissenschaftliche Güte dieser Daten (bezogen auf Methoden und mit ihnen erzeugte Forschungsergebnisse) ist dabei allerdings generell noch wenig ausgesagt. Zwar sollen Daten gemäß den FAIR-Prinzipien „community-spezifische Standards“ erfüllen, um „reusable“ – nachnutzbar zu sein (Kriterium R.1.3, vgl. Tabelle 3). Hinter dieser Anforderung stehen im Einzelfall jedoch komplexe Konglomerate offener Fragen der Forschungspraxis in den jeweiligen Communities, die Verständigungsprozesse erfordern und die durch Trainings allein nicht beantwortet werden können (vgl. dazu auch 1.2.1). Wie das Kriterium „reusable“ konkret eingefordert und umgesetzt werden könnte, bleibt offen.

Tabelle 3: FAIR Data Principles 2016.

TO BE FINDABLE:	
F 1	(meta)data are assigned a globally unique and eternally persistent identifier.
F 2	data are described with rich metadata.
F 3	(meta)data are registered or indesed in a searchable resource.
F 4	metadata specify the data identifier.
TO BE ACCESSIBLE:	
A 1	(meta)data are retrievable by their identifier using a standardized communications protocol.
A 1.1	the protocol ist open, free and universally implementable.
A 1.2	the protocol allows for an authentication and authorization procedure, where necessary.
A 2	metadata are accessible, even when the data are no longer available.
TO BE INTEROPERABLE:	
I 1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I 2	(meta)data use vocabularies that follow FAIR principles.
I 3	(meta)data include qualified references to other (meta)data.
TO BE RE-USABLE:	
R 1	meta(data) have a plurality of accurate and relevant attributes.
R 1.1	(meta)data are released with a clear and accessible data usage license.
R 1.2	(meta)data are associated with their provenance.
R 1.3	(meta)data meet domain-relevant standards.

Quelle: Force 11.[26]

[26] https://www.force11.org/group/fairgroup/fairprinciples (zuletzt geprüft am: 30.08.2019)

Verbreitung von FAIR

Die FAIR-Prinzipien fordern (und konstituieren) für sich genommen noch keine wissenschaftliche Datenqualität. Sie bieten allerdings ein Set an grundlegenden Kriterien, die geeignet sind, einen Prozess mit der akzeptierten Zielrichtung einer breiten Datennutzbarkeit und Datennutzung zu leiten. Zu diesem Schluss kommt auch eine von der Europäischen Kommission eingerichtete Expertenkommission. Sie weist darauf hin, dass mit der Implementierung der FAIR-Prinzipien weitere Anforderungen verbunden werden können.[27]

Aufgrund ihrer Eingängigkeit und vielleicht auch der Beschränkung auf Nutzbarkeitsvoraussetzungen sind die FAIR-Prinzipien sehr zügig von der Forschungspolitik adaptiert worden. In der *EOSC Implementation Roadmap* von 2018 spielt die Implementierung der FAIR-Prinzipien beispielsweise eine zentrale Rolle, und auch in Deutschland ist für die Nationale Forschungsdateninfrastruktur (NFDI) die Einhaltung der FAIR-Prinzipien bereits als Ziel postuliert worden.

1.3 ZWISCHEN TOP DOWN UND BOTTOM UP: DIE SCHWIERIGE SUCHE NACH WISSENSCHAFTSADÄQUATEN KONZEPTEN FÜR DATENQUALITÄT

Wissenschaftsbezogene Anforderungen an Datenqualität

In der Wissenschaft müssen – anders als in der datenintensiven Wirtschaft – die Anforderungen an Datenqualität in ein Verhältnis zur Leistungsfähigkeit von fachwissenschaftlichen Methoden gesetzt werden, darüber hinaus zu bereits vollzogener Forschung und zu den Möglichkeiten künftiger, heute noch unbekannter, Forschung. Dies bedeutet, dass Daten in der Wissenschaft sowohl in der Zeit- als auch in der Sachdimension höchst komplexen Anforderungen genügen müssen: Hinsichtlich ihres Explikationspotenzials bieten sie idealiter sowohl Anschlussmöglichkeiten für grundlegend Neues als auch den stetigen Rückbezug zur bisherigen Forschung (als Vergleichsmaß für den Erkenntnisfortschritt). „Standards" für das Qualitätsmanagement digitaler Forschungsdaten zu vereinbaren, setzt aufgrund dieser äußerst anspruchsvollen Anforderungen im Wissenschaftssystem ein komplexes Zusammenspiel von Bottom-up-Initiativen aus der Forschung selbst und Top-down-Beratungen voraus – unter Einbezug der Akteure, die die Governance der Wissenschaft und damit deren operative Rahmenbedingungen organisieren.

[27] „FAIR is not limited to its four constituent elements: it must also comprise appropriate openness, the assessability of data, longterm stewardship, and other relevant features." Hodson et al. (2018) – Fair Data Action Plan. Interim Recommendations, S. 3.

Herausforderungen für Begriffsbestimmung und Steuerung

Regulierungsformen (Policies) für die Qualität von Daten, die der Wissenschaft gerecht werden, umfassen insofern ein breites Spektrum aus wissenschaftspolitischen Vorgaben, Bestimmungen der Forschungsförderung, Leitlinien einzelner Einrichtungen sowie Normsetzungen in spezifischen Fachgemeinschaften beziehungsweise für bestimmte feldspezifische Gegenstandsbereiche.

Festzuhalten ist insgesamt, dass Datenqualität in der Wissenschaft nicht nur begrifflich schwer zu bestimmen ist, sondern sowohl in selbstorganisierter Form als auch über externe (politische) Rahmensetzung sehr schwer gesteuert werden kann. Nichtsdestotrotz werden heute von der Wissenschaft und für die Wissenschaft Qualitätsanforderungen für digitale Daten forschungspolitisch gesetzt beziehungsweise gefordert. Der expliziten Normierung von Datenqualität (DIN/ISO) kommt dabei eine geringere Bedeutung zu. Sie leistet in der komplexen Welt digitaler Forschung zu wenig und ist als eher hierarchisches Regulierungsregime der dezentralen und dynamischen Struktur von Forschungsprozessen kaum angemessen. Pragmatische Maßstäbe wie „fit for purpose“ können dagegen übermäßig flexibel sein. Eine wissenschaftspolitisch eingeforderte Vereinheitlichung (Standardisierung) von Dateneigenschaften durch die Etablierung kaum abdingbarer Prinzipien wie FAIR stellt, was die Verbindlichkeit angeht, einen Mittelweg dar. Aspekte der Normierung werden mit einer Pragmatik („fit for purpose“) verbunden, die vor allem die (technische) Nutzbarkeit in den Mittelpunkt rückt.

Über FAIRe Daten hinausdenken

FAIR hebt unter Qualitätsaspekten allerdings vor allem auf die Gesichtspunkte der Maschinenlesbarkeit und – damit eng verknüpft – der Auffindbarkeit von Forschungsdaten ab. Servicequalität für Daten, nicht aber wissenschaftliche Güte ist das Ziel, das hierbei im Vordergrund steht. Zur Steigerung der wissenschaftlichen Güte von Forschungsdaten, das heißt der Erschließung der in den Daten liegenden Möglichkeiten für künftige Nutzung und Rekombination in der Forschung sollte daher bereits heute über FAIR hinaus weitergedacht werden (s. Kap. 4).

2 HERAUSFORDERUNGEN FÜR DIE QUALITÄT VON DATEN – AUS DER PRAXISPERSPEKTIVE

2.1 IDEAL UND WIRKLICHKEIT: DATENQUALITÄTSPROBLEME IM FORSCHUNGSPROZESS

Dass sich unter digitalen Bedingungen eine Fülle von zum Teil neuartigen Qualitätsproblemen stellt, zeigen nicht nur die Anstrengungen, zu operationalisierbaren Standards für das Forschungsdatenmanagement zu kommen (vgl. Kap. 1). Es lässt sich auch sehr viel konkreter entlang des Datenlebenszyklus nachzeichnen. Damit ist die Frage gestellt, wie „unterhalb" von Qualitäts- und Qualitätssteuerungsmodellen die Wirklichkeit von Datenqualitätsproblemen in der Wissenschaft faktisch aussieht.

Datentransformation als dynamischer Prozess

Der Datenlebenszyklus beschreibt modellhaft mit den Arbeitsschritten im Forschungsprozess jeweils einhergehende Phasen des Datenmanagements – idealtypisch: von der Erhebung bis zur wissenschaftlichen Publikation und einer Archivierung, welche die Daten zugleich für eine erneute Nutzung bereitstellt. Für eine Nachnutzbarkeit digitaler Daten ist allgemein die Vorstellung etabliert, dass diese unabhängig, das heißt ohne Hinzuziehung der bereitstellenden Fachleute, verstehbar und verarbeitbar sein sollen.[28] In der Forschung ist freilich die Einholung eines Mindestmaßes an Kontextwissen zu den Daten unabdingbar, um deren Genese verstehen und deren Potenzial (aber auch: Begrenzung) für die Nachnutzung einschätzen zu können. Zudem sind in der Forschung mit fast jedem Arbeitsschritt Prozesse verbunden, in denen Daten Transformationen durchlaufen. Daten haben somit gewissermaßen Schritt für Schritt verschiedene „Aggregatzustände".[29] Der Dynamik der Datentransformationen stehen Begriffe wie Datenprodukt oder Zwischenprodukt gegenüber, die suggerieren, dass es in den verschiedenen Arbeitsschritten stabile und abschließbare Zustände gibt. Gleichwohl sind auch diese möglicherweise fragil beziehungsweise nachträglichen Korrekturen unterworfen.

Unter digitalen Bedingungen, die – durchaus suggestiv – auch als „data continuum" beschrieben werden, wäre idealerweise jeder Schritt des Forschungsprozesses mitsamt seinen „Zwischenprodukten" möglichst transparent und explizit zu reflektieren und zu dokumentieren: „Data must be linked in a way that ensures

[28] Vgl. CCSDS (2012) – Reference Model OAIS, Kap. 3.1.

[29] Alle Verfahren, die einen aus einer Datenquelle stammenden Datensatz (also eine Entität) verarbeiten, um einen neuen Datensatz zu erstellen, sind mit einer „Transformation" dieser Daten verbunden. Tatsächlich sind digitale Forschungsprozesse in diesem Punkt weniger robust als herkömmliche Forschungsprozesse, in denen Objekte vielfach eine nachhaltigere physische Identität besitzen, die verschiedene Prozessierungsschritte überdauert (nicht also in jedem Schritt auf neuer Informationsverarbeitung beruht).

the continuum can be traversed."[30] Dies ist ein Grundgedanke, der auch dem Gebot der Nachprüfbarkeit wissenschaftlicher Aussagen und der Belegfunktion wissenschaftlicher Daten entspricht. Dass auch das Forschungsdesign, dem Daten entstammen, Anforderungen unterliegt, Forschungsfragen eines stimmigen Zusammenhanges bedürfen und Methoden fachgerecht zum Einsatz kommen müssen, sind weitere Aspekte, die Datenqualität im Rahmen eines Forschungsprozesses praktisch beeinflussen können. Auch diesbezüglich lassen sich Anforderungen im Datenlebenszyklus zumindest idealtypisch verorten.

Nutzung eines Datenlebenszyklusmodells zur Veranschaulichung der vielfältigen Herausforderungen für Datenqualität

Nachfolgend nutzt der RfII das Modell des Datenlebenszyklus in kritischer (von der Wissenschaft aus gesehen: selbstkritischer) Absicht. Hierzu wird eine Variante des Modells gewählt, die sehr früh eine Phase des Datenteilens und der Archivierung vorsieht, da Open Access und Nachnutzbarkeit derzeit im Fokus wissenschaftspolitischer Maßnahmen rund um Forschungsdatenmanagement und das Publikationswesen stehen,[31] aber auch, weil mit dem Publizieren, dem Teilen und der Langzeitverfügbarkeit von Daten wichtige Nachhaltigkeitsinteressen der Wissenschaft selbst berührt sind (vgl. die folgende Abb. 2 und Abb. 3). Der Zyklus trägt auch Konstellationen Rechnung, in denen Forschende auf Daten anderer zugreifen, um eigene Erhebungen zu validieren oder zu ergänzen.

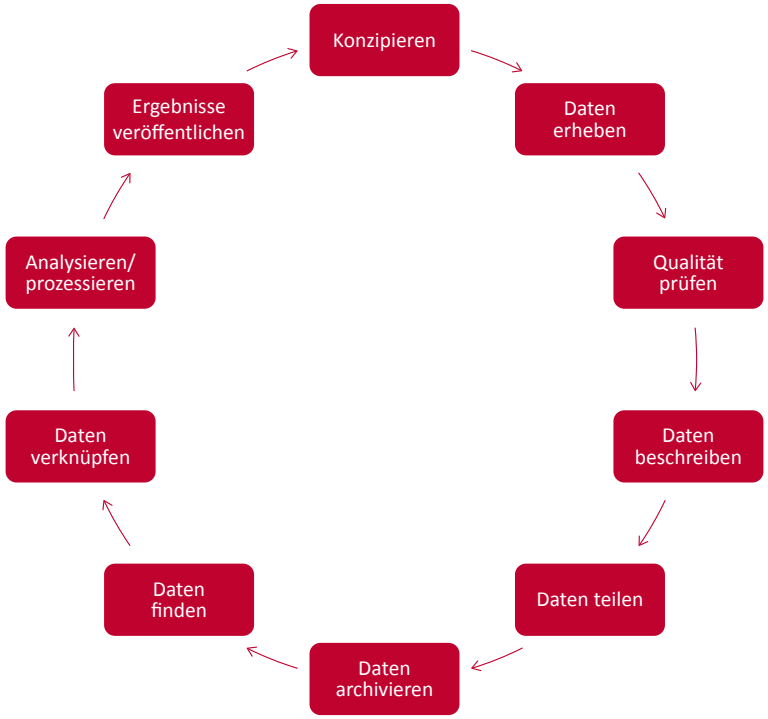

Abbildung 2: Der Datenlebenszyklus.
Quelle: eigene Darstellung in Anlehnung an German Federation for Biological Data (2019).[32]

[30] Field et al. (2013) – Common Challenges in Data Management, S. 6.
[31] Vgl. RfII (2019) – Stellungnahme aktuelle Entwicklungen Open Data.
[32] https://www.gfbio.org/training/materials/data-lifecycle/plan (zuletzt geprüft am: 30.08.2019)

Der Datenlebenszyklus wird in diesem Kapitel abgeschritten, um Probleme aufzuzeigen, die – unter den Bedingungen des digitalen Wandels – in der gelebten Realität der Forschungsprozesse und Forschungsformen der Umsetzung idealtypischer Qualitätsziele entgegenstehen. Insbesondere stellt das Ineinandergreifen der Prozessierung von digitalen und nicht-digitalen Daten eine Herausforderung dar. Denn nicht-digitale Daten wird es in Forschungsprozessen auch weiterhin geben – also etwa von physischen Objekten, aber auch von analogen Aufzeichnungsverfahren bis hin zu den habitualisierten Intellektualtechniken der Forschenden selbst.

Qualitätsideale und -herausforderungen

Neben den *Idealen* von Qualität lassen sich, so der Vorschlag des RfII, auch *Herausforderungen* an die Qualität von Daten mithilfe des Lebenszyklus benennen. Im Folgenden wird dies – pragmatisch und möglichst realitätsnah – versucht. Aufgrund der Komplexität der Forschungsprozesse und auch der Vielfalt der Forschungsformen können jeweils nur einzelne Schlaglichter auf die genannten Problemfelder geworfen werden.

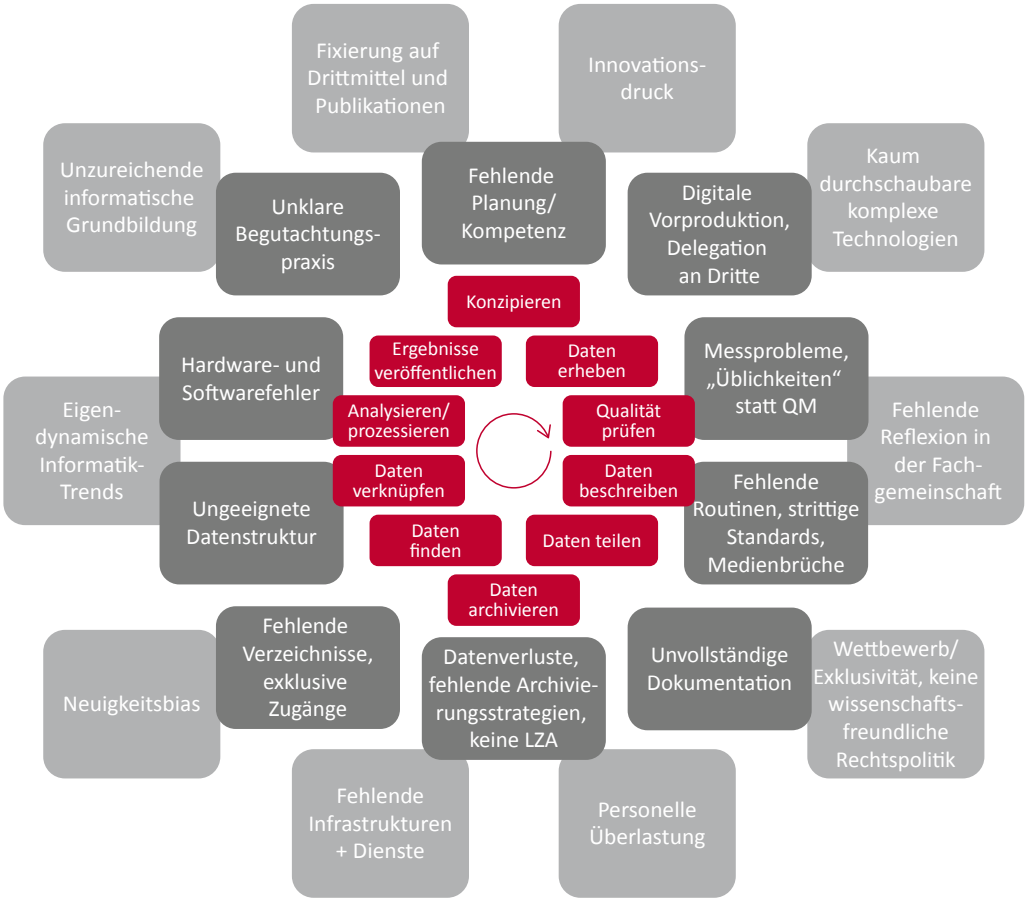

Abbildung 3: Herausforderungen für die Datenqualität im Datenlebenszyklus in (selbst-)kritischer Sicht. Legende
Innenkreis: Datenlebenszyklus, mittlerer Kreis: Probleme und Faktoren für Datenqualität entlang des Datenlebenszyklus, Außenkreis: potentiell hinderliche Rahmenbedingungen der Wissenschaft
Quelle: eigene Darstellung.

2.1.1 DATEN ERHEBEN

Daten werden idealtypisch zunächst erhoben oder sind abhängig von der Forschungsform irgendwann erhoben worden beziehungsweise wurden produziert. Dabei unterscheiden sich Methodik und Forschungsfrage (etwa: Ziele einer Studie) wie auch die Pragmatik von Datenerhebungen und die dabei berücksichtigten Qualitätskriterien. Grundlage können empirische/beobachtende, im technischen Sinne „einlesende" (also Daten übernehmende) Praktiken oder im hermeneutischen Sinne „lesende" Praktiken sein. Durch Digitalität sind hier teils vollständig neue Konstellationen entstanden.

Digitalität verändert Erhebungsbedingungen

Die automatisierte Datenerhebung durch digitale Sensorik, die „empirische" Neuinterpretation von visuellen Repräsentationen, Grafiken, Texten, Annotationen, die Produktion von Simulationsdaten oder auch die Erforschung von Nutzungsspuren in Digitalsystemen (einschließlich Maschine-Maschine-Interaktionen) sind Beispiele für Verfahren, die zu etablierten wissenschaftlichen Vorgehensweisen hinzutreten. Für Forschungsformen, in denen physische Objekte traditionell durch Beschreibung, Zeichnung oder Fotografie erfasst werden, stehen automatisierte Verfahren der Datenerhebung, zum Beispiel Scannen oder die 3D-Massendigitalisierung, zur Verfügung. Annotationen können bei automatisierter Erfassung sofort vergeben werden und treten an die Stelle von zuvor nachträglich erst hinzugefügten Beschreibungen. Auch Beobachtungsdaten sind immer schon – auch im Rahmen empirischer/beobachtender Forschungsformen – abstrahierende Darstellungen. Unter digitalen Bedingungen hängt ihre Qualität von einer Reihe von gerätebezogenen Faktoren ab, die sich auch als Formen einer digitalen (Vor-)Produktion beschreiben lassen. Auch typische Datenqualitätsprobleme haben hier bereits ihren Sitz.

Die folgenden Schlaglichter verdeutlichen einige der durch Digitalität substanziell veränderten Erhebungsbedingungen:

- *intransparente (proprietäre) Gerätesoftware:* Zur Datenerhebung werden vermehrt Geräte eingesetzt, deren Ergebnisse intransparente digitale Prozessierungen beinhalten, zumindest dort, wo die Wissenschaft auf Gerätesoftware keinen vollen Zugriff hat. Dies gilt nicht nur in den klassischen Forschungsfeldern, in denen gerätebasierte Messungen traditionell zum Einsatz kommen, sondern ganz besonders auch dort, wo Beobachtungsdaten Dritter verwendet werden (zum Beispiel Tracking-Daten). Soweit die komplexen Geräteparameter unbekanntes Firmeneigentum bleiben, stehen Forschende einer „Blackbox" gegenüber. Streng genommen werden so nur noch in einem sehr vermittelten Sinne „wissenschaftliche" Ergebnisse erlangt. Über wechselnde Software- oder Gerätegenerationen hinweg sind Erhebungsdaten nur dann wissenschaftlich valide und auch nachnutzbar, wenn entsprechendes Gerätewissen oder vergleichende Studien vorliegen.

Die von der DFG im Jahr 2019 aktualisierten „Leitlinien zur Sicherung guter wissenschaftlicher Praxis“ geben keinen Hinweis, wie mit diesem Problemkreis seitens der Forschungseinrichtungen und seitens der individuellen Wissenschaftlerinnen und Wissenschaftler umgegangen werden sollte.

- *dokumentationsarme Daten Dritter:* Forschung nutzt auch Daten, die Dritte erhoben haben. Die Datenerhebung ist einer Kontrolle nicht komplett zugänglich und auch die Dokumentation der relevanten Arbeitsschritte entspricht nicht ohne Weiteres wissenschaftlichen Standards, kann Lücken aufweisen, die sich nicht seriös füllen lassen, fehlt oder entspricht nicht ohne Weiteres wissenschaftlichen Standards. Ähnlich komplex stellen sich Qualitätssicherungsaufgaben dort dar, wo Forschung mittels entsprechender Unternehmensangebote (zum Beispiel Programmierungsschnittstellen) auf in Digitalmedien „frei“ entstandene Datenspuren eines sozialen Netzwerks zugreift. Punktuell suchen kommerzielle Anbieter und Plattformbetreiber zum Zweck der Datenanalyse und zur Verbesserung der Algorithmen durchaus Kooperationen mit der Wissenschaft, auch weil sich eigene Auswertungen als mangelhaft oder fehlerbehaftet erwiesen haben.[33] Zugänge, die gewährt werden, bleiben jedoch auf einen privilegierten Kreis von Forscherinnen und Forschern begrenzt – und diese dürfen oftmals nur als vertraglich gebundene, so genannte „embedded scientists“ tätig werden, die durch die Unternehmen über kommerzielle Filter- und Selektionsmechanismen der Datenerzeugung in unterschiedlichem Maße jeweils unvollständig aufgeklärt beziehungsweise im Rahmen von „Disclosure Policies“ in einer freien Analyse eingeschränkt werden.[34] Auch in diesen Fällen sind die Daten einer Nachprüfung nicht ohne Weiteres zugänglich, was ihre wissenschaftliche Qualität schmälert.

- *heterogene Datenmodelle vs. „Big Data“:* Bei der Erhebung von Daten stehen sich unterschiedliche Datenmodelle (und damit auf einer elementaren Ebene auch bereits unterschiedliche fachliche, methodisch geforderte oder sogar an die Forschungsfrage geknüpfte „Logiken“) gegenüber. Datenmodelle werden (zum Beispiel zwecks Standardisierung, aber auch aus inhaltlichen Motiven heraus) von Communities/Fachgemeinschaften entwickelt, sie beeinflussen wiederum die Details der Beobachtung und Erfassung. Eine solche

[33] Zum Negativbeispiel Google Flu Trends vgl. die klassische Studie von Lazer et al. (2014) – The Parable of Google Flu. Twitter bot 2014 pilothaft sog. „Data grants“ für ausgewählte Forschungseinrichtungen an, bei denen über einen „certified data reseller partner“ kostenfrei Datensätze für die Forschung bereitgestellt wurden. Facebook hat jüngst zusammen mit akademischen Partnern ein internationales, von verschiedenen Stiftungen gefördertes Kooperationsmodell entwickelt, das unter anderem Daten für die Demokratieforschung bereitstellt („Social Science One“, seit April 2018, online unter https://socialscience.one/our-facebook-partnership, zuletzt geprüft am: 30.08.2019); vgl. auch King/Persily (2019) – A New Model.

[34] Vgl. Pfaffenberger (2016) – Twitter als Basis wissenschaftlicher Studien, Kap. 4.

Präfiguration der Datenerhebung müssen breit angelegte, „übergreifend“ vorgehende Big Data-Ansätze ignorieren, bei denen strukturierte und unstrukturierte Daten kompiliert und so ihrerseits im Sinne einer „Erhebung“ beforschbar gemacht werden sollen. Die Qualität der Daten ist hier von strukturierenden Prozessen der Akteure abhängig, die durch die schiere Menge an Daten unmöglich nachzuvollziehen sind. In interdisziplinären Forschungsumgebungen vergrößert sich die Komplexität dadurch, dass sich eine Antwort auf die Frage nach (künftiger) Qualität erst in der Kombination unterschiedlich erhobener Daten ergibt. Das Problem von einer (nun: verdeckten) Ursprungsheterogenität besteht zugleich weiter. Dies kann nicht nur zu einer unklaren Validität von Aussagen, sondern auch zu Artefakten in der Auswertung führen und erfordert entsprechende Aufmerksamkeit.

- *unübersichtliche Vielzahl von Ontologien/Thesauri:* In beobachtenden und beschreibenden Forschungsformen werden kontrollierte Vokabulare als ein zentrales Mittel der Qualitätssicherung im Kontext der Datenaufnahme angesehen (Thesauri, Normdaten, Ontologien). Von diesen Vokabularen existieren jedoch sehr viele und die Forschung wendet sich inzwischen teils sehr simplen („generischen“) Standardisierungen zu. Der Qualität der Beschreibung von Daten bei ihrer Erhebung ist das nicht zuträglich.

Voraussetzungsvolle Dokumentation von Daten in den verschiedenen Phasen

In einer digital geprägten Arbeits- beziehungsweise Forschungsumgebung bestehen enorme Reichweiten für die Generierung von Daten. Bereits im Stadium der Datenerhebung existieren aber auch erhebliche Dokumentationsprobleme, die der wissenschaftlich geforderten Nachvollziehbarkeit von Arbeitsschritten entgegenstehen. Benötigt werden an die Verfahren angepasste Werkzeuge für die Dokumentation, welche zum Beispiel auch die Aufzeichnungen elektronischer Geräte integrieren. Hier sind in einzelnen Communities Lösungen in Ansätzen bereits vorhanden; eine flächendeckende Verbreitung steht allerdings noch aus.

2.1.2 QUALITÄT PRÜFEN

Herausforderungen für Prüfverfahren – Automatisierung als Lösung?

Als expliziter Qualitätssicherungsschritt gelten im Datenlebenszyklus Arbeitsschritte, in denen Erhebungsfehler und „Verunreinigungen“ beseitigt werden. Es handelt sich um ein genuines Kuratieren der Daten. Qualitätsprobleme können sowohl die Kriterien wie auch die Instrumente solcher Qualitätssicherungsmaßnahmen betreffen.

- *Explikation von Kriterien:* Auswahl, Präparation und Verdichtung der Daten können aufgrund angenommener Üblichkeiten (je nach Erhebungskontext nicht selten auch „händisch“) oder nach expliziten Kriterien erfolgen. Digitalität erzwingt ein deutlich höheres Maß an Explikation, damit Daten maschinell verarbeitbar sind. Qualitätskriterien können so einerseits konsequenter

angewendet werden, andererseits aber auch abstrakter (und damit ungewollt gröber) oder in der Anwendung risikoreicher sein, etwa wo Festlegungen zu Fehlertoleranzen getroffen werden oder wo „lernende“ (und also sich verändernde) Algorithmen zur Qualitätssicherung eingesetzt werden. Auch existieren in Communities/Fachgemeinschaften hinsichtlich geeigneter digitaler Qualitätssicherungsschritte – je nach Geschwindigkeit des digitalen Wandels – allenfalls Trends bezüglich geeigneter Kriterien. Diese wiederum müssen auf Risiko gewählt werden. Denn sie sind dem über Jahrzehnte eingespielten Wissen einer (auch) händischen Qualitätsprüfung nur begrenzt äquivalent beziehungsweise mit diesem nicht vermittelbar.

- *intransparente Werkzeuge:* Automatisierte Prüfverfahren können zum Einsatz kommen, wo Daten aufgrund der schieren Menge einer händischen Prüfung nicht mehr zugänglich sind, beziehungsweise wo dies aus Effizienzgründen geboten erscheint oder technisch einfach machbar ist. Beispiele sind logische Prüfungen zur Einhaltung bestimmter Werte-Intervalle in Tabellen, die sogenannten „Missing-Data-Techniken“ oder auch automatisierte Validierungswerkzeuge (vgl. 1.2.2). Wo kommerzielle Dienstleister vorprogrammierte Zwischenschritte anbieten und die Wissenschaft solche Voreinstellungen ungeprüft übernimmt, kann dies die Datenqualität auch negativ beeinflussen.

2.1.3 ANNOTIEREN

Daten sorgfältig zu dokumentieren, ist wesentlich nicht nur für die Datenqualität, sondern auch für die Nachvollziehbarkeit und somit die Qualität der Forschung an sich. Zumindest in den frühen Phasen der Datenerhebung kann die Beschreibung weder vom Erhebungszweck noch von der Forschungsmethode ganz getrennt werden. Da in der digitalen Welt Datenbeschreibungen auch wieder digital prozessiert werden, hat sich durchgesetzt, diese Zusatzinformationen als „Daten über Daten“, und zwar als „Metadaten“, zu bezeichnen. Für Metadaten hat die Forschung auf informationswissenschaftlicher Grundlage verschiedene, oftmals disziplinspezifische und auch transnationale Standards oder Metadatensysteme formuliert (vgl. 1.2.1 und Anhang A.1, Kap. 2.2).

Metadaten: Wichtige Zusatzinformationen – u. a. für die Maschinenlesbarkeit

Die Bedeutung der Metadaten und von (ggf. neuen) Metadatenstandards für die Qualität digitaler Forschungsdaten wird zu Recht vielfach betont.[35] In der digitalen Welt müssen in einem vormals unbekannten Ausmaß den Daten Zusatzinformationen zur maschinellen Prozessierbarkeit beigeschrieben werden. Automatisierte Verfahren arbeiten faktisch nicht auf Daten, sondern (allein) auf

[35] Vgl. unter anderem DINI (2018) – Thesen zur Informations- und Kommunikationsinfrastruktur.

Metadaten oder „Metadaten über Metadaten“, dies gilt für den Big Data-Bereich in besonderem Maße. Zugleich liegt auf der Hand, dass eine „vollständige“ Beschreibung von Daten nicht möglich ist und auf eine schrittweise sich verbessernde Dokumentation gesetzt werden muss. Metadaten bleiben deshalb selektiv, veränderlich und sind notwendigerweise nicht komplett. Die Praxis gestaltet sich entsprechend problematisch.

- *Routinen:* Für die analoge Welt hatte die Wissenschaft über Jahrhunderte etablierte und fortschreitend verbesserte Beschreibungsroutinen für die verschiedensten Datenarten und Medien (tabellierte Messwerte, Protokolle, Transkriptionen, Verzeichnisse für Proben/Archivgut, Audio- oder Diatheken, Bibliothekskataloge für Texte, Karten, Bilder, etc.). Hiervon ist vieles auf die digitale Welt nur begrenzt übertragbar, passt nur partiell, bedarf neuer digitaler Werkzeuge oder wird schlicht noch nicht systematisch gelehrt.

- *unklare Referenzierung Daten/Metadaten:* In über lange Zeiträume hinweg gewachsenen Datenbankstrukturen ist nicht immer unmittelbar zu erkennen, was die Bezugsgröße der Daten und Metadaten ist (was also genau individuiert wird). Beziehen sich etwa in einer digitalisierten Bilddatenbank Daten auf das abgebildete physische Objekt oder auf Repräsentationen (zum Beispiel Fotografien, Zeichnungen, Scans des Objekts)? Verfügen die digitalen Objekte nicht über eine eindeutige Identifizierung,[36] entstehen in der Praxis problematische Unklarheiten in der Einordnung der zugehörigen Informationen. Auch können die Urheberrechte an den beschriebenen Werken unklar sein. In wenigen Fällen können auch Metadaten im rechtlichen Sinn „Werke“ sein, jedenfalls wenn es sich um längere Texte handelt – was eine Identifikation des Autors beziehungsweise der Autoren erfordert.[37]

- *Provenienz und Transformation:*[38] Das Wissen über die Provenienz, also die Herkunft oder Entstehungsweise der Daten, ist eine wesentliche Information für das wissenschaftliche Arbeiten. Daten sind nicht zu trennen von Informationen zu verwendeter Software, Codes beziehungsweise Programmiersprachen, oft auch zu verwendeter Hardware. Entsprechende Beschreibungen erhöhen die Komplexität im Annotationsprozess. Die Erfassung von Metadaten zu Transformationsschritten gegebener Datensätze ist sogar über den gesamten

[36] Um das Problem zu lösen, wird heute die Vergabe einer registrierten Identifizierungsnummer empfohlen („persistent identifier“), zum Beispiel ein digital object identifier (DOI) oder die Uniform Resource Name (URN) für Internetobjekte. Wie lange die auf dem Markt derzeitig tätigen Registrierungsagenturen ihren Dienst aufrechterhalten können, ist allerdings eine offene Frage. Ebenso sind die Voraussetzungen einer DOI-Vergabe in der Diskussion.

[37] Siehe hierzu Klimpel (2015) – Eigentum an Metadaten.

[38] Gemeint sind Übersetzungs-, Umwandlungs- oder Verarbeitungsprozesse der analogen und digitalen Daten. Unter Transformationen sind nach Messung oder Erhebung zum Beispiel auch Modellierungen, abgeleitete Indikatoren oder Visualisierungen etc. zu verstehen.

Datenlebenszyklus hinweg und nicht allein bei der Datenerfassung und ersten Kuratierung ein Thema. Dass die Erfassung von Metadaten also hinreichend statisch und dynamisch gleichermaßen sein muss, macht gute Dokumentationen aufwendig und erfordert einen langfristig gedachten organisatorischen Rahmen.

- *händische und automatisierte Erfassung:* Wo Standards und gute Praxis des Beschreibens nicht konsentiert oder explizit formuliert sind, gilt es, als Forschender, Projekt oder Verbund selbst über Erfassungswege, Erschließungstiefe und Umfang der Dokumentation zu entscheiden. Metadaten werden oftmals händisch im Forschungsprozess erfasst oder erst im Nachgang (und dann unter anderen Gesichtspunkten) festgelegt, beispielsweise in der Vorbereitungsphase einer Publikation. Diese Praxis ist erheblich mühselig und fehleranfällig. In den experimentellen Wissenschaften wird forschungsunterstützende Dokumentationssoftware zunehmend als Option angestrebt, zum Beispiel in Form elektronischer Laborbücher. In anderen Forschungsbereichen wird der Einsatz solcher Werkzeuge erst gefordert.[39]

- *Versionierung und Datierung:* Metadaten unterliegen wie die Daten auch in der digitalen Welt typischerweise schnellen Veränderungen. Transformationen der Daten, Fehlerkorrekturen oder zusätzliche Informationen für neue Zwecke (zum Beispiel die Archivierung), machen es erforderlich, auch die Metadaten zu aktualisieren beziehungsweise zu datieren und zu versionieren. Die Zeitgestalt von Daten zum Beispiel einschließlich einer „Datierbarkeit" von Änderungen in wissenschaftlichen Prozessen zu dokumentieren, ist eine besondere Herausforderung des Annotierens.

- *Qualitätssicherungsbedarf für Metadaten:* Mit der steigenden Bedeutung der Metadaten im Forschungsprozess ist es notwendig, dass diese ebenfalls Gegenstand von Qualitätssicherungsprozessen werden. Wenn Datensätze vorliegen, in denen verschiedene Informationsebenen verbunden werden, stellt sich logisch wie pragmatisch die Frage nach Aufwand und Nutzen konkret vorgeschlagener Beschreibungssysteme. Aktuell vervielfältigen sich zudem die Beschreibungspfade und-sprachen. So existieren zwar transdisziplinäre Register für die eindeutige Identifikation sogenannter digitaler Objekte, deren institutionelle Nachhaltigkeit ist aber ungeklärt. Transparenten und fachlich erstellten, wissenschaftseigenen Ordnungssystemen steht zudem die Konkurrenz automatisierter Verfahren wie auf kommerzielle Zwecke optimierter „Suchmaschinen" und Linked Open Data-Anwendungen gegenüber, die gegebenenfalls auf andere Verknüpfungen setzen.

[39] Vgl. Peer/Green et al. (2014) – Committing to Data Quality Review, S. 275.

Zeitaufwand für Annotation konkurriert mit anderen Forschungsaufgaben

Insgesamt stellt sich der Beschreibungsaufwand in der Praxis als entscheidende Hürde dar. Metadaten, die unbeteiligten Dritten eine informierte Nachnutzung komplexer Datensätze ermöglichen, gehen weit über das hinaus, was zum Zeitpunkt der Datenerhebung dokumentiert wird. Für Forschende steht der Kuratierungsaufwand in Konkurrenz zu anderen Aufgaben, er verlangsamt die Forschungsabläufe tatsächlich unter Umständen erheblich. Die in der Wissenschaft üblichen, teils kurzfristigen Wechsel der institutionellen Anbindung in frühen Karrierephasen kommen erschwerend hinzu. Die Akzeptanz und Bereitschaft zur Annotation von Forschungsdaten hängt entscheidend auch von der Frage ab, welche Vorteile für den einzelnen Forschenden damit verbunden sind (zum Beispiel Reputationsgewinne, Zitationen) und ob er fachgerechte Unterstützung erhält. Entsprechend ausgebildetes unterstützendes Personal und geeignete Dokumentationshilfen fehlen im Arbeitsalltag allerdings oftmals.

2.1.4 DATEN TEILEN

Daten, die im Laufe des Forschungsprozesses entstehen, beziehungsweise angereichert und modifiziert vorliegen, werden unter Wissenschaftlern auf vielfältige pragmatische Weise zur Verfügung gestellt oder überlassen *(Data Sharing)*.

Data sharing: Von „peer-to-peer“ zu „open science“

Die traditionelle Form des Data Sharing ist die Weitergabe im persönlichen kollegialen Umfeld *(peer-to-peer)*. Mit dem digitalen Wandel gehen allerdings weitergehende Erwartungen an eine Datenverfügbarkeit einher, bis hin zu *open science/open data*, also der Verfügbarmachung der generierten Daten als eine Art wissenschaftlicher (wenn nicht gar gesamtgesellschaftlicher) Allmende.[40]

Teilen von Daten erhöht Vergleichbarkeit von Studien

Grundsätzlich ermöglicht das Teilen von Daten innovative Forschung und Vergleiche mit ähnlich gelagerter Forschung im gesamten Wissenschaftssystem. Es ist zudem wesentliche Grundlage für die Anwendung tragender Prinzipien guter wissenschaftlicher Praxis – unter anderem der (temporären) Validierung und der Falsifizierung von Forschungsergebnissen. Dennoch tun sich Teile der Wissenschaft aus nachvollziehbaren Gründen schwer mit der Anforderung, „alle“ Forschungsdaten frei zugänglich zu machen. Nicht für jede Form von Forschung und auch nicht für jedes Zwischenstadium und (Zwischen-)Ergebnis von Forschung ist die Offenheit der Datengrundlage zwingend geboten. Sinnvoll und notwendig wird sie dort, wo die Daten als solche für den weiteren

[40] Die Initiative Knowledge Exchange spezifiziert basierend auf Whyte/Pryor sechs Modi des Datenteilens, die sich zwischen den Polen „private management“ (sharing data with colleagues within a research group) bis zum „public sharing” (making data available to any member of the public) bewegen; vgl. KE (2014) – Sowing the Seed, S. 22; Whyte/Pryor (2011) – Open Science in Practice, S. 207.

wissenschaftlichen Fortschritt einen Mehrwert darstellen können – sofern sie für weitere Verwendungen nicht nur zugänglich, sondern auch im Hinblick auf Nachnutzbarkeit und Anschlussfähigkeit für weitere Forschung aufbereitet sind. Ist dies der Fall, dann stellen solche Daten und Datensätze eigenständige wissenschaftliche Produkte dar, die genauso wie die Ergebnispublikation als genuine wissenschaftliche Leistung zu bewerten sind.

Datenprodukte sind wissenschaftliche Leistungen

Datenprodukte können mittels dynamischer Systeme präsentiert werden (Datenkorpora, Datenkollektionen) oder in unterschiedlichen Formen archiviert und abgelegt sein („Archivpakete"). Es kann sich um empirisch erhobene Datensätze handeln, die für neue Auswertungen dienen oder mit früheren Datensätzen ähnlicher Art aggregiert werden, auch eine Kombination angebotener Daten mit zugehörigen Analysetools beziehungsweise Applikationen ist möglich. Das Produkt kann ebenso ein digitaler Katalog sein, der als Beschreibung, Datierung und Interpretation von Daten eine selbstständige wissenschaftliche Leistung darstellt. Im Kontext von Datenzentren sind Produktformate wie *Scientific/Public Use Files*[41] und *Data Reports*[42] bereits bekannt. Im Rahmen des wissenschaftlichen Publizierens haben die begleitend veröffentlichten Daten typischerweise eine Funktion als Argument oder Beleg für die präsentierten Ergebnisse (vgl. 2.1.9).

Digitalität erleichtert das Teilen von Daten maßgeblich. Dabei ist von Belang, ob und wie die Phase des Teilens in der Wissenschaft tatsächlich gelebt wird, denn hier finden gegebenenfalls wichtige Qualitätssicherungsschritte statt. Praxisprobleme sind dennoch zu konstatieren.

- *zugangsbezogene Qualität:* Die verschiedenen Datenprodukte unterscheiden sich in der Art und Weise, wie sie zugänglich und verarbeitbar sind, und ob der Zugang zu ihnen sicher ist und bleibt. Herausgaben in Eigenregie sind tendenziell stärker gefährdet, was die dauerhafte Verfügbarkeit betrifft. Sind die Datenprodukte online veröffentlicht, so sind sie gegebenenfalls nicht „offen", das heißt in entsprechend lizenzierten, maschinenlesbaren Formaten verfügbar.[43] Dies erschwert die weitere Verwendung (s. 2.1.7). Auch verbleibt der Großteil an Daten, die im Forschungsprozess gewonnen wurden, am Ort des Entstehens und ist kaum für Dritte zugänglich.

[41] Für Begriffsklärung siehe Statistische Ämter des Bundes und der Länder (Forschungsdatenzentren); https://www.forschungsdatenzentrum.de/de/zugang (zuletzt geprüft am: 30.08.2019).

[42] Zur Illustration seien zum Beispiel die standardisierten Datenberichte des Geoforschungszentrums Potsdam genannt, die vor Veröffentlichung eine interne Prüfung durchlaufen, vgl. http://dataservices.gfz-potsdam.de/portal/about.html (zuletzt geprüft am: 30.08.2019).

[43] Zur Definition „offen" nutzbarer Werke gehört zum Beispiel eine offene Lizenz, die Downloadbarkeit aus dem Internet und die Maschinenlesbarkeit; vgl. https://opendefinition.org/od/2.1/en/ (zuletzt geprüft am: 30.08.2019).

- *kontext- und darstellungsbezogene Qualität:* Daten müssen mit Informationen zu den Erhebungsmethoden, Prozessierungs- und Kuratierungsschritten, eingesetzten Werkzeugen und Systemen und damit letztlich auch Aussagen zu wichtigen Aspekten wie Datenintegrität und-authentizität versehen sein.[44] In der Praxis wird dieser Schritt oftmals durch das Fehlen von konsentierten Metadatenstandards und IT-gestützten Hilfsmitteln erschwert (vgl. auch 2.1.1 bis 2.1.3).

- *Datenschutz und Verfügungsrechte:* Das datenschutzkonforme Teilen von Daten erfordert ggf. besondere Aufbereitungsschritte, die die Auswertbarkeit der Daten für bestimmte Zwecke einschränken. Zudem ist vielen Forschenden nicht klar, wie sie ihre Daten bei gleichzeitiger Wahrung ihrer Interessen zugänglich machen können. Viele Datenbanken verfügen nicht über eine transparente Markierung der an den Daten hängenden Rechte beziehungsweise Lizenzen. Auch sind Verfügungsrechte an Daten oftmals nicht hinreichend geklärt beziehungsweise ausgewiesen.[45] Dies verunsichert vor allem in Bereichen, in denen die Daten möglicherweise einen kommerziellen Nutzen haben. Einsprüche anderer Projektbeteiligter, Haftungsfragen oder unerwünschte Formen der Aneignung durch Dritte bereiten Sorge.

- *schwierige Qualitätssicherung für „reine" Datenpublikationen:* Datenzentren oder Forschungs(-daten)-Infrastrukturen haben für Produkte, die in Eigenregie herausgegeben werden, Routinen interner Qualitätssicherung etabliert. Im Verlagswesen haben sich in den letzten Jahren sogenannte Data Journals[46] etabliert, welche Artikel über Datensätze („dataset articles" oder „data articles") veröffentlichen. Die eigentliche Datenpublikation erfolgt durch parallele Abgabe der Datensätze an Datenrepositorien oder-zentren. Oftmals erfolgt hier eine Qualitätssicherung aus Kapazitätsgründen nur kursorisch (vgl. dazu ausführlich 2.1.9 und 3.1.2).

- *eigene wissenschaftliche Reputation vs. kollektive Kuratierung:* Eigenständige Datenpublikationen sind eine Form digitaler Edition, die allerdings dynamisch sein kann. Eine Herausforderung besteht darin, die wissenschaftlichen Leistungen als „Ausgaben" zu versionieren und sie individuellen Forschenden zuzuordnen. In Analogie zum traditionellen Publikationswesen werden Konzepte wie Autorschaft und Datenzitation (oder „Referenzierung") diskutiert. Andererseits wird argumentiert, dass Daten(-produkte) idealerweise über

---

44 Einzelne Datenzentren haben beispielhafte Dokumentationen entwickelt, vgl. die „Product Types and Processing Levels" der European Space Agency ESA, https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/product-types-processing-levels (zuletzt geprüft am: 30.08.2019).

45 Lauber-Rönsberg et al. (2018) – Rechtliche Rahmenbedingungen FDM.

46 Beispiele von Data Journals finden sich unter https://www.forschungsdaten.org/index.php/Data_Journals (zuletzt geprüft am: 30.08.2019).

die Zeit und durch Nutzung Verbesserungen erfahren.[47] Gerade in der laufenden Pflege und Ergänzung bestehender Datensätze liegen erhebliche Potenziale für die Datenqualität. Dieser Prozess hat allerdings potenziell sehr viele Beteiligte über einen Zeitraum, in dem nicht unbedingt personelle Kontinuität besteht. Die Herausforderung, einen solchen „kontinuierlichen Verbesserungsprozess" zu managen, ist weitgehend ungelöst.

- *unbefugte Weitergabe:* Unter Druck stehen Kulturen des Teilens insbesondere dort, wo peer-to-peer-Prozesse von dritter Seite ausgenutzt werden, durch unfaire Konkurrenz oder andere Formen wissenschaftlichen Fehlverhaltens, durch Aneignung seitens wirtschaftlicher Akteure, durch verzerrende mediale (etwa netzöffentliche) Berichterstattung oder durch Industriespionage (s. a. Kap. 2.2 sowie 3.2).

Wissenschaftliche Kultur verändert sich langsam

Insgesamt bietet das Teilen von Daten vielfältige Chancen für wissenschaftliches Arbeiten – sei es im Kleinen (peer-to-peer) oder auch in Form gut ausgearbeiteter, in nachvollziehbaren Prozessen qualitätsgesicherter Datenprodukte. Dennoch verändert sich die wissenschaftliche Kultur hier erst langsam[48] und sie wird auch nicht durch die Schaffung geeigneter verwertungsfreier Räume von politischer Seite geschützt. Mangelnde Wertschätzung für die Erstellung von Datenprodukten im Vergleich zur klassischen Publikation und auch ein unklarer, ungeschützter Status des Teilens kommen also zusammen. Überdies ist das Teilen größerer oder erklärungsbedürftiger Datenmengen für die Datenproduzenten durchaus mit Aufwand verbunden. Auch legen Umfragen zu *Open Data* und *Data Sharing* unter Forschenden nahe, dass Forschende ihre Daten vorzugsweise erst zum Abschluss eines Projekts mit Blick auf die Ergebnispublikation ordnen. Im besten Fall wird die Bereitstellung/Archivierung der zugrunde liegenden Daten dann an dieser Stelle des Datenlebenszyklus nachgeholt. Kulturen des Teilens sind generell fragil. Wo allerdings der Forschungsprozess ein frühzeitiges Teilen von Daten zwingend erfordert – dies quasi Teil der Forschungspraxis ist – wird dieser Schritt auch in frühen Phasen gelebt.[49]

47 Parsons & Fox (2013) – Is Data Publication the Right Metaphor?, S. 39 f.

48 Belegt durch verschiedene Studien, unter anderem KE (2014) – Sowing the Seed; Fecher et al. (2015) – Academic Data Sharing; Wouters/Haak (2017) – Open Data; Stuart et al. 2018 – Practical Challenges in Data Sharing.

49 Beispiele finden sich in der Forschung rund um Großgeräte, beispielsweise der Auswertung von Satellitendaten in der Astronomie und der Erdbeobachtung, wie auch in den Biowissenschaften oder der textwissenschaftlichen Forschung.

2.1.5 ARCHIVIEREN

Die Langzeitsicherung digitaler Forschungsdaten ist ein intensiv diskutiertes Thema. Wird sie nicht professionell praktiziert, droht Datenverlust, auch deshalb, weil digitale Speichermedien verglichen mit anderen Medien kurzlebig sind sowie die Softwareversionen, auf denen Daten erstellt wurden, veralten. Wird Langzeitarchivierung praktiziert, sind Aufwand und Kosten angesichts rasant steigender Datenmengen und einer Vielzahl technischer Probleme aller Voraussicht nach erheblich. Was kommerzielle Partner für eine nachhaltige Wissenschaftsentwicklung im Bereich der Langzeitarchivierung leisten können, ist ebenfalls unklar.

Langzeitarchivierung als komplexe Aufgabe

Die Archivierung erfordert eigene Routinen der Aufnahme von Datenpaketen, ihrer Konvertierung zur sicheren Speicherung sowie der erneuten Bereitstellung für eine Nutzung. Hierfür existiert bereits ein international akzeptiertes Referenzmodell, das allerdings nur bei einer spezialisierten und mit Personal entsprechend ausgestatteten Einrichtung implementiert werden kann (vgl. Ausführungen zum OAIS-Modell im Anhang A.1., Kap. 2.2 und 2.4). In der Breite bestehen erhebliche Umsetzungsprobleme.

- *Zeitraum der Sicherung:* Die fachübergreifende Anforderung für die Aufbewahrung von Daten beträgt in Deutschland in der Regel zehn Jahre.[50] Die Bewahrung der fraglichen Primärdaten ist institutionell am Ort der Entstehung zu leisten und umfasst sowohl analoge als auch digitale Daten sowie deren Verknüpfung. Weder die hierfür erforderlichen Standards noch die (im Digitalbereich) entstehenden Kosten, geschweige denn ein System von verbindlichen Verantwortlichkeiten hat sich diesbezüglich im Wissenschaftssystem etablieren können. Richtlinien für das Forschungsdatenmanagement übertragen bislang abstrakte Verantwortlichkeiten, nicht aber konkrete Zuständigkeiten. Zwischen Ideal und Wirklichkeit dürfte daher selbst hinsichtlich der Mindestaufbewahrungsfrist eine Lücke klaffen. Insbesondere für die historisch arbeitenden Wissenschaften ist zudem eine Minimalanforderung dieser Art keine Lösung, Lösungen für eine echte Langzeitarchivierung (Aufbewahrung für die Ewigkeit) ist zumindest für eine Auswahl digitaler Artefakte unabdingbar.[51]

[50] Vgl. DFG (2019) – Leitlinien zur Sicherung guter wissenschaftlicher Praxis, S. 22, Leitlinie 17. Dieser Zeitraum wird für viele Felder als deutlich zu kurz kritisiert, andere betrachten ihn als zu lang. Im Vergleich zur Denkschrift der DFG von 2013 sprechen die jüngst aktualisierten Leitlinien im Haupttext (Leitlinie 17) heute von einem „angemessenen" Zeitraum der Aufbewahrung. Erst die Erläuterung verweist auf einen Regelzeitraum von zehn Jahren für „Rohdaten", der in begründeten Fällen aber auch verkürzt werden kann.

[51] Der RfII hatte hierzu unter anderem bereits einen Fachdiskurs über die Differenzierung zwischen projektlaufzeitnaher Speicherung sowie über deutlich länger zugeschnittene Archivierungsfristen angeregt, vgl. RfII (2016) – Leistung aus Vielfalt, S. 45, Empfehlung 4.3.

- *Auswahl und Selektion:* Während für den Umgang mit Schriftgut aus administrativen Prozessen Regeln für die Archivierung und Selektion existieren, fehlen solche transparenten Festlegungen in vielen Wissenschaftsbereichen. Welche Daten müssen archiviert und welche können oder müssen sogar gelöscht werden? Erst mit transparent formulierten Regeln kann bezüglich eines archivierten Datenbestandes eindeutig gesagt werden, wie repräsentativ, aussagekräftig und erhaltungswürdig er ist.

- *Regeln und Planung:* Vielfach sieht die Realität aufseiten der Datenproduzenten so aus, dass Daten ohne ausreichende Kenntnis über die Notwendigkeiten für eine Langzeitsicherung erhoben und prozessiert werden und auch die Ressourcen fehlen, anders vorzugehen. Aufseiten der Datenzentren fehlen wiederum die Kapazitäten und oftmals notwendigen Spezialisierungen, den heterogenen Datenbestand der vielfältigen Forschungsprozesse nachträglich zu kuratieren. Hier setzen Forderungen nach einer *Data Governance*, nach Forschungsdatenmanagementplänen und der Entwicklung einer Archivierungsstrategie an. Oft zeigen die Planungen allerdings, wie wenig davon realistisch machbar ist und somit verlangt werden kann.

- *fehlende Infrastrukturen und Dienste:* Obwohl jede forschende Einrichtung eine Archivierungsstrategie und Archivierungsmöglichkeiten sowohl für physische Artefakte (Proben etc.) als auch für Daten benötigt, um diese gemäß den Leitlinien guter wissenschaftlicher Praxis für die mindestens geforderte Spanne von zehn Jahren aufbewahren zu können, wird diese Aufgabe aufgrund fehlender Infrastrukturen standortabhängig oft nur ansatzweise gelöst. Auch der Zugang von Communities/Fachgemeinschaften zu geeigneten Diensten ist unterschiedlich gut – je nach Grad der Selbstorganisation, der Etablierung digitaler Methoden und der internationalen Vernetzung.

- *Kuratierung und Qualitätssicherung:* Fehlendes Personal führt zu einer ungenügenden Aufbereitung der Daten und Metadaten im Archivierungsprozess. Speziell die Harmonisierung in Bezug auf Struktur und Semantik der archivierten Datensätze ist im Einzelfall personell aufwendig. Beides ist allerdings maßgebend, um Daten aus unterschiedlichen Quellen effizient integrieren zu können (zu den Anforderungen an Datenintegration vgl. auch 2.1.7). Viele Aufgaben erfordern eine Kenntnis der wissenschaftlichen Domäne und persönlichen Kontakt mit den Datenproduzenten: Oftmals werden grundlegende Metadaten erst bei der Aufnahme ins Archiv nachgepflegt und müssen nachträglich erhoben werden. Ggf. sind bei wiederholter Datennutzung auch Änderungen oder Korrekturen am Datensatz zu dokumentieren und Querverweise zu den jeweiligen Nutzungen einzupflegen. Gerade Datenarchive und-repositorien können wichtige Beiträge zu einer laufenden Kuratierung und Pflege von Datenkorpora leisten, wenn sie personell entsprechend ausgestattet sind.

- *technische Infrastruktur:* Die Qualität der Basisinfrastruktur (Hardware, Betriebssysteme, Netzwerke, Sicherheit, Standards, Performanz, Datenmigration) hat ebenfalls Einfluss auf die Datenqualität. Die notwendigen laufenden Anpassungen der technischen Infrastrukturen können gerade in kleineren Einrichtungen kaum geleistet werden. Ähnlich problematisch kann es sein, im Personalbereich nicht über die notwendige Agilität und Innovationsbereitschaft zu verfügen. Ein Risiko sind zudem Datenverluste im Zuge von (Um-) Speicherung, Medien- und Formatwechseln (vgl. 2.2).

Von individueller zu kollektiver Verantwortung

Die Langzeitarchivierung wissenschaftlicher Daten ist für das Wissenschaftssystem sowohl in organisatorischer Hinsicht als auch mit Blick auf verfügbare Kapazitäten (Speicher, Kosten) insgesamt derzeit noch nicht ansatzweise befriedigend gelöst. Die für Deutschland im Aufbau befindliche Nationale Forschungsdateninfrastruktur (NFDI) wird für die Frage der Langzeitarchivierung nur Teillösungen bieten können. Angesichts erheblicher organisatorisch-institutioneller Fehlstellen ist eine Verlagerung der Archivierungspflichten auf die Schultern der Individualforschung, wie sie derzeit geschieht, nicht zielführend, um Datenqualität nachhaltig zu sichern.

2.1.6 DATEN FINDEN

Technische Möglichkeiten und disziplinäre Hürden bei der Auffindbarkeit von Daten

IT-gestützte Verfahren ermöglichen auf neue und sehr viel umfassendere Weise als zuvor das Auffinden von Information. Der Wissenschaft eröffnet das die Möglichkeit, neben der wissenschaftlichen Literatur auch vorhandene Datenerhebungen für neue Forschungsfragen und Auswertungen zu verwenden. So ist es im digitalen Zeitalter vor allem der Zugang zu Daten, der die Tiefe, Breite und Güte wissenschaftlicher Erkenntnisse bestimmt.

- *Suchen und Finden:* Das Potenzial, das in einer guten Auffindbarkeit von Daten liegt, bleibt vielfach ungenutzt, insbesondere auch deshalb, weil bestehende Sammlungen und Kollektionen entlang disziplinärer Grenzen und letztlich an einzelnen Communities/Fachgemeinschaften orientiert sind. Die Einbindung dieser Objekte in bestehende Katalogsysteme und Verzeichnisse wird angestrebt, die Reichweite dieser Bemühungen erfasst aber nur einen Teil der Datenlandschaft.[52] Trotz breiter Zustimmung, dass gute Wissenschaft auf guten Daten beruht, scheitern Wissenschaftler somit bereits häufig beim Auffinden der Daten. Erschwerend kommt im Einzelfall hinzu, dass Daten in Kooperationsprojekten (oder auch anderen Zusammenhängen wie Expeditionen) in fachlich oder institutionell verschiedenen Archiven aufbewahrt

[52] Die Ende 2018 von Google eingeführte Datensatz-Suche könnte Bewegung in die Situation bringen, vgl. Cousijn/Cruse/Fenner (2018) – Taking Discoverability.

werden. Ihre Zusammengehörigkeit und ihre Auffindbarkeit werden dadurch beeinträchtigt. Dies wäre technisch über eine Interoperabilität der Informationssysteme und Verknüpfungen zwischen eindeutig identifizierten Datenobjekten lösbar, wie sie unter anderem die FAIR-Prinzipien fordern. In der aktuellen Praxis ist dies jedoch meist nicht der Fall, sodass ein Wiederfinden zusammengehörender Datensätze oft unmöglich oder zumindest erschwert ist.

- Zugang: Oftmals ist die zugehörige wissenschaftliche Literatur der Schlüssel für das Auffinden eines Datensatzes. Soweit die Daten nicht in einem Repositorium oder anderweitig veröffentlicht sind, ist eine persönliche Anfrage beim Datenproduzenten erforderlich (Teilen von Daten *peer-to-peer*, vgl. 2.1.4). Das ist nicht immer erfolgversprechend – so können Ansprechpartner gewechselt haben, die aufbewahrende Organisation kann ihre Verfügungsrechte nicht wahrnehmen (oder sie sind unklar) etc. Auch der Zugang zu wissenschaftlich interessanten Daten aus dem Unternehmenssektor ist oftmals Verhandlungssache (vgl. 2.1.1). Bei sensitiven, zum Beispiel personenbezogenen Daten aus der medizinischen oder sozialwissenschaftlichen Forschung, wird die Zugänglichkeit durch rechtliche Rahmenbedingungen zum Schutz der Persönlichkeit zusätzlich kompliziert. Organisationsmodelle für einen Zugang zu solchen geschützten Daten existieren, sind aber weder wissenschaftsweit konsequent verbreitet noch homogen.[53]

2.1.7 DATEN VERKNÜPFEN

Verknüpfbarkeit erfordert fachliche Standards und Datenintegrität

Gerade die Verknüpfung von Daten aus unterschiedlichen Kontexten ist für wissenschaftliche Analysen attraktiv. Dieser Schritt ist jedoch im Einzelfall mit erheblichem Aufwand verbunden. Selbst wenn die Daten auffindbar und von Rechts wegen nutzbar sind, erschweren sowohl proprietäre Datenformate als auch die oftmals heterogene Struktur und Semantik der Daten deren Aggregation und Integration. Wo die Integration von Daten jedoch praktiziert wird, erfolgen gegebenenfalls auch wichtige Verbesserungen der kontextbezogenen Qualität, auch der Ursprungsdaten. Praxisprobleme betreffen vor allem Fragen von Datenstruktur und-integrität.

- *Struktur und Inhalte von Daten:* Die Struktur der Daten (zum Beispiel in einer Datenbank) ist oftmals nicht Ergebnis gezielter Planung, sondern mit bestehenden Methoden und über die Zeit hinweg pragmatisch weiterentwickelt worden (vgl. 2.2). Problematisch sind aber vor allem die Inhalte von Datenbankfeldern oder genauer gesagt deren Semantik. Exakte Begriffsdefinitionen sowie

[53] Vgl. Praxis der Forschungsdatenzentren in den Sozial- und Wirtschaftswissenschaften: RatSWD (2018) – Tätigkeitsbericht der Forschungsdatenzentren 2017.

die Beziehung von Begriffen zueinander in Form von Ontologien ist für die Verknüpfung von Datenbeständen essenziell. Hier fehlen teils fachliche Festlegungen, teils existiert – im anderen Extrem – eine verwirrende Vielzahl an anwendbaren Spezifikationen und Metadatenstandards, was es selbst Spezialisten erschwert, breit verfügbare Datenbestände aufzubauen. Forschungsprozess und Infrastrukturmanagement sind an dieser Stelle nicht gut verzahnt.

- *Integrität der Daten:* Technisch ist oftmals nicht klar, ob zum Beispiel die Integrität von archivierten Daten während der Speicherung sichergestellt wurde. Hinzu können Mängel im experimentellen Design kommen, die zum Beispiel eine statistische Auswertung verhindern, sowie undokumentierte Vorprozessierungen, die die Daten bereits in einer frühen Phase verändert haben (vgl. 2.1.1). Im schlechtesten Fall potenzieren sich Mängel aus früheren Phasen des Datenlebenszyklus und erschweren die Synthese von neuem Wissen.

- *händischer Aufwand:* Liegen Daten nicht in maschinenlesbarer Form vor, erfordert eine Integration hohen manuellen Aufwand mit entsprechender Fehleranfälligkeit – beispielsweise, wenn Daten aus der wissenschaftlichen Literatur extrahiert und in eine Datenbank überführt werden.

Die Verknüpfung von Datenbeständen birgt großes Potenzial für die Wissenschaft. Die Diskrepanz zwischen den theoretischen technischen Möglichkeiten und dem in der Praxis notwendigen Aufwand einer solchen Integration ist allerdings noch erheblich. Anläufe werden teils in Qualifizierungsarbeiten unternommen (hier ist die Strukturierung der Daten Teil der Lernleistung), teils werden ganze Projekte rund um die Integration relevanter Datenbestände konzipiert, beispielsweise in den Agrar- oder Umweltwissenschaften.

2.1.8 ANALYSIEREN UND PROZESSIEREN

Ursachen für Datenverfälschungen

Bei der Analyse und Prozessierung von Daten treten vergleichbare Probleme auf, wie bei der Datenerhebung und Qualitätssicherung von Daten. So wie bereits heutige digitale Messgeräte über Algorithmen gesteuerte Transformationsprozesse bei der Erhebung beziehungsweise Messung von Daten generieren, so haben die Algorithmen in Auswertungssoftware und -tools Einfluss auf die Ergebnisse. Die Analyse des Verhaltens von Algorithmen in komplexen Einsatzszenarien ist dabei selbst noch ein offenes Forschungsfeld. Zwecks einer detaillierten Betrachtung der möglichen Ursachen für Datenverfälschungen kann man grob zwischen Hardware- und Softwareproblemen unterscheiden.

Hardware als Fehlerquelle

Für viele Verarbeitungsprozesse, insbesondere mit Blick auf deren Replizierbarkeit, spielen bereits Unterschiede in der eingesetzten Computer-Hardware eine Rolle. Hiervon ist die langfristige Aufbewahrung von Daten, etwa im Fall der Lesbarkeit von Speichermedien, betroffen. Aber auch für Verarbeitungsprozesse müssen Hardware-Unterschiede präzise dokumentiert und in ihren Auswirkungen möglichst beherrschbar gemacht werden. Als Beispiele können genannt werden:

- *Hardwarefehler:* Die Ergebnisse von Berechnungen können von spezifischen Implementationsfehlern abhängig sein. Dies ist teils auf schlechte Software-Implementation zurückzuführen, die nur „zufällig" auf einer Hardware das gewünschte Verhalten zeigt, teils aber auch auf „korrektes" oder bekanntes Verhalten der Hardware.[54] Sofern Berechnungen von solchen Faktoren abhängen, lassen sie sich auf nicht betroffener Hardware nicht reproduzieren. Gerade für umfangreiche Berechnungen gewinnen auch Fehler der verwendeten Speicherbausteine an Bedeutung. ECC-RAM-Bausteine, die in bestimmten Grenzen solche Fehler erkennen und korrigieren können, sind für Server und Großrechner mittlerweile zum Standard geworden, auf Desktop-Rechnern und Notebooks bisher jedoch weniger verbreitet. Hardwarefehler können auch ein Sicherheitsrisiko sein, sodass beispielsweise die gezielte Manipulation von Berechnungen durch externe Angriffe möglich wird.

- *fehlende Emulierbarkeit:* Für ältere Hardware entwickelte Programme lassen sich auf neuerer Hardware nicht ohne Weiteres ausführen. Wo die Verwendung historischer Software erforderlich ist, müssen die Programme neu implementiert, konvertiert oder durch Emulatoren lauffähig gemacht werden, was eine Fülle von Detailproblemen mit sich bringt.[55] Wo Software für einzelne Arbeitsplätze nicht gekauft, sondern lizenziert wird, potenzieren sich bei Austausch der Hardware die Schwierigkeiten.

- *Erzeugung von Artefakten:* Computer-Hardware unterscheidet sich zum Beispiel hinsichtlich der Verfahren, die zur Implementation der Fließkommaarithmetik sowie der Erzeugung von Entropie/Zufallszahlen eingesetzt werden. Dadurch sind Berechnungen auf unterschiedlicher Hardware nicht unbedingt reproduzierbar.

- *physikalische Umwelteinflüsse:* Insbesondere im Fall der Wandlung analoger in digitale Signale (A/D-Wandlung) beeinflussen physikalische Umwelteinflüsse wie Temperatur, Druck, Luftfeuchtigkeit, elektromagnetische und radioaktive

[54] Ein Beispiel ist der „FDIV-Bug" der ersten Generation der Pentium-Prozessoren von Intel, der unter bestimmten Umständen zu einer deutlich geringeren Genauigkeit der Fließkommaeinheit führte.

[55] Zu nennen wäre hier vielleicht das eindrückliche Visual6502-Projekt, das Simulatoren für historische Prozessoren auf Transistor-Ebene entwickelt: http://www.visual6502.org/ (zuletzt geprüft am: 30.08.2019).

Einstrahlung, aber auch Erschütterungen und Beschleunigung die Verarbeitung. Dies kann zu erhöhtem Rauschen, zu Berechnungsfehlern und zu Systemabstürzen – etwa durch „kippende Bits“ – führen, weswegen für den Einsatz in besonderen Umgebungen teilweise gehärtete Hardware-Komponenten entwickelt wurden.

- *Alterung:* Zusätzlich zur Bedeutung subtiler Unterschiede der Umwelteinflüsse und der eingesetzten Hardware-Komponenten ist deren Alterung zu beachten. So verändern Sensoren auch jenseits stets bestehender Artefakte und Rauschquellen ihre Eigenschaften durch Alterung und Abnutzung, was beispielsweise zu Verschiebungen der Genauigkeit in verschiedenen Teilen des gemessenen Spektrums führen kann.

Professionelle Rechenzentren tragen dem Rechnung. Es verbleibt jedoch auch bei gutem Hardware-Management ein Restrisiko, das eruiert und gegebenenfalls beschrieben werden muss.

Bedeutung der Software für Datenqualität

Deutlicher noch als die Hardware spielt die zur Datenverarbeitung eingesetzte Software eine entscheidende Rolle für die Datenqualität:

- *Implementationsfehler:* Fehlerhafte Implementationen gefährden nicht nur die Stabilität von Software, sondern auch die Zuverlässigkeit der gewonnenen Daten und Analyseergebnisse. Die Korrektheit von Software lässt sich theoretisch nicht garantieren und kann selbst bei praktikablen Grenzen nur mit einem erheblichen Aufwand abgeschätzt werden.[56] Dabei können Implementationsfehler und mangelnde Softwarewartung zu Informationsverlust und zu Artefakten bei der Datenverarbeitung führen, was die Qualität der verarbeiteten Daten mindert.

- *Versionsunterschiede:* Ergebnisse von Datenverarbeitung und-analyse können sich bereits über verschiedene Versionen ein und derselben Software hinweg unterscheiden. Das Problem verstärkt sich, wenn identische Verfahren durch verschiedene Software-Pakete zur Verfügung gestellt werden, soweit komplexere Verarbeitungsverfahren sich in relevanten Implementationsdetails unterscheiden können. Die hier relevanten Unterschiede lassen sich nur schwer explizieren.

[56] Der Überprüfung von Software hinsichtlich Benutzeranforderungen (Validierung) und der formallogischen Korrektheit (Verifikation) sind praktisch und theoretisch enge Grenzen gezogen. Während die Verifikation sich insbesondere logischen Unentscheidbarkeitsproblemen gegenübersieht, bedarf die Validierung einer möglichst vollständigen, praktisch aber nur schwer leistbaren Explikation von Benutzeranforderungen, die sich zudem über den gesamten Lebenszyklus von Software hinweg beständig verschieben können. Vgl. Liggesmeyer (2009) – Software-Qualität.

- *Blackboxing:* Mittels proprietärer Software verarbeitete Forschungsdaten und daraus gewonnene Resultate lassen sich nur mit einer vollständigen Dokumentation der angewendeten Verfahren und möglicherweise nur durch den Einsatz der originalen Software selbst nachvollziehen oder replizieren. In der Praxis dürfte sich die Dokumentation der eingesetzten Verfahren regelmäßig als unzureichend erweisen. Zudem lassen sich Ergebnisse mit vertretbarem Aufwand auf Basis der Forschungsdaten allenfalls mit der Originalsoftware nachvollziehen, was die wissenschaftliche Nachprüfbarkeit beeinträchtigt.

- *maschinelles Lernen/lernende Algorithmen:* Auch der Einsatz von lernenden Algorithmen (künstliche Intelligenz, *Machine Learning*) schränkt die Nachvollziehbarkeit und Wiederholbarkeit von Berechnungen ein. Beispielsweise lassen sich Trainingsprozesse nur bei Verfügbarkeit der originalen Trainingsdaten exakt nachstellen. Resultate können in der Regel nicht befriedigend erklärt werden, da sich die aus den Trainingsdaten algorithmisch abgeleiteten Modelle durch den Menschen nur schwer oder gar nicht interpretieren lassen. So ist zwar ein detaillierter Nachvollzug einzelner Berechnungsschritte theoretisch möglich, liefert aber praktisch keine befriedigenden Erklärungsansätze. Die Erklärbarkeit (*explainability*) von Resultaten aus derartigen Algorithmen – gemeint ist die erklärende und begründende Beschreibung von Berechnungsergebnissen – ist ein dynamisches Forschungsfeld. Allerdings liegen bislang weder einsetzbare Tools noch überzeugende theoretische Konzepte für Erklärbarkeit und Verstehbarkeit vor.[57]

- *Sicherheitslücken:* Programme, die zum Prozessieren und Analysieren von Forschungsdaten eingesetzt werden, sind wie andere Programme auch von Sicherheitslücken betroffen. Hierdurch kann insbesondere die Verarbeitung von Daten gestört und deren Integrität sowie Vertraulichkeit beeinträchtigt werden. Die Vertraulichkeit von Forschungsdaten spielt zum Beispiel für die Einhaltung rechtlicher Bestimmungen (Datenschutzrecht, Urheberrecht/Lizenzbedingungen) eine wesentliche Rolle.

- *Anwendungsfehler:* Neben Bedienfehlern sind hier insbesondere unklare, schlecht dokumentierte oder ungeeignete Parametrisierungen zu nennen (Definition und Auswahl von Kategorien). Selbst bei korrekter Bedienung und geeigneter Parametrisierung der Software ist durch implizites Wissen mit einer Beeinträchtigung der Datenqualität zu rechnen, da implizites Wissen in der Regel nicht dokumentiert wird. Dies kann zum Beispiel die Abweichung von einer Default-Einstellung betreffen, mit der zwar stets verfahren wird,

[57] Vgl. etwa Lipton (2016) – The Mythos of Model Interpretability oder Samek et al. (2017) – Explainable Artificial Intelligence.

die aber nicht dokumentiert ist. Die Qualität der prozessierten Daten hängt aber auch von der Passung der eingesetzten Algorithmen ab. Diese Auswahl folgt im Einzelfall vielleicht eigendynamischen Trends und „Moden" der Informatik und weniger einer Anforderungsanalyse aus der jeweiligen Forschungsfrage heraus („genutzt wird, was da ist und einigermaßen passt").

Dokumentation in den Prozessstadien kann Fehlerquellen transparent machen

Viele dieser Probleme können durch eine verbesserte Dokumentation nicht unbedingt gelöst, aber zumindest transparent gemacht werden. Die Verwendung quelloffener und allen Wissenschaftlern frei verfügbarer Analysewerkzeuge ist ein weiterer derzeit gängiger Ansatz in vielen Wissenschaftsdisziplinen. Allerdings gilt das nur für Analysen, die keine zu großen Performanz-Anforderungen stellen. Außerdem bleiben Probleme der Reproduzierbarkeit zwischen unterschiedlichen Versionen und Hardware-Umgebungen ein offenes Problem. Die weitergehende, kollaborative Entwicklung von wissenschaftlicher Community-Software, die idealerweise auch Qualitätssicherungsverfahren umfasst, wie sie sich in einigen – meist simulations- und datenintensiven – Disziplinen etabliert, ist ein weitergehender Ansatz zur Schaffung von Transparenz und Reproduzierbarkeit.

2.1.9 WISSENSCHAFTLICHES PUBLIZIEREN

Veröffentlichung: Nicht nur Ergebnisse – auch Daten

Eine markante Station im idealtypischen Datenlebenszyklus stellt – nach Abschluss der Analysen – das Publizieren der wissenschaftlichen Ergebnisse dar. Zugehörige Daten werden als Beleg ebenfalls öffentlich verfügbar gemacht. Diese Praxis wird im Bereich der Fachzeitschriften durch Vorgaben der Verlage befördert. So kann in den jeweiligen Autorenleitlinien zum Beispiel eine Zusatzpublikation von Daten empfohlen oder Erklärungen zur Datenverfügbarkeit gefordert sein.[58] Ein weiterer Treiber sind die Forschungsdaten-Policies der Forschungsförderer (vgl. 1.2.3).

Erster Schritt zur Lösung: Enhanced Publications

- *Modelle:* Es hat sich eine Reihe von Modellen entwickelt, bei denen die Veröffentlichung von Daten zum Beispiel auf beigefügten Datenträgern (bei Monografien) oder als Supplement (zum Beispiel im PDF-Format) erfolgt. Auch Ansätze zur Publikation von Materialerschließungen in Datenbanken nehmen zu. Um 2010 wurde der Begriff *„Enhanced Publications"*[59] für eine Publikation geprägt, bei der über die sogenannten Digital Object Identifier (DOI) in den einschlägigen Datenbanken Verlinkungen zwischen der

[58] Exemplarisch ist die Übersicht verschiedener Policy-Stufen auf der Open-Data-Informationsseite von Springer-Nature, https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096 (zuletzt geprüft am: 30.08.2019). Fachzeitschriften anderer Verlage verfahren ähnlich.

[59] https://www.forschungsdaten.org/index.php?title=Forschungsdaten-Policies&oldid=3619 (zuletzt geprüft am: 30.08.2019).

Forschungspublikation und der ihr zugrunde liegenden Datenpublikation hergestellt werden können. Die *Enhanced Publication* soll mindestens die Nachvollziehbarkeit der präsentierten Ergebnisse, aber auch eine Nachnutzung der Daten durch Dritte verbessern helfen. Mit Blick auf die Datenqualität ist hier von Vorteil, dass in der kombinierten Publikation von Aufsatz/Monografie und zugehörigem Datensatz zusätzliche kontextuelle Informationen zu Erhebung und Forschungsfragestellung verfügbar sind. Probleme mit der Datenqualität können zum Beispiel dadurch bedingt sein, dass die publizierten Dateien zwar öffentlich, aber nur mit manuellem Aufwand in eine maschinenlesbare Form zu bringen sind. Die Zugänglichkeit ist somit eingeschränkt und bei einer Nachnutzung entsteht das Risiko von Übertragungsfehlern. Aktuelle Bestrebungen gehen daher in die Richtung, Daten in Repositorien statt in Supplement-Publikationen zu archivieren.[60]

- *Datenherkunft (Provenienz):* Zumeist werden zugehörige Daten in aggregierter beziehungsweise prozessierter Form publiziert, das heißt, sie haben bereits eine Reihe von Transformationen durchlaufen, oder es handelt sich um bereits selektierte Daten. Qualitätsprobleme in Enhanced Publications können dadurch bedingt sein, dass Transformationen, die die Daten von der Erhebung bis zu dem letztlich publizierten Stand durchlaufen haben, nicht dokumentiert sind. Mögliche Fehlerquellen sind dann nicht nachvollziehbar. Auch verwendete Software ist relevant für die Beurteilung der Qualität in der Weiterentwicklung von Datensätzen, sie steht in der Praxis aber ebenfalls nur selten zur Verfügung.

- *Statik vs. Dynamik:* Die derzeitige Praxis der *Enhanced Publications* führt dazu, dass die veröffentlichten Datensätze weitgehend statisch sind: Nachträglich identifizierte Probleme oder Fehler können zwar als Erratum publiziert oder in den Metadaten eines Repositorium-Objekts vermerkt werden. Wahrscheinlicher ist aber, dass solche Informationen undokumentiert bleiben, insbesondere weil handelnde Personen aus der Wissenschaft ausscheiden und in der Regel keine Vorsorge für die laufende Pflege von einmal publizierten Daten getroffen wird.

- *Compliance:* Der Umfang der zur Verfügung gestellten Daten und die Genauigkeit der Dokumentation kann eingeschränkt sein, wenn Forschende aufgrund anderer externer Faktoren – zum Beispiel dem Druck, weitere Artikel oder Publikationen zu produzieren – der Datenpublikation nur wenig Zeit

60 Vgl. Aufruf der „Coalition for Publishing Data in the Earth and Space Sciences“ (bereits durch eine Reihe von Verlagen unterzeichnet), http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences/ (zuletzt geprüft am: 30.08.2019). In anderen Fächern lassen sich ähnliche Überlegungen finden.

widmen können. Eine Steigerung der Datenqualität wäre hier durch eine hohe Priorisierung der Datenaufbereitung als Teil des wissenschaftlichen Publizierens zu erreichen – gegebenenfalls auch zu dem Preis einer kleineren Anzahl von Aufsätzen beziehungsweise Publikationen, die in einer Forschungseinheit produziert werden. Dies ist in den etablierten Reputationssystemen noch nicht wirksam verankert und gehört zu den gängigen Forderungen für gutes Forschungsdatenmanagement.

- *Qualitätssicherung:* Während Dissertationen, Fachartikel und Buchbeiträge einen standardisierten Peer-Review oder andere Formen der Qualitätssicherung durchlaufen, gilt dies für die zugehörigen Datensätze oftmals nicht: Viele Repositorien erlauben das Hochladen von Dateien ohne weitere inhaltliche Prüfung oder die Frage, ob diese vor Veröffentlichung gegebenenfalls ein internes Review durchlaufen haben. Auch in den sogenannten Data Journals wird auf eine ausführliche technische Prüfung meist verzichtet, sie ist schlicht zu aufwendig (vgl. 2.1.4 und 3.1.2). Über die Belegfunktion hinaus ist der Wert vieler dieser Datenprodukte möglicherweise gering.

- *neue Barrieren:* Eine Datenpublikation, die primär als Ergänzung zum klassischen Artikel beziehungsweise zur Monografie erfolgt, verbleibt im klassischen Publikationssystem und „erbt" gegebenenfalls Auffindbarkeits- und Zugänglichkeitsprobleme, die mit dem jeweils etablierten Subskriptionssystem verbunden sind.[61]

2.1.10 KONZIPIEREN

Herausfordernde Praxis im Forschungsdatenmanagement

Forschungsdaten und Publikationen sind Grundlage für die Planung und Beantragung neuer Forschungsprojekte und Datensammlungen. Im Abschreiten des Datenlebenszyklus wird deutlich, dass die Datenqualitätsfragen aufeinander aufbauen: Fehlende Informationen aus frühen Phasen, wie der Datenerhebung, werfen bei Archivierung, Analyse oder Publikation zusätzliche Qualitätsprobleme auf. Daher wird zunehmend gefordert, bereits in der Konzeptionsphase eines Forschungsvorhabens Dokumentation, Management und Archivierung/Publikation der Daten zu planen beziehungsweise in einer nachlesbaren Form niederzulegen (Datenmanagementpläne). Dem Ideal – Datenmanagement professionell geplant und umgesetzt – stehen Praxisprobleme gegenüber, die sich auch nachteilig auf die Sicherung der Datenqualität auswirken.

[61] Kritisch beleuchtet in Parsons & Fox (2013) – Is Data Publication the Right Metaphor?, S. 40. Die Publikation von Daten als Beleg zu einer wissenschaftlichen Publikation innerhalb des klassischen Publikationssystems tragen nicht per se zu offenerer Wissenschaft bei.

- *Akzeptanz:* Policies der Forschungsförderer – und in der Folge auch der Forschungseinrichtungen (vgl. 1.2.3) – haben innerhalb der Wissenschaft nicht nur die Aufmerksamkeit für das Thema Datenmanagement erhöht, sie macht sich auch strukturell in Form von neu eingerichteten FDM-Beratungsstellen oder der Einführung unterstützender Tools bemerkbar. Aus dem Kreis der Beratungsstellen wird aber auch von erheblichen Mühen berichtet, einzelne Forschende für die Erstellung von Datenmanagementplänen zu gewinnen – diesbezügliche Beratungsleistungen würden wenig nachgefragt. Als akzeptanzmindernd werden formale Vorgaben und der zusätzliche Aufwand angeführt.[62] Die Ursache für fehlende Akzeptanz der Planungsaufgabe liegt möglicherweise in der oftmals überbetonten Rationale begründet, Daten müssten mit Blick auf eine mögliche Nachnutzung durch Dritte kuratiert werden. Dies wird nicht von allen Forschenden als prioritär angesehen.

- *Bewilligung von Ressourcen:* Die Forschungsförderung selbst folgte in ihrer Bewilligungspolitik lange Zeit nicht immer konsequent ihren eigenen grundsätzlichen Forderungen. Mittel für das geplante Forschungsdatenmanagement und den zugehörigen Kuratierungsaufwand wurden nicht immer als förderwürdig eingestuft. In der grundsätzlichen Implementation und Umsetzung der Datenmanagementpläne stößt man zudem auf das grundsätzliche Problem, dass diese Tätigkeiten Daueraufgaben sind und keine befristeten Tätigkeiten. Datenmanagementpläne haben als Instrument nur dann qualitätssichernde Wirkung, wenn die damit verbundenen Aufgaben organisiert und Aufwände finanziert sind.

2.2 DATENINTEGRITÄT IM GESAMTEN DATENLEBENSZYKLUS

Übergänge und Übergabepunkte sind von Bedeutung

Der Nachvollzug des Datenlebenszyklus hat erkennen lassen, dass in jeder Phase typische Faktoren die Qualität von Daten negativ beinträchtigen können. Einige davon sind forschungsdatenspezifisch, andere generisch. Des Weiteren trat deutlich hervor, dass in verschiedenen Phasen ähnliche Herausforderungen auftreten, wie die Dokumentation von Transformationen, Wechsel der Medien und Formate oder das Erstellen von Beschreibungen. Diesbezüglich sind insbesondere die Übergänge zwischen zwei Phasen wichtig, die auch Übergabepunkte von einem Akteur zum anderen sein können (zum Beispiel Datenproduzent/Archiv).

Datenzyklen müssen den Forschungsformen entsprechen

Nimmt man den gesamten Datenlebenszyklus in den Blick, sind insbesondere zwei Aspekte von Bedeutung: die ausschnitthafte Wahrnehmung beziehungsweise Bedeutung des Datenlebenszyklus für bestimmte Fachgemeinschaften und

[62] Zum Beispiel Neuroth et al. (2012) – Langzeitarchivierung.

dort praktizierte Forschungsformen sowie das über den gesamten Lebenszyklus hinweg bedeutende Thema der Datenintegrität.

Ziel: Datenintegrität

In empirischen/beobachtenden Forschungsformen wird der Datenlebenszyklus oftmals abgekürzt, indem der Weg von der Datenerhebung direkt zur Analyse und Publikation führt. Die Kuratierung der Daten steht hier nicht im Mittelpunkt, die Bedeutung der fortlaufenden Dokumentation wird unterschätzt, sodass Informationen zur Provenienz der Daten und zu den vorgenommenen Transformationen lückenhaft sein können. In hermeneutischen/beschreibenden Forschungsprozessen werden Daten häufig nur minimalistisch bibliografisch beschrieben und gehen ohne Datenpublikation direkt in die Langzeitsicherung. Digitale Methoden zur Analyse werden nur punktuell angewandt. Bei Gedächtnisinstitutionen wie Sammlungen, Archiven und Bibliotheken stehen die Phasen zwischen „Erheben" und „Suchen" im Mittelpunkt. Zu den zentralen Herausforderungen für die Fachgemeinschaften gehört es also, dass sie analysieren, wie sie genau mit dem Datenlebenszyklus umgehen und wo sie sich dort jeweils aktuell befinden.

Dies ist umso bedeutsamer, als gute Wissenschaft die Aufrechterhaltung und Gewährleistung der Nachvollziehbarkeit, Genauigkeit und Konsistenz von Daten über deren gesamten Lebenszyklus hinweg fordert. Der wissenschaftliche Gegenstand muss unverfälscht derselbe bleiben, das Substrat der Forschungsarbeit muss darstellbar und sollte (möglichst alterungsfrei) auch materiell beständig bleiben. „Datenintegrität" lautet hierfür ein – informationswissenschaftliches – Stichwort.[63] Datenintegrität ist ein entscheidender Aspekt für den Entwurf, die Implementierung und die Nutzung jedes Systems, das Daten speichert, verarbeitet oder abruft. Fragestellungen des Risikomanagements, der Validierung sowie rechtliche Fragen zum Beispiel des Datenschutzes setzen hier ein. Durch die Vielzahl der (auch außerwissenschaftlichen) Akteure sowie das Ausmaß technologischer Komplexität ist digitale Wissenschaft auch unter dem Gesichtspunkt der Datenintegrität mit Herausforderungen konfrontiert:

- *Formatänderungen:* Im Zuge des Prozessierens von Daten werden diese regelmäßig in andere Datenformate überführt, worunter sowohl die Dateiformate im engeren Sinne fallen, die von Software-Komponenten verarbeitet werden, wie auch Absprachen über Verwendungen von Datenformaten im weiteren Sinne, zum Beispiel die Codierung von Zahlen und Zeichen, die Benennung von Feldern in Tabellen oder Formatierungen. Eine Reihe von Faktoren kann hierbei die Datenqualität beeinträchtigen:

63 Unter anderem zur „Integrität" von Datenprozessen, eine organisationsweite und daher besonders komplexe Aufgabe für Hochschulen, hat der Wissenschaftsrat Empfehlungen gegeben. Vgl. WR (2015) – Empfehlungen zu wissenschaftlicher Integrität.

Die Komplexität oder inkommensurable Paradigmen in der Modellierung von Informationen können bei der Konvertierung von einem Dateiformat in das andere die Qualität der codierten Daten beeinträchtigen. Proprietäre Dateiformate können gegebenenfalls von Anwendungen nicht vollständig oder nicht korrekt gelesen und geschrieben werden. Bekannt sind auch Datenverluste durch Kompressionsverfahren. Insbesondere Bild-, Audio- und Filmdaten werden oft komprimiert gespeichert (zum Beispiel als JPEG, MP3, MPEG), um den Speicherbedarf zu reduzieren. Durch mehrfache Verarbeitungsprozesse auf verlustbehaftet komprimierten Dateien summieren sich die Verluste allerdings auf, sodass derartige Dateiformate für die langfristige Verarbeitung und Nachnutzung von Daten nur eingeschränkt geeignet sind.

Innerhalb einzelner Verarbeitungsschritte können Datenformate die Qualität von Daten beeinträchtigen. Ein Beispiel ist die Encodierung von Zeichen, die durch eine unvollständige Implementierung des Unicode-Standards nicht korrekt verarbeitet werden können, etwa im Fall von Sortierungen. Ein anderes Beispiel ist die Verwendung ungeeigneter Farbraummodelle bei der Digitalisierung.[64] Beschrieben sind auch die durch Fließkommaarithmetik entstehenden Einschränkungen und Eigenschaften.[65]

- *fragile Datensicherheit:* Digitalität macht Forschungsdaten leicht erreichbar. So können Daten, sobald man sie vernetzt, vergleichsweise leicht verletzt und auch zum Objekt von „Angriffen“ werden. Je offener Wissenschaft zugänglich ist (was sie ja im Digitalzeitalter anstrebt), desto mehr exponiert sie sich und ihre Daten nicht nur als Zugriffs-, sondern auch als Angriffsziel. Mögliche Motive sind Schädigung/Kompromittierung von Forschenden, Forschungseinrichtungen oder der betreffenden Forschung, Industriespionage, Piraterie/Raub, aber auch andere Formen krimineller Sabotage. Sogar militärische Szenarien sind denkbar. Derzeit ist das Wissenschaftssystem in Deutschland (jenseits üblicher Schutzmaßnahmen) nicht zureichend auf diese Art von Gefahren eingestellt.

Datenintegrität umfasst bisher aber nicht alle nötigen Maßnahmen zum Schutz vor unautorisierter Modifikation. Der RfII hatte bereits in seinem Positionspapier LEISTUNG AUS VIELFALT empfohlen, dass die verantwortlichen Akteure

64 Zum Beispiel Farbraummodelle, bei denen hochauflösende Farbwerte für einen gröberen Farbraum ausgelegt sind. Ein Farbraummodell mit lediglich 256 Graustufen reicht für die Verarbeitung und Darstellung von Röntgenbildern nicht aus.

65 Der weitverbreitete Standard IEEE 754 zur Darstellung und Verarbeitung von Fließkommazahlen weist einige für bestimmte Berechnungen ungünstige Eigenschaften auf, die insbesondere durch Rundungsfehler zu ungenauen Ergebnissen führen können. In einer mehrstufigen Verarbeitung können sich die Rundungsfehler aufsummieren.

technisch-organisatorische Maßnahmen zur Datensicherheit deutlich stärker in den Fokus nehmen sollten.[66]

Während größere Rechenzentren im Rahmen ihrer Aufgabenwahrnehmung Konzepte zur Wahrung der Datenintegrität praktizieren, gibt es zahlreiche kleinere oder selbstorganisierte Forschungs- und Informationsinfrastrukturen, bei denen dieser Grad an Professionalisierung nicht erreicht werden kann.

2.3 FORSCHUNGSPROZESS UND DATENLEBENSZYKLUS INEINANDER INTEGRIEREN

Anders als es das anschauliche Modell des Datenlebenszyklus mit seinen Schritten beziehungsweise Phasen suggeriert, erweist sich das Datenmanagement im konkreten Forschungsprozess als ein Weg mit zahlreichen Hürden. Datenqualität entsteht zum einen im Forschungsprozess und bedarf der fachlichen Expertise. Zum anderen werden spezielle Expertisen für zahlreiche (informationstechnologische, technische und rechtliche) Schnittstellen benötigt.

Organisation an Schnittstellen

Als Herausforderung erweist sich die Komplexität und Beherrschbarkeit der Vielzahl von einflussnehmenden Faktoren und Schnittstellen. Im Prozess stecken verschiedene Szenarien von Arbeitsteiligkeit – zwischen „Menschen und Maschine“, zwischen „Lieferanten und Kunden“, aber auch zwischen Forschenden und wissenschaftsunterstützendem Personal, das teils am Ort der Forschung, teils in Infrastruktureinrichtungen tätig ist. Viele der Aufgaben erfordern eine Spezialisierung. Andererseits beinhaltet eine Auslagerung von Aufgaben des Forschungsdatenmanagements an hierauf spezialisierte Einheiten und Einrichtungen auch eine gewisse Abhängigkeit, zumindest aber eine partielle Abgabe von Kontrolle und Verantwortung seitens der Forschenden über wesentliche Teile ihrer Arbeit. Eine weitere Herausforderung besteht darin, die sehr heterogenen Modelle und Ansätze zur Sicherung der Datenqualität für die jeweils konkrete Forschungsaufgabe nutzbar zu machen. Hier sind lokale und Einzellösungen denkbar. Für die Wissenschaft insgesamt braucht es aber letztlich Konsensbildungsprozesse in Fachgemeinschaften, damit die gefundenen Lösungen skalieren und zu einer qualitätvollen Forschung insgesamt beitragen. Drittens braucht es eine adäquate Ressourcenausstattung. Datenqualität über den gesamten Forschungsprozess zu sichern und zu steigern, beinhaltet sehr umfassende Dokumentationsaufgaben. Teils lassen sich diese mithilfe von Software erleichtern. Gut ausgebildetes Fachpersonal ist entlang der Prozesskette jedoch unabdingbar.

66 RfII (2016) – Leistung aus Vielfalt, S. 62 f., Empfehlung 4.12.

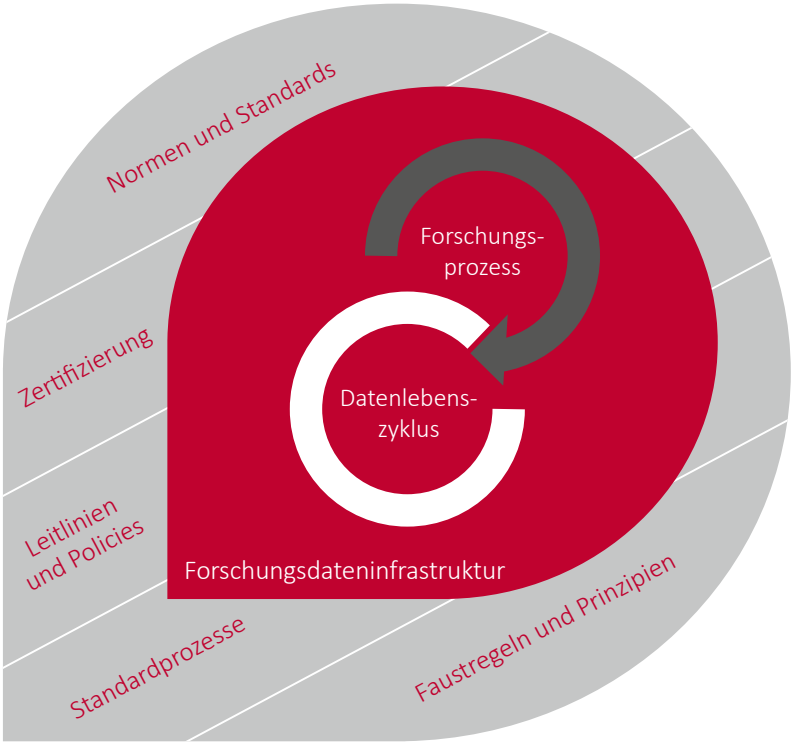

Abbildung 4: Multidimensionales Verständnis von Datenqualität.
Quelle: eigene Darstellung.

Datenlebenszyklus und Forschungsprozess verzahnen

Der aus dem einfachen zyklischen Modell entstehende Eindruck, zur Sicherung und Steigerung von Datenqualität könnten Wissenschaftlerinnen und Wissenschaftler den Forschungsdatenlebenszyklus quasi auf einem vollständig vorgezeichneten Pfad problemlos durchschreiten, ist eine Illusion. Der Datenlebenszyklus tritt neben den Prozesskreis der jeweiligen konkreten Forschungsaufgabe sowie das beteiligte Forschungspersonal beziehungsweise die Institutionen; beide zusammengenommen bilden einen multidimensionalen Raum mit vielfältigen Querbezügen zwischen den beiden Prozesskreisen (vgl. Abbildung 4).

Forschungsdaten-management als Aufgabe auf mehreren Handlungsebenen

Bezieht man die notwendige Regelungsebene mit ein, so erweist sich Forschungsdatenmanagement mit Blick auf die Datenqualität tatsächlich als komplexe Mehrebenenaufgabe. Für nachhaltige Fortschritte braucht es Aktivitäten aufseiten des individuellen Forscherhandelns, unterstützt durch ein organisationales und infrastrukturelles Umfeld, sowie auf der Ebene (internationaler) Fachgemeinschaften, in denen die grundlegenderen Diskurse zu methodischen Standards und wissenschaftlichen Gütekriterien stattfinden. Hierzu gibt der RfII in Kapitel 4 Empfehlungen.

3 DATENQUALITÄT UND DAS WISSENSCHAFTSSYSTEM

Die Beschreibung, Steigerung und Sicherung der Qualität von Forschungsdaten kann heterogenen theoretischen Modellen und Ansätzen folgen, das zeigte die Analyse in Kapitel 1. Ebenso vielfältig und facettenreich sind Faktoren und Prozesse, die in der heute gelebten Realität digitaler Forschung die Datenqualität steigern wie auch negativ beeinflussen beziehungsweise reduzieren können. An den Phasen des Datenlebenszyklus entlang wurde dies in Kapitel 2 skizziert.

Qualitätsdiskurs und krisenhafte Entwicklungen im Wissenschaftssystem

Eine dritte Sicht bezieht nachfolgend weitere Entwicklungen im Wissenschaftssystem (und seiner Umwelt) mit ein, die dazu beitragen, dass Forschungsleistungen und-erträge kritisch auf dem Prüfstand stehen. Dabei ist ebenfalls die Qualität der im Rahmen der wissenschaftlichen Wissensproduktion erzeugten Daten berührt. Mehrere krisenartige Tendenzen sind aktuell in der Diskussion: Die Rede ist von der Replikations- und der Reproduzierbarkeitskrise in Bezug auf veröffentlichte Forschungsergebnisse, von der Überlastung des Begutachtungssystems und von Fehlanreizen, welche die in vielen Disziplinen eingeführte Reputationsmessung mittels Publikationskennzahlen zu installieren scheint.

Einschneidende Veränderungen im Publikationswesen

Ebenso ändert sich das marktförmig organisierte Publikationswesen mit großer Geschwindigkeit: Einerseits versucht die Wissenschaft, Daten selbst zu publizieren, andererseits eignen sich Datenkonzerne große Marktsegmente des klassischen wissenschaftlichen Verlagsbereiches an und Daten verschwinden hinter Bezahlschranken, die teilweise nicht im Sinne der Wissenschaft sind. Auch politische Forderungen, die eine grundsätzlich „offene" Publikationspflicht für Daten aus der öffentlich geförderten Forschung postulieren, berühren die Frage nach deren Qualität unmittelbar.[67]

Digitalität tritt als systemische Herausforderung zu existierenden Krisen hinzu

Einige der genannten Entwicklungen haben bereits eingesetzt, bevor Digitalität als eine „systemische" Herausforderung für Datenqualität in der Wissenschaft auf den Plan trat. Die digitalen Optionen verschärfen allerdings die vorgängigen krisenhaften Entwicklungen weiter. Zugleich stellt der digitale Wandel – zumindest auf der technologischen Ebene – auch Instrumente bereit, die der Wissenschaft helfen können, Qualität zu steigern und gestärkt auch aus den öffentlich diskutierten „Krisen" hervorzugehen. Hierzu bedarf es allerdings nicht nur findiger Neuerungen auf der Ebene von Technologie (also etwa Prüfsoftware o. Ä.), sondern die Organisation des gesamten wissenschaftlichen Erkenntnis- und Erkenntnisverwertungsprozesses muss betrachtet werden. Für die Antwort auf die aktuellen „Krisen" und die Frage danach, was diese

[67] Über den Plan S wurde im Jahr 2018 intensiv diskutiert. Vgl. zur Qualitätsfrage im Rahmen der Umsetzung der PSI-Richtlinie (die Daten des öffentlichen Sektors betreffend) RfII (2019) – Stellungnahme aktuelle Entwicklungen Open Data. Zur PSI-Richtlinie generell siehe Kapitel 3.2.1.

verschärft beziehungsweise „treibt“, führt an einer systemischen Einbettung des Datenlebenszyklus kein Weg vorbei. Von einer solchen ganzheitlichen Einbettung kann dann die Rede sein, wenn man den Datenlebenszyklus ebenso wie Fragen der Harmonisierung und Standardisierung von Qualitätskriterien nicht losgelöst, sondern im Kontext sowohl des Forschungsprozesses als auch des Zusammenspiels der wissenschaftlichen Institutionen (Hochschulen, außeruniversitäre Forschung und explizite Infrastruktureinrichtungen) betrachtet. Zu den durch Digitalisierung induzierten Verflechtungen gehören auch bisher nicht gekannte außerwissenschaftliche und kommerzielle Einflüsse und Zugriffe auf die Wissenschaft. Ebenso sind hierzu neue, wissenschaftsfremde Verwertungsformen von und für Forschungsdaten zu zählen, auf welche die Wissenschaft selbst nur dann Einfluss hat, wenn sie sich mit dieser gewandelten Umwelt zügig und proaktiv auseinandersetzt.

Erhöhte politische Aufmerksamkeit für Datenqualität

Hierzu muss sie allerdings auch durch die Schaffung geeigneter wissenschaftspolitischer Rahmenbedingungen in die Lage versetzt werden. Bisherige Bemühungen um die Steigerung von Datenqualität in neu implementierten Informationsinfrastrukturen – vor allem aus der konkreten Forschung heraus geschaffene Datensammlungen, aber auch institutionelle Repositorien – hatten durch befristete Projektförderungsmöglichkeiten häufig einen kurzen Atem. Des Weiteren wurde die Erzeugung einer hohen Datenqualität auch in der Wissenschaftspolitik lange Zeit als selbstverständlich vorausgesetzt. Erst in jüngerer Zeit hat das Thema „nachhaltiges Forschungsdatenmanagement“ eine höhere Aufmerksamkeit gefunden, die sich zum Beispiel aktuell in der Etablierung einer Nationalen Forschungsdateninfrastruktur (NFDI) und analogen Gestaltungsbemühungen im Europäischen Forschungsraum (ERA), insbesondere in der European Open Science Cloud (EOSC) niederschlägt.

Krisen und Treiber fordern neue Arrangements zugunsten von Datenqualität

Im Folgenden wird kurz auf Krisen und Treiber eingegangen, die die Datenqualität in der Wissenschaft betreffen, wobei sie, durch die Digitalisierung begünstigt, sich zum Teil wechselseitig bedingen. Ihre Auswirkungen auf wissenschaftliche Forschung haben nicht nur nach Auffassung des RfII eine Größenordnung erreicht, welche die Schaffung neuer Arrangements in der Wissenschaft und für die Wissenschaft zu einer prioritären Gegenwartsaufgabe macht – gerade weil „Qualität“ und damit auch Datenqualität für den Mehrwert wissenschaftlich-methodengeleiteter Arbeit essenziell ist. Sowohl das individuelle und gemeinschaftlich-kollegiale Forscherhandeln als auch die Wissenschaftsförderung und-politik sind hier gefordert.

3.1 INNERWISSENSCHAFTLICHE KRISEN UND TREIBER

Zu den hier genannten Krisen zählen negative Entwicklungen, die aus der Wissenschaft und den Verfahren zur Produktion wissenschaftlichen Wissens selbst entstehen – und die sich mit Blick auf die exponentiell wachsende Produktion von Daten und Datenverfügbarkeit in den letzten Jahren kritisch verschärft haben.

3.1.1 MANGELNDE REPLIZIER- UND REPRODUZIERBARKEIT VON FORSCHUNGSERGEBNISSEN

In Disziplinen und Forschungsfeldern, in denen Replizierbarkeit methodisch möglich und geboten ist, liegen nicht wiederholbare Datenanalysen und Versuchsanordnungen mindestens in einer Grauzone des wissenschaftlichen Erkenntnisanspruchs. Als „Replikationskrise“[68] wird daher das Phänomen bezeichnet, dass wissenschaftliche Studien, Experimente und Simulationen statistisch signifikante Zusammenhänge und Kausalitäten behaupten, die in Nachfolgeuntersuchungen auf Grundlage unter nahezu gleichen Bedingungen erzeugter Daten nicht bestätigt werden können. Es liegt auf der Hand, dass hier ein Legitimationsproblem gerade für die öffentlich finanzierte, mit einem gesellschaftlichen Vertrauensvorschuss bedachte Wissenschaft besteht.

Forschungsdaten spielen für die Replizier- beziehungsweise Reproduzierbarkeit als Grundlage der statistischen Auswertung und anderer Formen der wissenschaftlichen Analyse eine wichtige Rolle. Entscheidend ist keineswegs allein die Qualität der erhobenen Daten, sondern der Umgang mit ihnen, vor allem die Qualität ihrer (Weiter-)Verarbeitung im Forschungsprozess. Entlang der Prozesskette der einzelnen Forschungsschritte stellen sich leicht Transformationseffekte ein, die zum Beispiel unter auch nur leicht veränderten Rahmenbedingungen in der Infrastruktur beziehungsweise den Analysewerkzeugen (zum Beispiel andere Hard- und Software-Versionen) dazu führen, dass Ergebnisse nicht wiederholbar sind.

Unkontrollierte Effekte der Datenverarbeitung beeinträchtigen Replizierbarkeit von Studien

So lässt sich beispielsweise hinterfragen, ob zahlreichen Ausgangsdaten nicht bereits im Zuge der Erhebung ein experimentelles Design zugrunde liegt, das in der statistischen Auswertung zwangsläufig nicht zu replizieren ist und damit nicht zu nachhaltig belastbaren wissenschaftlichen Ergebnissen führen kann. Hat man sich auf zu kleine Fallzahlen beschränkt, um schneller oder kostengünstiger voranzukommen? Stammen Daten gar aus nicht validierbaren Datenquellen? Hat man

[68] In der Auseinandersetzung um die adäquate Beschreibung dieses Qualitätsproblems – entweder als Replikations-, Replizierbarkeits- oder Reproduktionskrise mit je unterschiedlichen Ansprüchen an die Bedingungen der Wiederholbarkeit einer Studie oder eines Experiments – ergreift der RfII keine Partei. Für ihn sind diese Begriffe Ausdruck der gleichen Krisentendenz. Insofern werden hier die Begriffe synonym verwendet oder mit „Wiederholbarkeit“ umschrieben.

sich in der Datenerschließung und -verarbeitung auf instabil arbeitende Dienstleister verlassen? Wurden die Daten (eventuell durch *Blackboxing*-Effekte sogar unbemerkt) verändert, ohne dass dies transparent gemacht wurde und für Dritte nachvollziehbar ist? Oder ist die Arbeit mit „schlechten" Daten unproblematisch, wenn man sie nachträglich mit digitalen Mitteln nur „richtig" entschlüsselt? Werden Algorithmen genutzt, deren Verhalten aus wissenschaftlicher Sicht nicht kontrollierbare Effekte zeitigen kann? Lässt eine Datenbank-Software die Dokumentation all dessen, was dokumentiert werden müsste, wirklich zu – und wenn nicht: Wie ist mit Kompromissen umzugehen, die gemacht werden müssen? Wie ist mit Übersetzungsproblemen zwischen Computersprachen an „Schnittstellen" und überhaupt mit Aushandlungsprozessen bei der Schnittstellenprogrammierung zu verfahren? Und: Ist in ein Forschungsprojekt genug finanzierte Zeit für die Qualitätssicherung und zur Erstellung von Metadaten eingepreist und eingeplant?

Datenintransparenz kann durch Explikation und Rückbindung an Referenzobjekte beseitigt werden

Vor allem Fragen danach, warum genau im Einzelfall (ggf. unter dem Druck digital beschleunigter, unübersichtlicher gewordener Gegebenheiten) vieles nicht geschieht, zeigen einen neuen Qualitätsdiskurs in der Wissenschaft an, der in stärkerem Maße als früher – und ohne dass Standards bereits gefunden wären – Fragen der Datenqualität und der Prozessqualität ihrer Verarbeitung betrifft. Der Diskurs um Replikation unter digitalen Bedingungen ist einer, der keineswegs leicht abschließbar ist oder gar beliebig zu entscheiden wäre. Auch innerhalb einzelner Disziplinen kann die Frage zu Frontstellungen führen, wo genau die Grenzen des eigenen Replikations- beziehungsweise Reproduzierbarkeitsanspruchs liegen – sprich: Wo die methodisch gesehen „schlechte", die nachlässige oder sehenden Auges fehleranfällig organisierte (Daten-)Praxis beginnt.[69] Digitalität leistet insbesondere einem hohen Maß an Intransparenz Vorschub: Wie sehen die vielen, enorm komplexen Voreinstellungen aus, die in digitale Prozesse vorweg immer schon investiert, aber nicht expliziert sind? Und wie soll Reproduzierbarkeit gesichert werden, wenn nicht nur fehlende Explizierbarkeit, sondern auch lediglich „schwache" Formen der Reproduzierbarkeit im Digitalbereich vielfach der Normalfall sind? Dass im Zuge des digitalen Wandels der Zugriff auf relativ naiv, teilweise ohne jede weitere Dokumentation von Entstehungs- und Verarbeitungskontexten in Repositorien abgelegtes Datenmaterial anderer Forschender und Forschergruppen möglich ist, hat den Eindruck einer Replikationskrise auch jenseits der experimentellen Forschung, etwa bei klinischen Daten oder im Bereich der Bereitstellung von Digitalisaten von Kulturgütern zusätzlich angeheizt. Umso wichtiger ist heute auch die Rückbindung von Forschungsergebnissen an gut dokumentierte Daten in Sammlungen – zumindest dort, wo dies möglich ist. Dies bezieht ausdrücklich auch analoge beziehungsweise physische Sammlungen und Sammlungsgegenstände ein, die zum Zwecke der Validierbarkeit der auf ihrer Analyse beruhenden Forschungsergebnisse durch digitale

[69] Vgl. hierzu auch: DFG (2017) – Replizierbarkeit von Forschungsergebnissen.

Metadaten hervorragend ausdokumentiert sein müssen. Wo dies geschieht, können Sammlungen regelrechte Treiber guter wissenschaftlicher Praxis sein.

3.1.2 PROBLEME DES BEGUTACHTUNGSSYSTEMS

Peer Review und die Schwierigkeiten der Datenprüfung

Die kollegiale Begutachtung (im Folgenden: Peer-Review) hat sich in vielen wissenschaftlichen Feldern als Standard für die Qualitätsprüfung nicht nur der Beiträge, sondern auch für den bibliometrischen „Wert" einer Fachzeitschrift als ganzer durchgesetzt. Die Zahl der Begutachtungsvorgänge ist in den vergangenen Jahrzehnten enorm gestiegen – nicht eingerechnet die zahlreichen Evaluationen und Projektbegutachtungen, die von Wissenschaftlerinnen und Wissenschaftlern ebenfalls zu bewältigen sind. Die Verteilung einer wachsenden Zahl von eingereichten Manuskripten auf den relativ konstanten Stamm an einschlägig ausgewiesenen Gutachtern führt zu Überlastungen, die eine zeitnahe Veröffentlichung von Forschungsergebnissen erschweren, aber auch das Niveau der Arbeit der Peers bedrohen. Wird das Review durch miteingereichte und ebenfalls zu prüfende Datensätze anspruchsvoller, verlängert sich entweder die Wartezeit weiter oder der Qualitätsanspruch an die unter diesen Bedingungen leistbare Begutachtung sinkt. Es muss bei einer „Plausibilitätsprüfung" bleiben. Gleiches gilt für Forschungsergebnisse, deren Datengrundlage sich schon aus quantitativen Gründen nur schwierig qualitativ beurteilen lässt, zum Beispiel im Bereich komplexer Simulationen oder der hochauflösenden Elektromikroskopie.

Hohe Anforderungen an das „Data Reviewing"

An die Qualitätsprüfung von Daten werden oftmals deutlich komplexere Anforderungen als an diejenige von Texten (vgl. dazu auch 2.1.4 und 2.1.9) gestellt. Im Zweifel müssen Gutachterinnen und Gutachter die Datengewinnung mitsamt den konkreten digitalen Verarbeitungsschritten vollständig nachvollziehen können. Aufseiten der Forschenden müssen dazu sehr umfangreiche Dokumentationen bereitgestellt werden. Hier greifen gewissermaßen die Begutachtungskrise und die sogenannte Replikationskrise kritisch ineinander.

Unklare Leitlinien der Überprüfung

Das Problem wird auf verschiedenen Ebenen diskutiert, sowohl was Leitlinien für den Peer-Review als auch alternative Begutachtungsformen betrifft. Klar ist, dass ein technisches und wissenschaftliches Begutachten von Datensätzen durch kompetente Dritte („Peers") aufwendig ist und bei dem rasanten Aufwuchs publizierter Datensätze kapazitätsmäßig nicht skaliert. Hinzu kommt der Umstand, dass dort, wo Datensätze bereits in eine Begutachtung wissenschaftlicher Ergebnisse eingeschlossen werden, die Leitlinien der Begutachtung oft unklar sind. Nicht selten behelfen sich Verlage und Herausgebergremien mit Begutachtungsvorgaben, die dem bekannten Feld der Ergebnispublikationen entliehen sind – Originalität von Ergebnis und Argumentation, Beitrag zum fachlichen/wissenschaftlichen Fortschritt etc. –, die aber im Bereich der Begutachtung von Datensätzen in die Irre führen und sowohl die Gutachterinnen und Gutachter als auch die Begutachteten geradezu entmutigen.

Eigens für die Datenbegutachtung entworfene Leitlinien, die eher Kriterien der Akkuratesse von Datenerhebung und-prozessierung, dem Dokumentations- und Aussageniveau der Metadaten und – damit eng verknüpft – der Anschlussfähigkeit der Daten für die Nachnutzung (für weitere Forschung wie auch für Zwecke der Validierung der mit den Daten bereits geleisteten Forschung) enthalten, sind zwar vorhanden, kommen aber im wissenschaftlichen Publikationswesen kaum zum Einsatz.[70]

Datenrezensionen als eigenständige Publikationsform

Eine Variante der Begutachtung sind die aufkommenden „Datenrezensionen" als eigenständige Publikationsform. Hier beurteilen Peers (subjektiv) Datensammlungen oder-produkte unter wissenschaftlichen Gesichtspunkten.[71] Teils übernehmen auch Datenrepositorien oder andere Forschungsinfrastrukturen Teilaufgaben der Qualitätsprüfung vor der Publikation, oder verantworten diese in Gänze, wenn sie eigene Datensätze öffentlich bereitstellen. Auch dies deckt die Landschaft nur teilweise ab. Unter dem Schlagwort „reuse is the peer-review of data" wird die Vorstellung kultiviert, dass die Intensität der Nachnutzung als Indikator für qualitativ hochwertige Datensätze dienen könne. Dies funktioniert jedoch nicht in beide Richtungen: Geringe Nutzung von Datensätzen ist kein Indikator für schlechte Qualität und im Zweifel kosten nicht unmittelbar ersichtliche Qualitätsprobleme Nachnutzende viel Zeit oder schlagen direkt auf die Qualität der darauf basierenden Forschung durch. Vertrauenerweckender erscheint die Idee von „Data reviews", die im Sinne wissenschaftlicher Rezension unabhängig zu einem Datensatz oder einer Ressource veröffentlicht werden.

Kollektive Kuratierung von Datensätzen

Dem Netzwerkgedanken folgend ist auch eine arbeitsteilige oder kollektive Kuratierung von Datensätzen denkbar, wie sie das Wikidata-Projekt, die Geo-Wiki-Plattform für Naturbeobachtung oder – basierend auf dem Engagement von Akademiemitgliedern – mit Blick auf die Kuratierung von historischer deutschsprachiger Literatur das Deutsche Textarchiv (DTA) organisiert.[72] Sollbruchstelle scheint hier die Formierung einer Community zu sein, die eine solche Aufgabe über einen langen Zeitraum verlässlich wahrnehmen kann und will. Insgesamt ist jedoch zu konstatieren, dass die Limitierungen des Begutachtungswesens im Bereich der Datenpublikationen derzeit noch weitgehend toleriert werden. Strengere Auflagen für Autorinnen und Autoren, der konsequente Einsatz technischer Hilfsmittel oder eine Reform des Begutachtungswesens in einem hierfür unerlässlichen internationalen Maßstab gibt es bislang nicht.

---

70 Carpenter (2017) – What Constitutes Peer Review of Data.

71 Beispielsweise ist in den Digital Humanities eine eigene Rezensionszeitschrift entstanden, die sich digitalen Editionen und Ressourcen widmet und ein kritisches Forum für ihre Besprechung bietet (einschließlich einer Kriterienliste); https://www.i-d-e.de/publikationen/ride/ (zuletzt geprüft am: 30.08.2019). Weitere Beispiele finden sich in den Wirtschafts- und Sozialwissenschaften – hier allerdings als Teilbereiche bzw. Rubriken in einschlägigen Fachzeitschriften wie etwa dem Journal of Contextual Economics – Schmollers Jahrbuch oder dem European Sociological Review.

72 https://www.wikidata.org/wiki/Wikidata:Main_Page, https://www.geo-wiki.org/, http://www.deutschestextarchiv.de/ (alle zuletzt geprüft am: 30.08.2019).

3.1.3 FEHLENTWICKLUNGEN DES WISSENSCHAFTLICHEN PUBLIZIERENS

Fehlanreize durch Quantifizierung von Forschungsleistung

Hinter dem überlasteten Begutachtungssystem steht eine durch verschiedene Faktoren getriebene quantitative Überdehnung des Publikationsoutputs insgesamt. Fehlanreize in der Forschungsförderung und -finanzierung haben über längere Zeit dazu geführt, dass das Schielen auf die Quantität in der Wissenschaft selbst Einzug gehalten hat, unter anderem in Form informeller Leistungsanforderungen („drei Publikationen pro Jahr"). So ziehen insbesondere Forscherinnen und Forscher in der Qualifikationsphase die zügige Ausschlachtung eines Ergebnisses in mehreren Häppchen einer umfassenderen, gut dokumentierten Ergebnispublikation vor. Demgegenüber würde eine gut dokumentierte und maschinenlesbare Datensammlung als Supplement oder gar als eigenständige Edition komplementär zu der Ergebnispublikation mehr beziehungsweise zusätzliche Zeit und Arbeitskraft erfordern. Ein nicht zu unterschätzender Treiber dieser auf Quantitäten setzenden Entwicklung ist auch die zunehmende internationale Konkurrenz. Insbesondere aufholende „Wissenschaftsnationen" setzen hier gezielt Anreize zur Steigerung des Publikationsoutputs für „ihre" Wissenschaftseinrichtungen und Forschenden, um zum Beispiel in internationalen Rankings ihren Tabellenplatz zu erhöhen und damit nach innen und außen ihre (im Spiegel dieser Indikatoren) zunehmende Leistungsfähigkeit zu demonstrieren.

Qualitätssteigerung durch Reduzierung des Publikationsoutputs?

Während auf der einen Seite Kapazitäten in die Begutachtung gesteckt werden müssen, geht mit der steigenden Zahl von Publikationen auch die berechtigte Sorge einher, dass – schon was den Aufwand angeht, der seitens der Autorinnen und Autoren zeitlich geleistet werden konnte – die Qualität der Publikationen selber sinkt. Aus der Forschung heraus wird daher bereits seit längerem eine deutliche Reduzierung des Publikationsaufkommens im Sinne einer Konzentration auf wenige aber qualitativ sehr hochwertige Publikationen diskutiert.[73] Auch forschungspolitisch haben im deutschen Wissenschaftssystem ein Fälschungsskandal in der Krebsforschung 1998 und die Plagiatsskandale bei Promotionen 2011 zu einer kritischen Reflexion des Verhältnisses von Quantität und Qualität von wissenschaftlichen Publikationen geführt.[74] Der Wissenschaftsrat formulierte 2015 noch einmal eindrücklich eine grundsätzliche Kritik an der heute gängigen Publikationspraxis des „publish or perish" und ihren Folgen.[75] Die DFG hat auf die zu stark quantitativ ausgerichtete Selbstdarstellung von Forschenden mit einer drastischen Reduktion der in Drittmittelanträgen aufzulistenden „eigenen" Publikationen reagiert.

---

[73] So stimmten in einer Umfrage des Deutschen Hochschulverbandes nahezu 82 % der Teilnehmenden einer Forderung zu, die Zahl wissenschaftlicher Publikationen zu halbieren. Die Forderung wurde von Helga Nowotny formuliert, der ehemaligen Präsidentin des Europäischen Forschungsrates ERC. Vgl. Deutscher Hochschulverband, Barometer, https://www.hochschulverband.de/frage-des-monats.html?&tx_jkpoll_pi1%5Bno_cache%5D=1&tx_jkpoll_pi1%5Buid%5D=118#_ (zuletzt geprüft am: 30.08.2019).

[74] Kleiner (2010) – Qualität statt Quantität.

Auch die Evaluationsverfahren und Vorgaben anderer Forschungsförderer gehen dazu über, nur noch eine Auswahl der besten (oder aber unmittelbar einschlägigen) Publikationen als Nachweis von „past merit“ anzufordern. Nachsteuerungsversuche dieser Art haben insgesamt aber noch nicht grundsätzlich zur Überwindung des teils einseitig auf Quantität ausgerichteten Publikationsverhaltens geführt, was auch die einschlägigen internationalen Universitäts- und Wissenschaftsrankings zeigen. Gutachterinnen und Gutachter holen überdies nun auf informellem Wege die offiziell nicht mehr eingeforderten Gesamtdarstellungen ein, um sich für ihre Entscheidung ein Bild zu machen. Ein „Kulturwandel“ zeichnet sich hier noch nicht ab.

Data Review kann Druck im System erhöhen

Wird neben der Publikation von Ergebnissen auch eine (ggf. unabhängige) Publikation von Forschungsdaten gefordert, erhöht sich der Druck auf ein ohnehin überdehntes System. Ein als „Publikationsleistung“ gerahmter, wettbewerblicher Markt für Datenpublikationen wirkt, wenn diese als Forschungsleistung quasi „on top“ erstellt werden müssen, direkt zurück auf die Forschenden wie auch auf die Begutachtenden. Im Wege einer allzu druckvollen Forderung nach Publikation von Daten zeichnet sich ein Risiko ab, die bereits eingetretene Fehlentwicklung noch zu verschärfen und die angestrebte Reproduzierbarkeit und Nachnutzbarkeit von Ergebnissen nicht zu erreichen.

3.2 KRITISCHE EFFEKTE UNZUREICHENDER RAHMENSETZUNG FÜR DIE WISSENSCHAFT

Folgen der Internationalisierung für die Datenqualität

Angesichts der vielfältigen internationalen Kooperationen und der daraus resultierenden globalen Verflechtungen in Politik, Recht, Wirtschaft und öffentlichem Leben verändern sich auch die Rahmenbedingungen der Wissenschaft und ihrer Organisationen.[76] Komplexe internationale Beziehungen werfen neue Fragen auf, die weit über das grundsätzliche Thema Digitalität hinausgehen und das Ideal des freien Zugangs zu Bildung, Wissen und Kulturgütern sowie die internationale Vernetzung von Forschung betreffen. Wissenschaft operiert global, gleichwohl muss sie nationale Randbedingungen berücksichtigen. Sind in einer sich verändernden politischen Landschaft beispielsweise Spiegelserver für die Datensicherung in einem anderen Land oder auf einem anderen Kontinent angeraten? Welche Rolle spielen globale, kommerzielle Anbieter von Data Space (bzw. Cloud-Dienste)? Wie sollen sich nationale Wissenschaftseinrichtungen

[75] „Der Wissenschaftsrat weist erneut darauf hin, dass unrezipierbare Mengen von Publikationen den eigentlichen Sinn der Publikationspflicht konterkarieren, die ursprünglich der Kommunikation und Überprüfbarkeit neuer Forschungsbeiträge durch die wissenschaftliche Gemeinschaft dienen sollte. Alle Akteure sind dazu aufgerufen, den langfristig notwendigen Wandel hin zu einer stärker qualitativen Forschungsbewertung und damit Reduktion der Publikationsmasse zu befördern.“ WR (2015) – Empfehlungen zu wissenschaftlicher Integrität, S. 32.

[76] Vgl. WR (2018) – Internationalisierung von Hochschulen.

und-organisationen zur Forderung nach „Offenheit“ von Daten positionieren, wenn zugänglich gemachte Daten in anderen Ländern in eine kommerzielle Nutzung überführt, dort eventuell angeeignet und/oder angereichert werden und anschließend nicht mehr offen zur Verfügung stehen?[77] Wie schützt man Forschungsdaten, die frei zirkulieren, überhaupt wirksam vor Manipulation und Sabotage sowie umgekehrt die Wissenschaftlerinnen und Wissenschaftler vor der versehentlichen Rezeption von digital erzeugten Falschinformationen? Das „Open“-Paradigma kann hier selbst in eine Krise geraten, denn es setzt zweierlei voraus:

- eine methodisch seriöse Haltung aufseiten der Nutzerinnen und Nutzer von Forschungsdaten und damit eine weltweit gleiche und womöglich idealisierte Norm von „Wissenschaftlichkeit“,

- vergleichbare (und auf internationale Anschlussfähigkeit angelegte) nationale Anstrengungen zum Infrastrukturaufbau für tatsächlich *qualitätsgeprüfte* „offene“ Daten.

Asymmetrie im Zugang zu Forschungsdaten

Die gegenwärtig vorherrschende, am Output orientierte Form wissenschaftlicher Konkurrenz kann hingegen zu Asymmetrien im Zugang zu qualitativ hochwertigen Forschungsdaten führen, wenn sich einzelne zunächst einen Wettbewerbsvorteil verschaffen, indem sie die Offenlegung von Daten anderer für sich nutzen, selbst aber entweder gar nicht oder nur vordergründig das Leitbild der „Openness“ bedienen.

Forderung nach Datensouveränität

Darüber hinaus kann eine wirtschaftliche Nutzung von „offenen“ Forschungsdaten unter der Bedingung internationaler Konkurrenzlagen zu Wettbewerbsverzerrungen führen: Einzelne Nationen nutzen die zur Verfügung gestellten Forschungsdaten, halten ihre eigenen aber unter Verschluss. Im Zusammenhang mit dem Aufbau der EOSC wird dieses Problem diskutiert. Für Deutschland wurde in diesem Zusammenhang „Datensouveränität“ gefordert; der RfII hat dieses Stichwort vor allem im Sinne einer Souveränität des deutschen Wissenschaftssystems hinsichtlich seiner eigenen Forschungsdaten aufgegriffen.[78] In seiner Stellungnahme zu den aktuellen Entwicklungen rund um Open Data und Open Access fordert er politische Strategien für ein Ordnungssystem ein, das den Wissenschaftlerinnen und Wissenschaftlern, aber auch den Wirtschaftsakteuren Handlungssicherheit bezüglich des Teilens und Verwertens von Forschungsdaten bietet.

[77] Zu prüfen ist, ob sich dem mit entsprechenden Lizenzen wirksam begegnen lässt oder dies andere „Nebenwirkungen“ hat – siehe Creative Commons Lizenz CC-BY-NC, https://creativecommons.org/licenses/by-nc/3.0/de/ (zuletzt geprüft am: 30.08.2019).

[78] RfII (2016) – Leistung aus Vielfalt, S. 35.

3.2.1 UNKLARE REGULATORISCHE RAHMENBEDINGUNGEN VON „OPENNESS“

Suche nach rechtskonformen Standards

Zahlreiche rechtliche Regelungen werden derzeit an die neuen Bedingungen einer digitalisierten Welt angepasst. Für Deutschland und Europa wird von forschungs- und wirtschaftspolitischer Seite „Offenheit“ von in öffentlicher Trägerschaft erstellten Datenbeständen angestrebt. Damit steht auch die Wissenschaft neuen Fragen der Kenntnis und des Ausweises der Daten-Provenienz, der Urheberschaft, der Sicherheit und des Schutzes von Daten (sowie Forderungen nach „Besitz“ oder gar Eigentumsrechten an Daten) gegenüber. Die Suche nach Standards, die auch rechtlichen Ansprüchen genügen, beginnt für Forschende bereits mit der Frage, wie sie ihre wissenschaftliche Leistung der Erhebung und Kuratierung von Datensätzen auch mit Blick auf eine langfristige Sicherung dokumentieren können und müssen (vgl. 2.1.5). Wann werden in der Nutzung anderer Daten Rechte verletzt und wie kann man eigene Daten vor missbräuchlicher oder schädigender Nutzung schützen? Die Freigabe von Daten hat auch komplexe rechtliche Implikationen. Da digitale Daten multipel verwendbar sind, ist das Bild der „Publikation“ womöglich irreleitend.[79] „Nutzung“ kann vieles sein. Wer darf dann wie über die Daten verfügen?[80] Auch die generelle Unsicherheit eines rechtskonformen Umgangs mit Daten wird als Hürde für gute wissenschaftliche Praxis anerkannt. Ein echtes Krisenphänomen liegt insofern vor, als Forschende mangels greifbarer rechtlicher Konturen ihre Daten nicht selten beliebig weitergeben oder auch verloren gehen lassen und sich insbesondere bei der Nutzung kommerzieller (kostenloser oder kostengünstiger) Datendienste um das „Kleingedruckte“ gar nicht erst kümmern. Auf diese Weise fließen Forschungsdaten in die Rechtssphäre großer Datenkonzerne sowie in andere Zonen unkontrollierter Nachnutzung ab.

Forschungsdaten als Regulierungsgegenstand

Konkret fassbar wird der Einfluss externer Regulierungen auf das Forschungshandeln auch in einschlägigen Daten-Policies der Europäischen Union. Gemäß der im Juni 2019 beschlossenen PSI-Richtlinie[81] zählen erstmals auch Forschungsdaten zu denjenigen Daten des öffentlichen Sektors, die prinzipiell diskriminierungsfrei und unter transparenten Bedingungen weiterverwendbar sein sollen. Ziel ist es, bereits in Repositorien öffentlich verfügbare Forschungsdaten aus staatlich finanzierter Forschung für eine – insbesondere gewerbliche – Weiterverwendung nutzbar zu machen. Darüber hinaus enthält die Richtlinie jedoch auch die Vorgabe für die Mitgliedstaaten, eine eigene Strategie für den offenen Zugang zu Forschungsdaten vorzulegen. Der RfII hat in einer Stellungnahme die damit verbundene Absicht begrüßt, eine Harmonisierung im Bereich der in Europa mittlerweile gängigen Forschungsdatenrepositorien zu

79 Parsons & Fox (2013) – Is Data Publication the Right Metaphor?
80 Vgl. Lauber-Rönsberg et al. (2018) – Rechtliche Rahmenbedingungen FDM.
81 EU (2019) – Neufassung PSI-Richtlinie 2019/1024/EU.

erreichen.[82] Der RfII empfiehlt, anstelle eines quantitativen Wachstums vor allem das qualitative Ziel hochwertiger Datenbestände und Datendienste zu verfolgen. Undifferenzierte „Publizitätspflichten" – etwa auch für alle vorläufigen Zwischenprodukte auf dem Weg zu einem Ergebnis – würden dem spezifischen Leistungsanspruch und der Qualitätsverantwortung der Wissenschaft nicht entsprechen. „Openness" kann in diesem Sinne kein Selbstzweck sein.

3.2.2 ABHÄNGIGKEIT VON KOMMERZIELLEN PRODUKTEN UND DIENSTEN

Auch die zunehmende Abhängigkeit des wissenschaftlichen Produktionsprozesses von digitalen Dienstleistungen, Infrastrukturprodukten und Werkzeugen, die in der Regel von kommerziellen Anbietern stammen, treibt den Diskurs um Datenqualität voran. Von einer Krise durch *Blackboxing* ist zwar öffentlich (noch) nicht die Rede. Der digitale Wandel gibt aber dem Faktum, dass Forschung selbstverständlich vielfach eng mit den Herstellern von Spezialgeräten zusammenarbeiten muss, ein neues und verändertes Gewicht.

Zyklen des Veraltens von Hard- und Software beschleunigen sich

Der in den Natur- und Ingenieurwissenschaften immer schon bekannte Zyklus des Veraltens von Forschungsgeräten wird von dem viel schnelllebigeren Zyklus des Veraltens von Software deutlich übertroffen. Ebenso wird Software nur zum Teil für die wissenschaftliche Nutzung geschaffen beziehungsweise optimiert. Evolutionsprozesse im Softwarebereich tragen zudem oftmals komplexen Verflechtungen ganzer Softwarewelten Rechnung; Unternehmensstrategien sind entsprechend volatil und für die Wissenschaft als Kunden kaum durchschaubar. Dies hat weitreichende Implikationen zum Beispiel für die oben beschriebenen Schwierigkeiten bei der Replikation beziehungsweise Reproduzierbarkeit von Studien und Experimenten (vgl. 3.1.1). Oftmals führt die Verwendung unterschiedlicher Software-Versionen auf den gleichen Geräten zu unterschiedlichen Ergebnissen. Daher werden in vielen Einrichtungen mit einigem Aufwand Altsysteme weiterbetrieben (zu Hard- und Software-Problematik s. a. 2.1.8).

Aushandlung von Nutzungskonditionen als kollektive Aufgabe der Wissenschaft

Zur Dokumentation von Forschungsdatensätzen gehören immer auch Angaben zu benutzten Geräte- und Softwareversionen – und zwar für alle Verarbeitungsstadien im Datenlebenszyklus. Diese Anforderung ist in der Praxis kaum zu erfüllen, wenn eine Abhängigkeit von der Dokumentation der Anbieter besteht. Ein (zum Beispiel für die Wissenschaft zu Forschungszwecken eröffneter) Zugang zu kommerzieller Software ist selten. Im Idealfall ist verwendete Software nach wissenschaftlichen Kriterien gestaltet und publiziert – was vor allem dann der Fall ist, wenn das betreffende Produkt in enger Kooperation mit Forschenden oder durch die Wissenschaft selbst entwickelt wurde. Zu lindern wäre das Problem auch durch die Aushandlung

82 RfII (2019) – Stellungnahme aktuelle Entwicklungen Open Data.

besserer Nutzungskonditionen für die Produkte von kommerziellen Partnern. Beides ist jedoch eine kollektive Aufgabe; Individualforschende verfügen im Forschungsalltag oftmals weder über Zeit noch Ressourcen dafür.[83] Gelegentlich fehlt auch das Bewusstsein, dass eine Minimierung von Abhängigkeiten (und, wo dies nicht möglich ist, deren Dokumentation) einen wichtigen Beitrag zur guten wissenschaftlichen Praxis leistet.

Problemfeld Datensicherheit

Ein weiteres Problem in diesem Zusammenhang betrifft die Datensicherheit. Stellen externe Anbieter bestimmte Betriebssysteme und Applikationen beziehungsweise deren Wartung ein, schlägt dies auch auf Computer und internetfähige (Mess-)Geräte im Forschungsbereich durch. Ist aus welchen Gründen auch immer zum Beispiel kein Update auf eine neuere Windows-Version möglich, stehen viele Forschungseinrichtungen vor dem Problem, dass sie internetfähige Geräte vom Online-Zugriff abkoppeln oder in eigens geschützte Netze überführen müssen. Solch aufwendige (Not-)Lösungen stellen gerade kleinere Einrichtungen vor erhebliche Herausforderungen.

Risiko von Datenverlusten

Die beschriebenen Engpässe in der IT-Ausstattung und im IT-Management betreffen auch und insbesondere die Langzeitarchivierung. Ohne eine vorausschauende Strategie riskieren gerade kleinere Forschungseinrichtungen oder Lehrstühle nicht selten vollständige Datenverluste, wenn Daten auf veralteten Speichermedien gelagert werden. Ressourcen und Personal sind bislang allenfalls bei Großforschungs- beziehungsweise Infrastruktureinrichtungen vorhanden. Im Rahmen des Aufbaus der NFDI wird die Problematik der digitalen Langzeitarchivierung sicherlich adressiert werden können. Eine Inhouse-Lösung ist nicht zuletzt aus technischen Gründen kaum in Sicht. Auch angesichts des Kompetenzerwerbs hierfür nötigen Personals besteht dringender Handlungsbedarf.

3.2.3 PREKÄRE FINANZIERUNGSPERSPEKTIVEN VON DIENSTEN

Enge finanzielle Spielräume für Langzeitsicherung

Eine kaum im Licht der Öffentlichkeit stehende Rahmenbedingung für die Produktion, Verarbeitung und vor allem Langzeitsicherung qualitativ hochwertiger Forschungsdaten war bislang deren öffentliche Finanzierung. Gerade in Deutschland sind die Spielräume eng, um Forschungs- und Informationsinfrastrukturen, die sich aus der Forschung an Universitäten heraus entwickelt haben, als stetige Vorhaben in den Hochschulen selbst weiterzuentwickeln und zu finanzieren.[84] Auch an den Universitätsbibliotheken sind jüngere Ansätze in der

[83] Fallbeispiele für solche Abhängigkeiten gibt es viele, nicht zuletzt auch, was den Zugang zu Daten kommerzieller (oder behördlicher) Anbieter angeht. Zugänge für die Wissenschaft erfordern oftmals langwierige Verhandlungen.

[84] RfII (2016) – Leistung aus Vielfalt, S. 37 ff., Empfehlung 4.1.

Infrastrukturentwicklung in hohem Maße von zumeist befristeter Projektförderung abhängig. Mehr Möglichkeiten für dauerhaften Auf- und Ausbau bieten sich bislang im Rahmen der großen Wissenschaftsorganisationen und ihren Infrastruktureinrichtungen, die mit längerem Atem internationale Standards vorantreiben und deren gesicherte Planungshorizonte gute Voraussetzungen für langfristig angelegte Qualitätssicherung bieten.

Schwierigkeiten, Forschungs- und Informationsinfrastrukturen an den Hochschulen zu halten

Dort, wo beispielhafte Informationsinfrastrukturen mit exzellenter Datenqualität an den Hochschulen aufgebaut werden konnten – von epidemiologischen Studien und Krebsregistern bis hin zu großen sozialwissenschaftlichen Umfragestudien und sprachwissenschaftlichen Datenbanken –, war nach langjähriger wechselhafter Projektförderung häufig der Wechsel in eine außeruniversitäre Verstetigung das einzige Mittel zur Strukturerhaltung. An den zahlreichen Bruchstellen in diesem Prozess (markiert durch die jeweiligen Projektförderenden) ist immer wieder wichtiges Personal und damit auch wissenschaftliche Daten- und Infrastrukturkompetenz weggebrochen. Diese erratischen Prozesse sind in der Wissenschaft oft als nicht zweckdienlich thematisiert worden, wenn es darum geht, rund um Forschungsdaten dauerhafte forschungsnahe Dienste aufzubauen und zu erhalten – auch und gerade im Bereich der „kleinen Fächer", die sich nicht umstandslos in außeruniversitäre Institute überleiten lassen. Gerade kleinere, in der Forschung entstandene Datensammlungen, sind nicht selten vollständig verloren gegangen, wenn auf einer Professur die Stelleninhaberin beziehungsweise der Stelleninhaber wechselte. Zwar haben sich in den vergangenen Jahren im Gefolge des europäischen ESFRI-Prozesses auch im Bereich der Geistes- und Sozialwissenschaften zahlreiche Verbesserungen ergeben. Eine kohärente Politik zeichnet sich bislang jedoch nicht ab, um hervorragende Forschungs- und Informationsinfrastrukturen auch an den Hochschulen halten zu können. Die Etablierung der NFDI ist ein erster Schritt, um projektförmig finanzierten und organisierten Diensten zumindest den Weg zu weisen, wie sie sich mittelfristig an institutionell bereits gesicherte Zusammenhänge annähern oder sich gar in diese integrieren könnten. Die Hochschulen haben sich zuletzt sehr offen gezeigt, diesen Weg mitzugehen und zur Ausgestaltung der NFDI aktiv beizutragen.

3.3 LATENTE PROBLEME DER WISSENSCHAFTSPRAXIS

Es ist wesentlich für die wissenschaftliche Praxis, dass sie fach- und aufgabenspezifisch die Qualität von Basisprozessen wie Protokollierung, Dokumentation, Edition und Repräsentation/Präsentation nicht nur sichert, sondern dies auch reflektiert.[85] Dabei geht es wesentlich um die Nachvollziehbarkeit von Auswertungen, aber

[85] Vgl. Daston/Galison (2007) – Objektivität.

auch um die Anpassung vorhandener Bestände an sich wandelnde Forschungsfragen und Methoden. Ähnlich gelagerte Probleme gibt es quer zu etwaigen Forschungsformen in der nachträglichen Digitalisierung (Retrokonversion) bereits normiert erfasster Daten wie Texte oder Bilder, aber auch Kataloge in Bibliotheken und Archiven. Die Notwendigkeit des Editierens oder Kuratierens solcher Daten – Begriffe, die aus den Kontexten der philologischen Edition oder der Forschung in Museen stammen – treten wieder neu in den Mittelpunkt. Zugleich kommen aber weitere digitalitätstypische Probleme hinzu:

- Heterogenität und Kurzlebigkeit verwendeter Werkzeuge,
- trotz abstrakter Reflexion die faktische Nichtumsetzung von Kriterien hinsichtlich des Überganges von analogen zu digitalen Verfahren,
- eine fehlende Dokumentationskultur für „digitalitätsgetriebene“ (Zusatz-) Entscheidungen,
- ungeklärter Methodenbezug neu eingeführter Werkzeuge (die teils nicht für die Wissenschaft, sondern zu anderen Zwecken entwickelt und lediglich übernommen wurden),[86]
- eine unvermeidliche Verlustseite von digitaler Aufbewahrung (etwa entfallen materialgebundene Gebrauchsspuren als Informationsträger).

Kooperation zwischen fachlicher Forschung und wissenschaftlichen Bibliotheken notwendig

Ebenso hat sich der Aufwand, der nötig ist, um zum Beispiel digitalisierte Sammlungen bei professioneller Langzeitarchivierung über längere Zeit hinweg nutzbar zu halten, als hoch erwiesen. Namentlich Bibliotheken und Archive benötigen hierfür jedoch nicht nur Ressourcen, sondern dringend auch informierte Zuarbeiten seitens der Wissenschaft, mit denen sie eine dauerhafte Bearbeitung dieses Aufgabenfelds sicherstellen können.[87] Eine nachhaltige Pflege würde zum Beispiel voraussetzen, dass relevante objektbezogene Metadaten, bereits im Forschungsprozess systematisch erstellt werden und nicht nachträglich eingefordert werden müssen, damit dieses Aufgabenspektrum nicht als krisenhafte Überforderung erlebt wird. Insbesondere die wissenschaftlichen Bibliotheken sehen sich in diesem Zusammenhang nicht lediglich als Zulieferer von Fachinformationen. Vielmehr definieren sie ihre Rolle zunehmend auch als Treiber in der digitalen Transformation der Wissenschaft, der in der Sicherung und Steigerung von Datenqualität insbesondere im Bereich der Metadatenerstellung und -pflege eine aktive Rolle zukommt. Dies beinhaltet den Wunsch nach einem erkennbaren Interesse an Kooperationen auch aufseiten der Forscherinnen und Forscher.[88]

---

[86] Ein Beispiel sind hier die Verwendung von Suchmaschinen- oder Recommender-Software im Katalogbereich oder auch Text- und Bilddarstellungen, die für portable Endgeräte optimiert sind.

[87] Der RfII hat hierzu empfohlen, Infrastrukturbereiche und Forschung auch personell enger zu verzahnen. Siehe RfII (2019) – Digitale Kompetenzen, S. 27 f., insbesondere Empfehlung 4.5.

[88] Siehe hierzu DBV (2018) – Positionspapier Wissenschaftliche Bibliotheken 2025.

Neben generischen Herausforderungen, vor denen alle Disziplinen und die in ihnen praktizierten Forschungsformen in nahezu gleichem Maße stehen, gibt es auch je nach Forschungsform spezifische Herausforderungen für die Datenqualität. Aus welchen Konstellationen heraus sich diese entwickelt haben, an welcher Stelle im Datenlebenszyklus sie sich am dringlichsten stellen und mit welchen Instrumenten oder institutionellen Vorkehrungen Lösungen gesucht werden, ist dabei sehr unterschiedlich.

3.3.1 HERMENEUTISCH-INTERPRETIERENDE FORSCHUNGSFORMEN

Gesteigertes Interesse an „klassischen" Datenbeständen

In den Geisteswissenschaften kultiviert die „Philologie" in einer bis in vormoderne Zeiten zurückreichenden Geschichte unter anderem Editionen von Texten und Bildern in „Korpuswerken" und Katalogen oder auch Wörterbüchern. Im 19. Jahrhundert entstanden große Editionsprojekte. Diese als Langzeitunternehmen durchgeführten Vorhaben gerieten in die Kritik, als mit der Digitalisierung der schnelle und unbeschränkte Zugriff auf Informationen möglich wurde. Editionsprojekte mit bis in das 19. Jahrhundert zurückreichenden Traditionen schienen nicht mehr aktuellen Vorstellungen zu entsprechen und der Ressourceneinsatz für diese Projekte wurde infrage gestellt. Das Ringen um eine sinnvolle Verbindung zwischen dem traditionellen Umgang mit dem Material und der Nutzung neuer digitaler Technologien ist jedoch nicht auf Editionen von Texten und Bildern beschränkt, sondern betrifft auch technisches Wissen (etwa um Baumaterialien, den Instrumentenbau oder klassische Wege der Farbherstellung) oder die Kartografie (historische Luftaufnahmen einschließlich). Qualitätssicherung unter den Bedingungen von Digitalität setzt daher Reflexion über Basisprozesse und Standards sowie Fragen der systematischen (Neu-)Erschließung auch des „analogen" Wissens voraus. Mit dem Interesse an digitalen Datensammlungen geht daher nicht selten ein teils gesteigertes Interesse an (zu verknüpfenden) klassischen Datenbeständen einher. Für naturwissenschaftliche und medizinische Sammlungen ist nicht zuletzt im wissenschaftshistorischen Kontext und im Museumsbereich eine Methodenreflexion prägend, die in der Kombination von „analog" und „digital" ein besonderes Qualitätsmerkmal sieht.[89]

Erste Digitalisierungswelle war noch nicht nachhaltig konzipiert

Mit der Digitalisierung ging zunächst jedoch kein „neuer" Qualitätsdiskurs einher. Digitale und automatisierte Verfahren der Dokumentation und Präsentation sah man in den 1990er Jahren vielmehr als Chance für mehr Quantität. Sie schienen

[89] Vgl. zum Beispiel Aktivitäten der Fachgruppe Dokumentation des Deutschen Museumsbunds: https://www.museumsbund.de/fachgruppen-und-arbeitskreise/fachgruppe-dokumentation/arbeitsgebiete/ oder der Koordinierungsstelle für wissenschaftliche Universitätssammlungen: https://wissenschaftliche-sammlungen.de/de/ (beide zuletzt geprüft am: 30.08.2019).

schlicht Wege aufzuzeigen, um weitaus größere Mengen an Informationen als in den traditionellen Verfahren systematisch zu erschließen. Kosten-Leistungsverhältnisse für das Zugänglichmachen schienen weitaus besser zu sein. Ebenfalls beeindruckte die Erleichterung des Datentransfers. Gesteigerte Geschwindigkeiten und Datenvolumina schienen also entscheidende Effekte von Digitalität zu sein. Die „Digitalisierungswelle" der ersten Jahre hat daher in vielen Disziplinen – und insbesondere in den Geistes- und Kulturwissenschaften – zu einer latenten Qualitätskrise geführt: Viel wurde investiert, um Daten schnell zugänglich zu machen – allerdings häufig in Formaten, die ihre langfristige Verwertbarkeit beeinträchtigen. Die anfängliche Euphorie ist so der Erkenntnis gewichen, dass auch für digitale Daten letztlich die fachlich-wissenschaftliche Qualität entscheidend ist. Diese Qualität wiederum muss durch konzeptionelle Ansätze und redaktionelle Eingriffe aktiv sichergestellt werden.

3.3.2 BEOBACHTENDE UND EXPERIMENTIERENDE FORSCHUNGSFORMEN SOWIE SIMULATION

Bereits früh Fokus auf den Aufbau von Datenbanken und Repositorien

In einigen Teildisziplinen der Naturwissenschaften wurden bereits früh Datenbanken oder -repositorien für die Abgabe von Forschungsdaten eingerichtet, zum Beispiel in den Erd- und Umweltwissenschaften, der Genomforschung oder in der Medizin. Zu nennen ist hier das World Data System mit seinen bereits in den 1950er Jahren gegründeten Vorläufern, den „World Data Centres".[90] In den Lebenswissenschaften gilt seit Mitte der 1990er Jahren die Übereinkunft, neue Gensequenzen möglichst binnen 24 Stunden in einem der drei weltweiten Datenrepositorien zu veröffentlichen. Die Cochrane-Zentren für evidenzbasierte Medizin wirken ebenfalls seit Mitte der 1990er Jahre mit Mitgliedern aus über 130 Ländern daran mit, aus wissenschaftlichen Studien generierte Gesundheitsinformationen zu erstellen und zugänglich zu machen, die frei sind von kommerzieller Förderung (zum Beispiel durch die pharmazeutische Industrie).[91]

Zeitlicher Publikationsdruck vs. Aufwand für die Datenaufbereitung zur Nachnutzung

Die wissenschaftliche Publikationskultur bleibt jedoch auch in den Naturwissenschaften überwiegend davon geprägt, Ergebnisse in Artikelform schnell in den Forschungsdiskurs einzuspeisen. Diese Praxis erlaubt es in der Regel nicht, die zugrunde liegenden Daten ausführlich vorzulegen; allein eine Auswahl bereits prozessierter Daten prägt den Diskurs und dient als Referenz (vgl. 2.1.9). Die dem Analyseprozess zugrunde liegenden Daten stehen auf diese Weise kaum für andere Forschungsfragen zur Verfügung, und selbst dort, wo Daten publiziert werden, bestehen erhebliche Hürden für weitere Studien. So liegen Daten

[90] Vgl. ICSU- World Data System, http://www.icsu-wds.org/organization (zuletzt geprüft am: 30.08.2019).

[91] Vgl. Website der Cochrane Foundation, https://www.cochrane.de/de/cochrane (zuletzt geprüft am: 30.08.2019).

– etwa in der Chemie – zwar unter Verweis auf die internationalen Normierungskomitees standardisiert aber zum Beispiel nicht maschinenlesbar vor und müssen bei Bedarf händisch, also mit hohem personellem Aufwand aus Publikationen extrahiert werden (s. u.). Sind sie in Datenbanken publiziert, so sind sie oftmals nicht im heute wünschenswerten Umfang mit Kontextinformationen (Metadaten) versehen – diese werden dann, wiederum mit hohem Aufwand, gegebenenfalls den zugehörigen Publikationen entnommen oder bei der durchführenden Institution erbeten. Nicht immer ist allerdings klar, wer die zugrunde liegenden Daten archiviert und Zugang gewähren darf. Auf diese Weise bewegt sich die Forschungsdiskussion unter Umständen – trotz intensiver Prozessierung von Daten – auf einer Ebene, die von den ursprünglich erhobenen Daten von vornherein in (zu) hohem Maße abstrahiert.

Regelung der Zugriffsmöglichkeiten auf Daten zunächst prioritär

In den Sozial- und Wirtschaftswissenschaften erfolgte die Einrichtung zentraler Datenbanken beziehungsweise Repositorien für statistisch relevante Daten zur Gesellschaftsentwicklung deutlich später. Hier lag das Ausgangsproblem darin, dass Daten, die von öffentlichen Einrichtungen wie statistischen Ämtern und Sozialversicherungsträgern gesammelt wurden, der Forschung überhaupt erst zugänglich gemacht werden mussten. Das gleiche Prinzip wurde – getrieben durch die Möglichkeiten des digitalen Fernzugriffs – auch für Erhebungsdaten aus großen Paneluntersuchungen ermöglicht, die bis dato nur „vor Ort“ und unter zum Teil hohen Nutzungsauflagen anderen Forscherinnen und Forschern von den erhebenden Personen beziehungsweise Einrichtungen zur Verfügung gestellt wurden. Auch hier war zunächst das Motiv der Öffnung beziehungsweise des Zugangs einer intensiven Qualitätsprüfung der Daten und Datensätze vorgängig. Dort, wo empirische Daten heute in Forschungsdatenzentren – vor allem in außeruniversitären Einrichtungen – zur Verfügung gestellt werden, hat sich der Fokus von Zugriff auf aktive Qualitätssicherung im Zuge der Aufbereitung für die Forschungsinteressen Dritter bereits stark gewandelt. Dieser Wandel beinhaltet auch zunehmenden Einfluss von gewählten Repräsentanten aus den Fachverbänden auf die Ausgestaltung der Erhebungssamples (zum Beispiel Fragebogengestaltung, Sonderthemen für Teilpopulationen im Rahmen von Panels mit konstanten Erhebungswellen) – und damit ganz allgemein auf die Frage, welchen Ausschnitt der Realität die Primärdaten repräsentieren sollen.

Standards für maschinenlesbare Daten notwendig

In Teilen der Naturwissenschaften haben seit längerem die internationalen Fachvereinigungen eine wichtige Rolle, die Standards und Nomenklaturen für die Verwendung fachzentraler Begriffe und Messwerte in Tabellen und Formeln in eigens hierfür eingerichteten Komitees festlegen, die sich durch eine hohe Akzeptanz in der Forschungspraxis auszeichnen. Vereinigungen wie die International Union of Pure and Applied Chemistry (IUPAC) und International Union of Pure and Applied Physics (IUPAP) reagieren allerdings inzwischen auch auf den Umstand, dass standardisierte Beschreibungsformen für Daten noch keine hinreichende Grundlage für nachnutzbare Daten sind. Hierfür bedarf es hinreichender Standards für

maschinenlesbare Formel- und Datensätze, die im Zweifel eine schnelle Nachvollziehbarkeit und damit Replizierbarkeit eines Forschungsexperiments ermöglichen. Auch wird in Ergänzung zu den bereits etablierten Referenzdatenbanken – in denen lediglich eine Auswahl geprüfter Daten zu Validierungszwecken vorgehalten wird – die Einrichtung von Datenrepositorien diskutiert, um Forschungsergebnisse offen bereitzustellen.[92]

Innovative Formen der Forschungsdokumentation gefragt

Nur schwer lässt sich in der experimentellen Forschung die klassische qualitätssichernde (und Garanten-)Funktion des Laborbuchs substituieren. Ähnliches gilt beispielsweise in den Altertumswissenschaften für Grabungsbücher oder für Forschungstagebücher über teilnehmende Beobachtung in der ethnologischen und sozialanthropologischen Feldforschung. Reproduzierbarkeit von Ergebnissen ist ohne einen Nachvollzug des Experiments oder der Datenerhebung selbst freilich nicht oder nur bedingt möglich. Nicht nur den experimentell arbeitenden Wissenschaften fehlen hier teils an digitale Verfahren angepasste Werkzeuge für eine geeignete Dokumentation. „Elektronische Laborbücher" sind zwar (als kommerzielle Softwareprodukte) erhältlich, konkurrieren aber miteinander und sind dem fachspezifischen Bedarf nicht immer angepasst. Auch können Laborbücher Arbeitsschritte, die teils vollautomatisch erfolgen, nicht angemessen dokumentieren. Dies gilt analog auch für die oben angeführten Dokumentationsformen im kultur- und sozialwissenschaftlichen Spektrum.

3.3.3 DATENKURATIERUNG FÜR DIE NUTZUNG ÜBER FACHGRENZEN UND DOMÄNEN HINAUS

Qualitätssicherung auch für außerwissenschaftliche Zielgruppen notwendig

Mit dem öffentlichen Zugang zu Forschungsdaten oder der „Offenheit" von wissenschaftlichen Datenbeständen (vgl. 3.1.3) verbindet sich in verstärktem Maße die Erwartung einer Qualitätssicherung, die auch interessierten Wissenschaftlerinnen und Wissenschaftlern anderer Disziplinen und externen Fachkräften zum Beispiel aus dem Bereich des Wissenschaftsjournalismus Daten nutzbar macht. Inwiefern für diese erweiterten Zielgruppen eine wissenschaftlich wünschenswerte Qualitätsprüfung in der dafür geforderten Breite erfolgen kann, ist derzeit noch nicht absehbar. Die Notwendigkeit des Editierens oder Kuratierens von Daten bestreitet auch in den Natur- oder Ingenieurwissenschaften niemand. Die Wege zur Erschließung von Daten für einen Nutzerkreis, der über die Domäne hinausreicht, sind jedoch schwierig und ressourcenintensiv – auch wenn sich das Ergebnis lohnen kann, wie das Beispiel der Satellitenbeobachtung des Polareises zeigt: Waren die diesbezüglichen Beobachtungsdaten in den 1980er Jahren nur einem kleinen Kreis von Fachleuten zugänglich, so arbeitete ab den 1990er

[92] Koepler et al. (Aug 2018) – Thesenpapier NFDI4Chem.

Jahren auch die Klimaforschung und später die Biodiversitätsforschung mit diesen Daten – dies dann in für diese Zwecke aggregierter beziehungsweise modellierter Form, wie zum Beispiel Zeitreihen. Heute stehen auch für Journalisten und die interessierte Öffentlichkeit spezielle Datenaufbereitungen zur Verfügung.[93]

Vergleichbarkeit und Anschlussfähigkeit von Datenveröffentlichungen

Fortschritte im Bereich der fachnahen Datenkuratierung setzen Zeit und Ressourcen voraus, um entsprechende Grundsätze in den Fachgemeinschaften auszuhandeln und sie in den Hochschulen und Forschungseinrichtungen zu operationalisieren und zu implementieren. Vielerorts fehlt es schlicht an unterstützendem Fachpersonal für die Dokumentation digitaler Forschungsprozesse.[94] Hinzu kommt die Aufgabe, Formen der Datenveröffentlichung zu entwickeln, die auch von Forschenden außerhalb der eigenen Disziplin sinnvoll genutzt werden können. Durch digitale Verfahren werden zudem existierende fachwissenschaftliche Standards (teils durch kommerzielle Angebote) infrage gestellt. So tauchen unter dem Stichwort „Semantic Web" konkurrierende Vorschläge für Verschlagwortungen als Ersatz zum „kaskadierenden Katalogwesen" auf. Dennoch werden auch für avancierte Textmining- und Mustererkennungsverfahren solche kontrollierten Vokabulare oder Thesauri (weiter-)genutzt (vgl. 1.2.1), die in der analogen Welt entwickelt worden sind. Dies ist nötig, um Vergleichbarkeit und Anschlussfähigkeit zu garantieren. Vergleichbarkeit und Anschlussfähigkeit sind dabei auf mehreren Ebenen gefordert und machen die Problematik der Erzeugung von Datenqualität ebenso dringlich wie voraussetzungsvoll. Große Herausforderungen für die Erschließung von Forschungsdaten stellt zum Beispiel die stimmige Verzahnung von „alten" und „neuen" Daten dar. Wie geschildert, bereitet unter anderem das Veralten von Speichermedien große Probleme für die Kontinuität von Langzeitstudien. Auch der Transfer und die Verknüpfung von Daten nicht nur über Fächergrenzen, sondern auch über die engere Sphäre des Wissenschaftssystems hinweg stellen enorme Qualitätsansprüche. So ist es beispielsweise zur Erforschung der großen Volkskrankheiten notwendig, Gesundheitsdaten aus dem medizinischen Bereich, mit Daten aus sozioökonomischen Panelstudien und mit georeferenzierten Daten zu verknüpfen, um zu Aussagen über Krankheits- und Verlaufsmuster für unterschiedliche Populationen zu kommen. Nicht zuletzt sind die Vergleichbarkeit und Anschlussfähigkeit von Daten aus unterschiedlichsten Forschungsquellen essenziell für die Trainingssets von Prozessen maschinellen Lernens, also für das Lernen von künstlicher Intelligenz (KI). Qualitativ „schlechte" beziehungsweise nicht verknüpfbare Datensätze können hier notwendigerweise nicht zu validen Lernergebnissen führen.

93 Baker et al. (2015) – Scientific Knowledge Mobilization.

94 Der RfII hat sich dazu 2019 in seinen Empfehlungen zu Berufs- und Ausbildungsperspektiven geäußert, vgl. RfII (2019) – Digitale Kompetenzen.

4 EMPFEHLUNGEN ZUR WEITERENTWICKLUNG VON DATENQUALITÄT IN DER WISSENSCHAFT

Definitionen und die Verwendungen des Begriffs „Datenqualität“ sind durch eine große Heterogenität und Diversität gekennzeichnet. Das Wort hat einerseits einen technischen, einer kleinteiligen Normierung zugänglichen Sinn; andererseits tangiert es aber gleichermaßen das Ethos jeglichen wissenschaftlichen Handelns. Die Entwicklung von Forschungsdateninfrastrukturen verlangt eine Verständigung über die zu speichernden, zu übertragenden und zu verarbeitenden Daten – sowohl hinsichtlich der für ihre Generierung verwendeten Methoden als auch ihrer Typen und Formen. Dies betrifft Anforderungen unter anderem an die Auffindbarkeit, Verwendbarkeit und Zuordnungsfähigkeit ebenso wie Anforderungen an die Struktur beziehungsweise das Format sowie an die Qualität der Daten selbst in einem weiten Sinne. Zugleich tritt genau in einer solchen Konkretisierung die Schwierigkeit hervor, zur Beschreibung der Anforderungen einen überzeugenden (akzeptablen), brauchbaren (funktionalen), hinreichend weiten (umfassenden) und zugleich auch im Detail aussagekräftigen Begriff von Datenqualität zu gewinnen, der angesichts des rapiden technologischen und wissenschaftlichen Wandels zumindest kurz- bis mittelfristig allgemeinen Konsens ermöglichen könnte.

4.1 FÜR EINEN DYNAMISCHEN UND PROZESSBEZOGENEN BEGRIFF VON DATENQUALITÄT

Mehrdimensionales Verständnis von Datenqualität gefordert

Die hier gegebenen Empfehlungen des RfII setzen ein mehrdimensionales Verständnis von Datenqualität voraus, das zum einen – jedoch nicht allein – den Datenlebenszyklus sowie – als zweiten Prozesskreis – den jeweiligen konkreten Forschungsprozess sowie die dabei beteiligten Institutionen und individuellen Forscherinnen und Forscher miteinbezieht (s. Abbildung 3). Von einer einfachen (normativen) Definition des Begriffs Datenqualität nimmt der Rat daher Abstand. Ein solcher Ansatz würde Gefahr laufen, dass Lösungen für die Erstellung und Handhabung von Datenqualität (a) unterkomplex, also enttäuschend, (b) für die konkrete Anwendung unpassend und somit nicht akzeptabel oder aber (c) quasi beliebig weit in die Zukunft verschiebbar erscheinen. Aus wissenschaftspolitischer Perspektive würde dies auf einen nicht abschließbaren Prozess hinauslaufen. Des Weiteren ist zu berücksichtigen, dass der durch Digitalität angetriebene Umbruch auch den Forschungsprozess selbst verändert und damit den maßgeblichen Kontext, aus dem heraus überhaupt erst sinnvoll über Datenqualität gesprochen werden kann. Forschung ist selbst sowohl Treiber als auch Gegenstand der aktuellen Transformation. Stattdessen schlägt der RfII vor, die Erarbeitung eines nur vorläufig bestimmbaren Begriffs von Datenqualität zum Gegenstand des kontinuierlichen Methodendiskurses in allen wissenschaftlichen Communities/Fachgemeinschaften zu machen.

Alle Akteure des Wissenschaftssystems sind angesprochen

Empfehlungen in diesem Bereich richten sich an alle Akteure des Wissenschaftssystems, die den Forschungsprozess und seine Rahmenbedingungen durch Vorgaben und Konsensbildungsprozesse prägen: Neben den individuellen Forscherinnen und Forschern selbst sind dies die Fachgemeinschaften, die Wissenschaftsorganisationen, die Forschungsfördernden, aber auch die Organe der wissenschaftlichen Kommunikation, insbesondere die Fachzeitschriften mit ihren Gutachtern und Gutachterinnen sowie Herausgebergremien. Diese Adressaten müssen aber auch materiell sowohl im Rahmen einer zuverlässigen institutionellen Grundsicherung wie auch durch spezifische Leistungsanreize, die von Reputationssystemen ausgehen, durch eine vorausschauende Wissenschaftspolitik in den Stand versetzt werden, handeln zu können. Es gilt, die zusätzlichen Anstrengungen bewältigen zu können, die im Zuge der Digitalisierung in vielen Disziplinen nötig sind, um die Flut der erzeugten und verfügbaren Daten über einzelne Gruppen sowie Fach- und Domänengrenzen hinaus auf hohem Niveau nutzen und sichern zu können.

4.1.1 DOKUMENTATION ALS KERNELEMENT BEGREIFEN

Offenlegung der Verfahrensschritte und Analyseinstrumente in der Forschung notwendig

Digitalität erfordert es, Daten nicht nur zu „nutzen“ beziehungsweise Werkzeuge „auf“ Daten einzusetzen, sondern die Qualitätssicherung der Daten (in der jeweils fachlich gebotenen Weise) als Teil des Forschungsprozesses anzusehen. Kernelement eines hinreichend dynamischen Verständnisses von Datenqualität ist nach Ansicht des RfII die genaue Dokumentation und Offenlegung der Maßnahmen, Werkzeuge, der eingesetzten Forschungssoftware und der Verfahrensschritte zur Generierung, Verarbeitung und Bereitstellung der Daten.

4.1.2 WISSENSCHAFTLICHE DATENKULTUR PROZESSBEZOGEN ENTWICKELN

Methoden am gesamten Datenlebenszyklus ausrichten

Der Mehrwert (und das Alleinstellungsmerkmal) einer wissenschaftlichen Datenkultur liegt darin, dass sich ihre methodischen Anstrengungen im Grundsatz auf den gesamten Datenlebenszyklus richten. Daher ist der RfII der Überzeugung, dass wissenschaftliche Informationsinfrastrukturen Datenqualität einzufordern und zu sichern haben, damit sich das Gesamtsystem Wissenschaft tatsächlich auf höchstem Niveau forschungsgetrieben weiterentwickeln kann. Die Herausarbeitung disziplinen- und domänenspezifischer wie auch generischer Verfahren und Benchmarks zur Sicherung wissenschaftlicher Datenqualität ist nicht zuletzt eine Daueraufgabe für die kommende Nationale Forschungsdateninfrastruktur (NFDI) in Deutschland sowie die Aktivitäten in der European Open Science Cloud (EOSC). Erfolgreiche Anstrengungen einiger wissenschaftlicher Communities in der Entwicklung global genutzter Open-Source-Programme können dabei als Beispiel dienen.

4.2 INTEGRATION IN DAS WISSENSCHAFTLICHE METHODENVERSTÄNDNIS

Forscherinnen und Forscher sind die maßgeblichen Akteure, die im jeweils eigenen Interesse eine mit der Qualität ihrer Forschungsdaten unmittelbar verknüpfte gute wissenschaftliche Praxis im Methodenkanon ihrer jeweiligen Disziplinen und Forschungsformen verankern müssen. Die hiermit einhergehende individuelle und kollektive Verantwortung ist ein wesentlicher Bestandteil des wissenschaftlichen „Berufsethos“.[95]

4.2.1 DATENQUALITÄT ALS GRUNDWERT GUTER WISSENSCHAFTLICHER PRAXIS HERAUSSTREICHEN

Wissenschaftliche Daten müssen sich durch transparente Qualitätssicherung auszeichnen

Der RfII empfiehlt, das Thema Datenqualität als unabdingbaren Grundwert guter wissenschaftlicher Praxis noch nachhaltiger als bislang im fachwissenschaftlichen Methodenverständnis zu verankern.

- Akzeptierte methodische Standards und eine qualitätsgesicherte Verarbeitung im laufenden Verarbeitungs- und Erkenntnisprozess verleihen wissenschaftlichen Daten – inmitten der „Datenflut“, die das digitale Zeitalter generell auszeichnet – ihre besondere Validität und damit der Wissenschaft eines ihrer wesentlichen Alleinstellungsmerkmale. Engagement zur Schaffung beziehungsweise Steigerung von Datenqualität betrachtet der RfII in diesem Sinne auch immer als einen Beitrag zur Stärkung des gesellschaftlichen Vertrauens in die Wissenschaft und als ein Stück wissenschaftliche Leistung (vgl. 4.7.1).

- Zu den Mindestanforderungen an alle wissenschaftlichen Communities und Disziplinen, aber auch an die wissenschaftsnah arbeitenden technischen Infrastrukturdienste gehören dabei grundlegende Kenntnisse über die Anforderungen datenschutzrechtlicher Regulierungen der Erzeugung und Verarbeitung von personenbezogenen und weiteren Daten. Die Beachtung der relevanten rechtlichen Vorgaben ist wesentlicher Bestandteil einer hochwertigen Datenerhebungskultur. Hierzu gehört insbesondere aber auch eine Kultur des Explizierens (siehe auch 4.8.3), die sich zunächst als eine Kultur des Explizierens von verbleibenden Unsicherheiten oder auch potenziellen Fehlerquellen in einem Datensatz zeigen sollte. Eine Beachtung solch allgemeiner Mindeststandards würde gerade die Phase der Datenerhebung, die oftmals (und zu Unrecht) als bloße „Technik“ aufgefasst wird, deutlich aufwerten.

[95] Vgl. DFG (2019) – Leitlinien zur Sicherung guter wissenschaftlicher Praxis, S. 9 f., Leitlinie 2.

4.2.2 REPLIKATIONSSTUDIEN FÖRDERN

Replikationsstudien sind ein Anreiz zur Steigerung von Datenqualität

Der RfII empfiehlt den Fachgemeinschaften, disziplin- und forschungsfeldspezifische Gütekriterien für Datenqualität zu entwickeln, die diese – dort, wo dies noch nicht der Fall ist – stärker in das jeweilige fachliche Methodenverständnis zu integrieren. Dies schließt – dort, wo der Forschungsprozess dies zulässt beziehungsweise die Fachkultur dies gebietet – die gezielte Förderung von Replikationsstudien ein, die geeignet wären, die Validität von Daten, Datensätzen und „Datenprodukten" (siehe Empfehlung 4.4) zu sichern. Replikationsstudien könnten so einerseits einen direkten Anreiz zur Steigerung von Datenqualität und hiermit zusammenhängender ausführlicher Dokumentation von Forschungsdaten bieten. Andererseits würde ihre Aufwertung auch die Internalisierung und Auffrischung der Grundregeln guter wissenschaftlicher Praxis in allen Phasen einer wissenschaftlichen Karriere befördern. Nicht zuletzt sollten entsprechende Förderungen auch als Ermutigung begriffen werden, Replikationsstudien aufzuwerten und mit Reputationschancen für diejenigen zu versehen, die sie auf hohem methodischen Niveau durchführen.

4.2.3 WISSENSCHAFTLICHE DATENQUALITÄT IM FORSCHUNGSPROZESS FORTLAUFEND SICHERN

Qualitätssicherung als permanente Herausforderung begreifen

Datenqualität zu sichern und zu steigern ist eine permanente Herausforderung über den gesamten Forschungsprozess und muss von den Forscherinnen und Forschern entsprechend reflektiert werden. Informationswissenschaften und Infrastruktureinrichtungen können technische Prinzipien oder Standards als Leitplanken für Qualität setzen. Letztlich ist für die wissenschaftliche Qualität von Daten aber eine Pluralität und auch Dynamik von Kriterien kennzeichnend. Diese können im Detail nur die wissenschaftlichen Fachgemeinschaften selbst definieren und beschreiben beziehungsweise in die Forschungspraxis integrieren. Harmonisierung und Standardsetzungen im Forschungsdatenmanagement erfordern Festlegungen seitens der Forschung beziehungsweise fachspezifische Antworten auf Qualitätsfragen. Allerdings sind immer die interdisziplinäre Anschlussfähigkeit und potenzielle Transfermöglichkeiten im Auge zu behalten: Breite (wissenschaftliche) Nachnutzbarkeit ist kein abschließendes Kriterium, aber ein wichtiger Aspekt von Datenqualität.

Aufgabenfelder für NFDI und EOSC

Der RfII sieht gerade in diesem Aspekt der Qualitätssicherung von Datenbeständen auch für eine interdisziplinäre Verwendung eine der wichtigsten Aufgaben der künftigen NFDI- und auch der EOSC-Konsortien: Ihre Aufgabe ist es, eine Balance zu finden zwischen der Setzung von allgemeinen disziplin- und domänenübergreifenden Leitlinien und den fachspezifischen Qualitätsdiskursen. Grundsätzlich sollten Harmonisierungen und Standardsetzungen – sofern sie nicht rein

technische Fragen betreffen – immer aus den Fachgemeinschaften kommen beziehungsweise im Rahmen der NFDI an deren Resonanz gekoppelt werden.

4.2.4 VERANTWORTUNG FÜR DIE OPERATIONALISIERUNG VON QUALITÄTSKRITERIEN ÜBERNEHMEN

Die Fachgemeinschaften sind in der Verantwortung

Der RfII ermutigt die wissenschaftlichen Fachgemeinschaften, für die Definition und Beschreibung der Pluralität und der Dynamik von Qualitätskriterien, aktiv Verantwortung zu übernehmen – und dies namentlich auch im Rahmen notwendiger Aushandlungsprozesse in den und zwischen den NFDI-Konsortien. Er empfiehlt, grundsätzlich die enge Verbindung digitaler Forschungsprozesse mit digitalen Werkzeugen und Diensten (also: die „Infrastrukturdimension") in den fachwissenschaftlichen Methodendiskursen intensiver als bisher zu reflektieren. Dies gilt in besonderem Maße für Disziplinen, die mit qualitativen Forschungsmethoden beziehungsweise in hermeneutisch-interpretierenden und beobachtenden Forschungsformen arbeiten – insbesondere große Teile der Geistes- und Sozialwissenschaften. Daten, die zum Beispiel auf Feldbeobachtungen oder offenen Befragungen und Interviews basieren, stellen wiederum andere Herausforderungen an die methodische Qualität und Dokumentation ihrer Entstehungszusammenhänge als solche, die auf Quelleninterpretationen beruhen. Auch zwischen den Naturwissenschaften gibt es große Unterschiede: Datenbestände, die über bildgebende Verfahren durch Sensoren, Detektoren oder Scanner gewonnen werden, stellen andere Herausforderungen an Qualitätssicherungsprozesse als Daten, die über biochemische Reaktionen im Labor aufgezeichnet werden oder solchen, die im Rahmen der translationalen Medizin zum Beispiel im Zusammenspiel von epidemiologischen Studien und individualisierten Therapien entstehen. Vielfach muss auch die Verknüpfung von „analogen" Daten und digitalen Artefakten im Fokus von Datenqualitätskonzepten stehen – so in den Ingenieurwissenschaften, aber auch in der naturwissenschaftlichen Empirie. Der RfII sieht auch hier die NFDI-Konsortien in der Verantwortung, wechselseitiges Verständnis zu fördern und bei aller fachlichen Unterschiedlichkeit auf einige gemeinsame Leitlinien für gutes Forschungsdatenmanagement hinzuwirken. Solche zu erarbeiten, schließt die Gewissheit ein, dass es über den gesamten Datenlebenszyklus hinweg keine die Disziplinen übergreifenden „One size fits all"-Lösungen geben kann.

Muster für FDM-Pläne entwickeln

Zur Operationalisierung der qualitätsbezogenen Leitlinien für das Datenmanagement in einzelnen Disziplinen oder Fächergruppen empfiehlt der RfII, Muster für Forschungsdatenmanagementpläne (FDM-Pläne) zu erstellen und weiterzuentwickeln. Dies sollte in Zwiesprache von Fachgemeinschaften, Hochschulen und außeruniversitären Forschungseinrichtungen mit Infrastrukturaufgaben erfolgen. Entsprechende Muster, die auch nach Forschungsformen differenzieren

würden, könnten lokal an die Anforderungen und Aufgabenstellungen konkreter Projektzusammenhänge angepasst werden. Der RfII sieht FDM-Pläne keineswegs als eine zusätzliche bürokratische Zumutung für den Forschungsprozess. Im Gegenteil können sie – sofern sie den disziplinären Bedürfnissen und Forschungsformen adäquat angepasst sind – den Forschungsprozess entlasten: Sie machen den Aufwand sichtbar, den Forschende in Qualität investieren, und stellen eine wichtige Grundlage für die adäquate Bemessung von Drittmitteln sowie für die Einhaltung der Prinzipien guter wissenschaftlicher Praxis dar.

4.2.5 FORSCHUNGSINFRASTRUKTUREN ALS ERMÖGLICHUNGSSTRUKTUREN

Güte der Daten abhängig von Infrastrukturqualität

Die Qualität von Forschungsdaten wird auch vom *state of the art* der Forschungsinfrastrukturen bestimmt, mit deren Hilfe (namentlich digitale) Daten analysiert, gespeichert und geteilt werden. Ebenso stellt die Qualität der Forschungsdaten einen entscheidenden Maßstab für die Entwicklung guter Infrastrukturen dar. Dieses Zusammenspiel von Datenqualität und Infrastrukturqualität muss in zahlreichen akademischen Disziplinen viel ausdrücklicher als bisher erkannt werden. Es gehört zum Grundverständnis des jeweiligen Fachs und sollte bereits im Rahmen des Studiums vermittelt werden. In das Methodenverständnis in den Disziplinen sollten in diesem Zusammenhang grundsätzlich Kenntnisse über datenerzeugende,-prozessierende und-speichernde Infrastrukturen – von wichtigen institutionellen Einrichtungen, Prozessschritten (Datenlebenszyklus) bis hin zu Hard- und Softwarekenntnissen – einbezogen werden (siehe auch 4.6 zur Personalentwicklung). Eine gute Kenntnis der Stationen des gesamten Datenlebenszyklus und des dort jeweils stattfindenden Wechselspiels von Forschungshandeln und Infrastrukturgebundenheit des Forschungsprozesses würde dazu beitragen können, bereits während der Forschung Überlegungen zur Datendokumentation anzustellen, die für die spätere Langzeitarchivierung und Pflege einer Datensammlung unabdingbar sind.

4.2.6 HÖHERE ANERKENNUNG FÜR DIE ARBEIT MIT FORSCHUNGSDATEN

Qualitätssicherung von Daten als unverzichtbare Leistung im Wissenschaftssystem anerkennen

Der RfII hält es für unabdingbar, dass einem wachsenden Problembewusstsein für die mit der Digitalität verbundenen Herausforderungen an gute wissenschaftliche Praxis entsprechende Fähigkeiten und Kapazitäten aufseiten der Wissenschaft aktiv entwickelt und wertgeschätzt werden müssen. Er begrüßt daher den von zahlreichen Fachgemeinschaften inzwischen forciert geführten nationalen und internationalen Diskurs zur Anerkennung von qualitätssichernden Verfahren als Teil digitaler „Methoden“. Entsprechende Ansätze, wie sie künftig im Rahmen der NFDI und diverser EOSC-Initiativen zu entwickeln sind, müssen nachhaltig zur Legitimation von Qualitätssicherungsverfahren und zur Anerkennung hierauf bezogener akademischer Tätigkeiten führen. Letztere sind nicht allein

„gute Praxis“, sondern unverzichtbare Leistungen im Forschungsprozess, die Anerkennung und Reputation verdienen, weil sie eine wesentliche Grundlage für die Validität und Robustheit von Forschungsergebnissen darstellen. Die Herstellung von Datenqualität ist in allen Disziplinen ein „Positivziel“ von Forschung, dessen Erreichung als reputationswirksame Leistung im gesamten Datenlebenszyklus zu kultivieren ist.

Qualitätssicherung als Thema in der wissenschaftlichen Ausbildung forcieren

Überall dort, wo hohes Aufkommen und Toleranz gegenüber ungeprüften Daten, die sich auf vielfältige gesellschaftlich relevante Fragestellungen beziehen können, die „traditionell“ hohen Qualitätskriterien der öffentlich finanzierten Forschung herausfordert, kann und soll die Wissenschaft dies zum Anlass nehmen, auch die Rahmensetzung für ihre eigenen Erfolgsbedingungen neu zu verhandeln. Dies bedeutet auch, Qualitätssicherung von Daten als selbstverständlichen Gegenstand der wissenschaftlichen Ausbildung in allen Disziplinen ernst zu nehmen und entsprechende Studien- und Ausbildungsangebote weiter auszubauen. Hierbei empfiehlt der RfII, besser zusätzliche Angebote in die Curricula der Fachwissenschaften zu integrieren als transdisziplinäre Lehrstühle ohne Kontakt zur jeweiligen disziplinären Forschungsbasis neu einzurichten. Nur die unmittelbare Integration in Fachcurricula wird letztlich die gewünschte Erweiterung des fach- oder feldspezifischen Methodenverständnisses gewährleisten können.

4.3 QUALITÄTSSICHERUNG IM DATENLEBENSZYKLUS ALS WISSENSCHAFTLICHE AUFGABE ANNEHMEN

Dokumentationsaufgaben sind Bestandteil guter wissenschaftlicher Praxis

Im Datenlebenszyklus entstehen in allen Stadien und an allen Schnittstellen spezifische Probleme, die sich negativ auf die Datenqualität auswirken und anschließend im Zyklus weitergereicht werden können. Am Ende der Kette sind so eventuell Forschungsergebnisse durch Ungenauigkeiten, Fehler, einen „Bias“ oder auch durch fehlende Nachhaltigkeit gefährdet. Ebenso sind wichtige Bedingungen wie (je nach konkreter Konstellation) Validität, Reproduzierbarkeit, Authentizität etc. nur auf der Basis guter Datenprozessierung gegeben. Die Idee der guten wissenschaftlichen Praxis bleibt in dieser Hinsicht auf den Rahmen einer ebenso „gut“ gelebten Verantwortungskultur von Fachgemeinschaften bezogen. Ebenso bedarf es einer Konkretisierung, wie sie an und zwischen jeder Station in einem – je nach Disziplin – unterschiedlich ausgestalteten Datenlebenszyklus gelebt werden kann. Der RfII empfiehlt in diesem Zusammenhang den Fachgemeinschaften genauso wie den individuellen Forscherinnen und Forschern, Dokumentationsaufgaben in jedem Stadium des Datenlebenszyklus als wesentlichen Beitrag zu guter wissenschaftlicher Praxis ernst zu nehmen und die Folgen von datenbezogenen Entscheidungen im Forschungsprozess vor dem Hintergrund späterer Archivierung, Zugänglichkeit, Pflege und Verwertung für weitere wissenschaftliche Fragestellungen (auch jenseits des eigenen Forschungsfeldes) zu reflektieren.

4.3.1 ANFORDERUNGEN AN DATENBESCHREIBUNG UND -DEKLARATION PRÄZISIEREN

Fachliche Beurteilung der Datenqualität sicherstellen

Die analytischen Teile des Positionspapiers (Kap. 1 und 2) haben gezeigt, dass aus dem Anspruch auf Wissenschaftlichkeit grundlegende Anforderungen an die Datenbeschreibung und Deklaration resultieren: Die Generierung von Daten muss so beschrieben und ausgewiesen werden, dass mindestens innerhalb der jeweiligen fachlichen oder methodischen Domäne eine Beurteilung ihrer Qualität möglich ist. Diese Anforderungen bestehen unabhängig davon, ob die Daten im Rahmen guter wissenschaftlicher Praxis archiviert oder in irgendeiner Form als Produkt für die Nachnutzung durch Dritte verfügbar gemacht werden.

Hierbei empfiehlt der RfII, folgende Anforderungen zu präzisieren:

- Forscherinnen und Forscher müssen – gegebenenfalls mit professioneller Unterstützung – eine Entscheidung treffen, wo sie auf Quantität und wo auf Qualität der abgelegten Daten fokussieren. Auch stark „verrauschte" Datensätze sind wissenschaftlich wertvoll. Die jeweils vollzogenen Qualitätssicherungsschritte und auch verbleibende Unwägbarkeiten müssen transparent ausgewiesen sein.

- Datensätze müssen in ihrer Entstehungsgeschichte ausgewiesen werden und – dort wo der Forschungsprozess dies zulässt – nachverfolgbar sein. Leitbild ist das eines „Provenienzkontinuums" entlang des Datenlebenszyklus, inkl. Angaben zu verwendeter Software und Codes sowie wichtiger Transformationsschritte, wie zum Beispiel Anonymisierung.

- Bei personenbezogenen Daten sind Datenschutzgarantien und entsprechende Kennzeichnungen unabdingbar: Dazu gehört auch ein für Dritte nachvollziehbares Einwilligungsmanagement, zum Beispiel, ob Daten zu Validierungszwecken oder für weitere Forschungen an Dritte weitergeben werden dürfen.

- Verfügungsrechte müssen dokumentiert sein: Wer ist berechtigt, die Daten im Bedarfsfall abzugeben beziehungsweise über eine Veröffentlichung zu befinden oder gar Korrekturen zu autorisieren? Ebenso müssen die Konditionen für eine innerwissenschaftliche Nutzung beziehungsweise gegebenenfalls wirtschaftliche Verwertung hinterlegt sein.

Hinreichende Datenbeschreibung durch Metadaten ist zwingend

Gerade für digitale Forschung gilt: Daten sprechen nicht für sich, sondern können in ihrer Qualität nur dann beurteilt werden, wenn ihr Entstehungszusammenhang über Metadaten hinreichend beschrieben ist. Nur durch differenzierte Qualitätskonzepte (jenseits von problematischen Vereinfachungen) leistet Wissenschaft einen Beitrag, um populistischen „Fake"-Botschaften entgegenzutreten.

4.3.2 SCHNITTSTELLENKOMMUNIKATION VERBESSERN

Datenqualität aktiv gestalten – Schnittstellenmanagement verbessern

Der RfII setzt sich dafür ein, dass auf allen institutionellen Ebenen des Wissenschaftssystems die Kommunikation an den Schnittstellen des Datenlebenszyklus verbessert und durch vorausschauendes Forschungsmanagement aktiv unterstützt wird. Neben den individuellen Forscherinnen und Forschern sieht der RfII hier in erster Linie die Fachgemeinschaften sowie die Wissenschaftsorganisationen in der Pflicht. Die Hochschulen und die außeruniversitären Forschungsorganisationen sollten ihre durch Zielvereinbarungen und das System der Pakte sich eröffnenden Handlungsmöglichkeiten nutzen, um gemeinsam an Datenmanagementstrategien über den Datenlebenszyklus hinweg zu arbeiten. Ebenso sollten sie gezielt Ressourcen poolen. Alle Akteure sind zudem aufgefordert, rechtzeitig Transparenz zu schaffen hinsichtlich der Aufwände und Kosten für die Datenarbeit in den jeweiligen Stationen des Datenzyklus. „Datenqualität" ist ein Thema, das in hohem Maße aktiv gestaltet werden muss.

Im Folgenden verweist der RfII auf einzelne Schritte, die bei der Umsetzung eines verbesserten Schnittstellenmanagements zu beachten wären:

- Datenproduzenten sollten bei der Erhebung und Generierung von Daten spätere Schritte des Datenlebenszyklus entsprechend den Möglichkeiten ihrer Fachkultur antizipieren und dabei jeweils professionell von Ansprechpartnern in ihrem institutionellen Umfeld unterstützt werden. Im Rahmen beispielsweise der NFDI und der EOSC sollten sich die Forschungsakteure dort, wo Bedarf besteht und noch keine entsprechenden Konzepte oder Umsetzungen vorliegen, über die Entwicklung einer nachhaltigen und jeweils fach- beziehungsweise fächergruppenadäquaten *Data Governance* austauschen. Entsprechende Strategien sollten eine gemeinsame Verantwortungsübernahme ermöglichen (wer ist wie für die jeweiligen Schritte im Datenlebenszyklus ansprechbar/verantwortlich, welche Dienste sind vorhanden, etc.).

- Wo immer es möglich ist, empfiehlt der RfII, Verabredungen hinsichtlich der akzeptierten Formate, Vokabularien und Ontologien oder der Verwendung einheitlicher Vorlagen zu treffen. Die Nomenklatura- und Standardisierungskomitees internationaler Fachverbände in den Naturwissenschaften leisten hier entsprechende Koordinierungsarbeiten und geben verbindliche Empfehlungen, die internationale Akzeptanz und Anwendung finden. In Disziplinen und Fächergruppen, in denen die Forschung in stärkerem Maße idiosynkratisch, multiparadigmatisch oder einfach weniger arbeitsteilig organisiert ist, sollten die einschlägigen Fachverbände zumindest prüfen, wie eine Konsentierung übergreifender Standards der Datenbeschreibung gestaltet werden könnte, um neben Wiederauffindbarkeit auch eine breitere wissenschaftliche Anschlussfähigkeit von Daten sicherstellen zu können. Die FAIR-Prinzipien bieten in diesem Zusammenhang eine gute Orientierungshilfe.

- Im Wissenschaftssystem sollten Kapazitäten für Beratungsangebote zur rechtssicheren Ausgestaltung des Umgangs mit Daten, das heißt zu spezifischen Fragen des Urheberrechtes, von Copyright- und anderen Verwertungsbestimmungen, Datenschutz, Dienstrecht, „geistigem Eigentum“ sowie zur guten wissenschaftlichen Praxis geschaffen werden.[96] Diese Beratungskapazitäten sollten insbesondere auf den in der Regel internationalen Charakter von datenrechtlichen Problemstellungen eingestellt sein. Sie müssen nicht notwendigerweise lokal aufgebaut werden, sondern können auch überregional oder in Verbünden installiert werden. Die NFDI-Konsortien wären auf nationaler Ebene wichtige Akteure, um entsprechende Anstöße zu geben.

4.3.3 TECHNISCHE PRÜFVERFAHREN WEITERENTWICKELN UND SYSTEMATISCH ZUM EINSATZ BRINGEN

IT-gestützte Verfahren zur Qualitätsprüfung konsequenter nutzen

Der RfII ist der Auffassung, dass die Potenziale von IT-gestützten Verfahren zur Integritäts-/Konsistenz- und Qualitätsprüfung von Daten noch zu wenig genutzt werden. Sie könnten die geforderten Dokumentationsschritte im Forschungsprozess erheblich erleichtern und damit für einen spürbaren Fortschritt im Datenmanagement sorgen.

Im Rahmen der Forschungsförderung auf nationaler und europäischer Ebene und in Abstimmung mit entsprechenden Initiativen der künftigen NFDI sollten gezielt Projekte zur Weiterentwicklung und Erprobung von Diensten und Verfahren gefördert werden. Perspektivisch sollten erprobte Verfahren und Methoden der Datenprüfung im Sinne von Best-Practice-Beispielen über einzelne Einrichtungen und Fachgemeinschaften hinaus der Wissenschaft zur Verfügung gestellt werden.

Verfahren, die künftig im Rahmen der Forschungsförderung Berücksichtigung finden sollten, wären zum Beispiel (ohne Anspruch auf Vollständigkeit):

- automatisierte Qualitätsprüfungen für digitale Daten („validators“) und Plausibilitätsprüfungen,
- Verfahren zur Sicherung der Dokumentation von Datenprovenienz („Provenance Tracking“),
- Verfahren zur Dokumentation von Datentransformationen,
- Verfahren zur Fehlerbehebung,
- Verfahren einer auch „historisch“ weit zurückreichenden Auswertung von Datenträgern hinsichtlich der Art ihrer Nutzungsverfahren, um abgeleitete Datensätze vergleichbar zu machen,
- Verfahren zur Aufdeckung von Datenfälschungen und Datensabotage.

[96] Vgl. RfII (2016) – Leistung aus Vielfalt, S. 61 ff., Empfehlungen 4.11 und 4.12.

4.3.4 KONTROLLE UND TRANSPARENZ VON SOFTWAREPRODUKTEN

Blackbox-Effekte beeinträchtigen Datenqualität

Ein großes Problem in der Nutzung proprietärer Hard- und Software im wissenschaftlichen Forschungsprozess ist die Unkenntnis der Forschenden über das „Innenleben" der Maschinen, mit denen sie Experimente und Analysen durchführen sowie dessen Einfluss auf die Datenerzeugung, -verarbeitung und -analyse. Im schlimmsten Fall wäre ein Forschungsergebnis ein schlichtes Artefakt des maschinellen „Innenlebens" eines Messgeräts oder des Verarbeitungsalgorithmus der eingesetzten Software. In Unwissenheit über den „withinput" der benutzten Instrumente erzielte „Forschungsergebnisse" sind nicht replizierbar und konterkarieren das Ideal einer guten wissenschaftlichen Praxis (siehe 2.1.8). Das als *Blackboxing* bekannte Phänomen ist nicht leicht zu beheben, da der Wissenschaft oft die Ressourcen fehlen, um eigene Geräteentwicklungen zu forcieren, die sie dann auch kontrollieren könnte. Helfen können hier – zumindest auf dem Feld des Software-Einsatzes – wissenschaftseigene Open-Source-Entwicklungen, bei denen der Quelltext transparent gemacht wird. Nichtsdestotrotz wird *Blackboxing* die Wissenschaft dauerhaft als kritisches Phänomen begleiten.

Fachgemeinschaften müssen gemeinsam Anforderungen an Hersteller artikulieren

Die kommerziellen Hersteller von Forschungsgeräten und Laborwerkzeugen verweisen auf ihre Eigentumsrechte und versuchen, sich vor Technologiepiraterie zu schützen. Dennoch gibt es Handlungsoptionen zumindest dort, wo wissenschaftliche und klinische Einrichtungen die einzigen Abnehmer von Industrieprodukten sind. Neben dem gemeinschaftlichen Testen der Funktionen und Folgen verschiedener Produkte (Benchmarking) sollten sich Fachgemeinschaften, Hochschulen und Wissenschaftsorganisationen zusammenschließen und – ähnlich wie in der Auseinandersetzung mit den Publikationsoligopolen – gemeinsame Bedarfe sichten und artikulieren.

Der RfII empfiehlt in diesem Zusammenhang:

- eine breite Offenlegung von Intransparenz externer Geräte und Produkte sowie der Ergebnisse von disziplinären Benchmarking-Prozessen, mit welchen sich die betroffenen Forscherinnen und Forscher vielfach bereits selbst behelfen;
- eine möglichst gemeinschaftliche Auswahl des unter Transparenzgesichtspunkten für eine Domäne beziehungsweise ein Forschungsfeld „besseren" Produkts;
- die Entwicklung wissenschaftseigener Lösungen – quasi Infrastrukturbausteine als „Commons" – wo die Größe des wissenschaftsinternen Absatzmarktes dies auch unter ökonomischen Gesichtspunkten rechtfertigt;
- die Schaffung einer Clearing-Stelle im deutschen Wissenschaftssystem, um gegebenenfalls Kauf-, Wartungs- und Nutzungskonditionen mit Herstellern auszuhandeln und verbindlich zu vereinbaren (vgl. zum Bedarf an Gesprächsverbünden Empfehlung 4.5.4).

4.4 DATENPRODUKTE ENTWERFEN UND AUSDIFFERENZIEREN

Datenprodukte steigern die wissenschaftliche Wertschöpfung

Der RfII sieht in gut dokumentierten und kuratierten „Datenprodukten“ ein großes Potenzial für die wissenschaftliche Wertschöpfung. Bereits jetzt haben sich verschiedene Formate von Datenprodukten herausgebildet (siehe auch 2.1.4). Solche Datenprodukte haben einerseits den Status von eigenständigen Erkenntnisprodukten, ohne die wissenschaftliche Durchbrüche und der Transfer von Forschungserkenntnissen in das Innovationssystem nicht denkbar wären. Andererseits erfüllen Datenprodukte eine wichtige Belegfunktion im Wissenschaftssystem. Datenbanken und Datenzentren beispielsweise können in diesem Sinne auch in der Tradition der wissenschaftlichen Sammlungen betrachtet werden: Sie erfüllen eine wichtige Funktion der Validierung und Rückversicherung von Forschungsergebnissen und sorgen damit für Zeitstabilität sowie überhaupt für Nachhaltigkeit des wissenschaftlich geprüften Wissens.

4.4.1 DATENPRODUKTE AUSDIFFERENZIEREN UND VERBREITEN

Erstellung von Datenprodukten mit wissenschaftlicher Anerkennung versehen

Der RfII empfiehlt den Forschenden und ihren Fachgemeinschaften wie auch den Wissenschaftsorganisationen gemeinsam mit den Informationsinfrastrukturen, die Anfertigung von Datenprodukten nachhaltig zu unterstützen. Diese sollten als eigenständige wissenschaftliche Leistungen Wertschätzung erfahren. Um den o. a. Anforderungen im Wissenschaftssystem gerecht zu werden, müssen Zweck und Zielgruppe für diese Produkte klar bestimmbar sein, idealerweise sind sie auch an explizierten, (fach-)wissenschaftlich akzeptierten Standards ausgerichtet. Welche Form von Datenprodukt für welchen wissenschaftlichen und disziplinären Zweck auch von der NFDI oder der EOSC befördert und bereitgestellt werden kann, sollte mittelfristig im Rahmen der NFDI-Konsortien entschieden werden. Bereits kurzfristig sieht der RfII die Forschungsfördernden in der Pflicht, Anreize für die Entwicklung von Datenprodukten zu setzen (siehe 4.7.1). Künftige Datenprodukte können sich an folgenden bereits entwickelten Formaten orientieren beziehungsweise diese weiter ausbauen:

a. Abgabe eines Datensatzes in eine existierende wissenschaftliche Sammlung
Die Abgabe eines Datensatzes an ein – im besten Falle zertifiziertes – Archiv beziehungsweise eine Datensammlung mit fortlaufender Kuratierung stellt die niedrigschwelligste Variante eines Datenprodukts dar. In der Regel wird der abgebende Forschende den Datensatz so aufbereiten müssen, dass er mit den Leitlinien beziehungsweise Regeln zur Einlieferung in die Sammlung kompatibel ist. Für die Fachverbände derjenigen Disziplinen und Fächergruppen, in denen potenziell nachnutzbare Daten im Forschungsprozess entstehen, empfiehlt der RfII, diese Praxis als Mindestanforderung an eine verantwortliche Sicherung der Daten zur Maxime zu machen.

b. Ko-Publikation von Ergebnis und zugehörigem Datensatz („Enhanced Publication“)

Diese Kombination von Ergebnispublikation und einem geeignet verlinkten Datensatz ist im Sinne einer Qualitätssicherung von Forschung grundsätzlich sinnvoll und förderwürdig. Allerdings erfüllt die bislang vielfach noch gelebte Praxis der Supplement-Publikation im PDF-Format die Anforderungen an Zugänglichkeit und Interoperabilität nicht. Datensätze für „Enhanced Publications“ sollten – mit der gleichen Sorgfalt und kontextbezogenen Qualität – maschinenlesbar aufbereitet sein, damit sie für etwaige Replikationsstudien oder anschließende Forschungsfragestellungen leichter nutzbar sind (s. hierzu auch Kap. 2.1.9). Zu konstatieren ist allerdings auch, dass sich die Probleme des Publikations- und Begutachtungswesens im Bereich der „enhanced publications“ potenzieren (vgl. 3.1.2 und 3.1.3). Insbesondere die Begutachtungspraxis für Daten muss aus Sicht des RfII dringend verbessert werden. IT-gestützte Verfahren zur Überprüfung der technischen Datenqualität sind ein möglicher Teil der Lösung. Interne Qualitätssicherungsverfahren, die Datensätze vor Publikation durchlaufen, können eine sinnvolle Ergänzung der externen kollegialen Begutachtung sein, soweit sie transparent geregelt und für Dritte nachvollziehbar sind.

c. Digitale Editionen

Die Erstellung „digitaler Dateneditionen“ geht mit Blick auf Aufwand und Güte der Dokumentation über Enhanced Publications noch deutlich hinaus. Ziel einer „Edition“ ist es, Daten jenseits ihrer Belegfunktion oder der Möglichkeit, sie wiederholt für ähnliche Zwecke zu nutzen, so einzurichten, dass sie auf längere Zeit und für möglichst viele Forschungsfragen nutzbar werden. Dazu zählt unter anderem eine zeitstabile Einrichtung (ggf. Langzeitarchivierbarkeit), eine domänenübergreifende Annotation mit Metadaten, die interaktive Verknüpfung mit externem Archivmaterial oder die Analyse und Visualisierung von Textphänomenen mithilfe digitaler Tools und Dienste. Ein solches Datenprodukt lässt sich nicht nur für Sprach- und Bilddaten, sondern auch für Messergebnisse, quantitative Umfragedaten etc. erstellen, um den Kontext der Datensätze aufzuzeigen und eine Einordnung zu ermöglichen. Der RfII empfiehlt, diese Produktform als anerkannte wissenschaftliche Leistung zur Datendokumentation zu fördern. In den wissenschaftlichen Fachgemeinschaften sollten hierzu Standards für geeignete Formate entwickelt werden, ebenso sind Fragen der technischen Erhaltung und Langzeitverfügbarkeit zu klären.

d. Anbieterseitige Data Reports

Data Reports sind Datenprodukte, die aus Anbieterperspektive (hier: Wissenschaftlerinnen und Wissenschaftler) für Forschung genutztes oder nutzbares Datenmaterial darlegen und beschreiben. Sie sind häufig im Rahmen von Großforschungseinrichtungen, infrastrukturtragenden außeruniversitären Forschungsinstituten oder größeren Forschungsdatenzentren gängige Instrumente, um die interessierte Fachöffentlichkeit kontinuierlich mit einer

laufenden und qualitätsgesicherten Datenberichterstattung aus langfristigen Forschungszusammenhängen zu versorgen – von Panel-Untersuchungen über Daten aus fortgesetzten Beschleunigerexperimenten bis hin zu astrophysikalischen Beobachtungsdaten der großen Radioteleskopanlagen. Der RfII empfiehlt, solche Reports im Sinne eines mitlaufenden Daten-Monitorings auch im Rahmen langfristiger Forschungsprojekte (zum Beispiel SFBs oder Exzellenzclustern) vorzusehen beziehungsweise je nach Fach- und Feldspezifik deren Einsatz zu prüfen.

e. Aufbau kuratierter Datensammlungen

In Aufbau und Pflege einer kuratierten Datensammlung sieht der RfII eines der umfänglichsten Datenprodukte. Eine kuratierte Sammlung zeichnet sich durch dynamische Pflege und Aufbereitung von Forschungsdaten aus, die eng an aktuellen Forschungsfragestellungen orientiert und oftmals kollektiv organisiert ist. Der Aufbau solcher Datenbestände setzt die Kenntnis beziehungsweise Einhaltung von Standards schon bei der Erhebung der Daten voraus. Aus der Kuratierung heraus können sich auch neue Produktformate und Standards entwickeln, die wiederum eigenständige wissenschaftliche Leistungen darstellen. Gleichzeitig müssen sich die Kuratorinnen und Kuratoren offen zeigen für neue standardsetzende Entwicklungen, die aus der Wissenschaft selbst kommen. Der RfII begrüßt, dass im Rahmen der systematischen Sammlung von Forschungsdaten auch deren Aufbereitung für weitere Forschungszwecke beziehungsweise die Öffentlichkeit oft einen großen Stellenwert hat, zum Beispiel in Form von *Scientific* beziehungsweise *Public Use Files* oder die Präsentation von Daten in Kombination mit Software beziehungsweise einfach zu nutzenden „Apps“.

4.4.2 HERAUSGABE UND OFFENLEGUNG VON DATEN DIFFERENZIERT HANDHABEN

Unterschiedliche Fächerkulturen und Forschungsformen berücksichtigen

Die Offenlegung von Forschungsdaten in Form von Datenprodukten ist eine anspruchsvolle und voraussetzungsreiche Aufgabe. Namentlich in den Geistes- und Sozialwissenschaften (aber auch in Teilen der Lebenswissenschaften und der klinischen Forschung) muss fallweise entschieden werden, ob und in welchem Umfang mit der Verpflichtung zur Herausgabe und Offenlegung erhobener Daten die Schwelle für einen gelingenden Feldzugang beziehungsweise für die Interviewbereitschaft von Personen und Gruppen höher gelegt wird. Dies muss anhand geeigneter empirischer Studien überprüft werden. Da im Bereich hypothesenfreier Forschung auf großen digitalen Datenmengen sowohl Personenbezug als auch andere (etwa gesellschaftliche Gruppen exponierende) kritische Formen der Datennutzung sehr einfach möglich sind, ist auch hinsichtlich der Nutzung von Sozial- und Sprachdaten durch die informatische Grundlagenforschung der begonnene Fachdiskurs über ein Einwilligungsmanagement sowie gegebenenfalls über „ethische“ Grenzen von Datenanalytik wichtig und erforderlich. Auch hier ist zu berücksichtigen, dass eine für Forschungszwecke

praktizierte Nachverfolgung von Datenspuren, die von Personen im Internet hinterlassen werden, zur Verletzung von Persönlichkeitsrechten führen und den Feldzugang mittelfristig weiter limitieren kann. Die Empfehlung zur Erstellung von Datenprodukten sollte daher nicht per se mit einer Veröffentlichung gleichgesetzt werden. Im Einzelfall sind zweckdienliche und angemessene Zugangsregelungen zu finden.

4.4.3 REZENSIONSKULTUR FÜR FORSCHUNGSDATEN FÖRDERN

Datenrezensionen aus der Nische holen – in führenden Fachzeitschriften lancieren

In dem Maße, wie sich Datenprodukte als gleichwertige Formate neben der Ergebnispublikation etablieren, hält der RfII auch die Förderung einer geeigneten Rezensionskultur für sinnvoll, um die Bekanntheit dieser Ressourcen in den Fachgemeinschaften zu steigern sowie Interaktionen mit den forschenden Nutzern zu stimulieren. Hierbei können in geeigneten Forschungsfeldern eigene Rezensionsorgane für Forschungsdaten sinnvoll sein.[97] Um die Integration von Datenqualität in das allgemeine Methodenverständnis nachdrücklich anzuregen, wäre allerdings in den meisten Fällen darauf hinzuwirken, der Forschungsdatenrezension in den führenden Zeitschriften der Fachgemeinschaften mehr Raum und gegebenenfalls eine eigene Rubrik zu geben.

4.4.4 FAIRE PARTNERSCHAFTEN MIT DIENSTLEISTERN

Markt für Datenprodukte wissenschaftsgerecht ausgestalten

Die Aufbereitung von Daten erfordert Zeit und – je nach Format des Datenprodukts – informatische Expertise unterschiedlicher Tiefe. Dies ist durch Forschende, Projektbeteiligte oder auch wissenschaftliche Einrichtungen nicht in jedem Fall vollumfänglich zu leisten. Absehbar wird sich daher ein Markt für die Erstellung von Datenprodukten entwickeln, ähnlich wie im Bereich der Forschungssoftware. Entsprechende Bewegungen sind bei etablierten Verlagen ebenso zu beobachten wie im Bereich der wissenschaftsnahen Ausgründungen. Dabei sind alle beteiligten Parteien zu einem hohen Maß an Fairness im Interesse des offenen Zugangs zu Forschungsdaten und-ergebnissen aufgefordert. In entsprechenden Vertragsgestaltungen müssen die Wissenschaftseinrichtungen und ihre Träger dafür Sorge tragen, dass kommerzielle Partner und Dienstleister die ihnen anvertrauten Daten weiter zugänglich halten und – im Verhältnis zu Dritten – die Datensätze der Forscherinnen und Forscher vor möglicherweise unberechtigtem Zugriff schützen. Der RfII weist in diesem Zusammenhang auf die Bedeutung der Souveränität der Forschung über „ihre“ Daten als einer wesentlichen Grundlage für die Funktionsfähigkeit des Wissenschaftssystems und die damit verknüpften gesellschaftlichen Erwartungshaltungen hin.

[97] Vgl. auch die Ausführungen in 3.1.2 mit einigen Beispielen.

4.5 FORSCHUNGS- UND INFORMATIONSINFRASTRUKTUREN ALS GARANTEN FÜR QUALITÄTSSICHERUNG

Dialog zwischen Infrastrukturexperten und Forschungsakteuren vertiefen

Der RfII empfiehlt den Datenarchiven und -repositorien sich als Kompetenzzentren für den Umgang mit wissenschaftlichen Daten und damit als institutionelle Formen der Qualitätssicherung und Qualitätsförderung im Wissenschaftssystem zu begreifen und sich gegebenenfalls in diese Richtung weiterzuentwickeln. Sie sollten dort, wo dies noch nicht der Fall ist, durch Wissenschaftlerinnen und Wissenschaftler in den Forschungsprozess systematisch unterstützend eingebunden werden. Keinesfalls dürfen sie als bloße „Ablagen" oder „Datenspeicher" behandelt werden, die lediglich befüllt werden, um beispielsweise Anforderungen institutsinterner oder von Forschungsförderern aufgestellter Vorgaben zu genügen. Im Gegenteil: Für eine umfassende Datenkultur im Wissenschaftssystem ist ein permanenter Dialog zwischen Informationsinfrastrukturvertretern und Forschungsakteuren zwingend erforderlich.

4.5.1 VERLÄSSLICHE INSTITUTIONELLE ANBINDUNG ALS NOTWENDIGE RAHMENBEDINGUNG

Dienste mit dem Forschungsprozess verschmelzen

Der RfII setzt sich dafür ein, dass datenbezogene Dienste unter Bedingungen von Digitalität zunehmend und im wechselseitigen Austausch von Infrastrukturen und Fachgemeinschaften mit dem Forschungsprozess selbst verschmelzen. Dies ist in Einrichtungen der Forschung mit Großgeräten wie zum Beispiel bei CERN oder DESY schon seit Langem der Fall und eine der wesentlichen Erfolgsbedingungen globaler Forschungserfolge zur Entschlüsselung der elementaren Bausteine der Materie. In den Sozialwissenschaften ist es durch selbstorganisiert initiierte und zertifizierte Forschungsdatenzentren bei Trägern öffentlicher Datenressourcen (wie zum Beispiel den statistischen Ämtern) sowie bei Leibniz-Instituten mit Infrastrukturaufgaben ebenfalls gelungen, nachhaltige Informationsinfrastrukturen aufzubauen, die den Fachgemeinschaften relativ gut zugänglich sind. Auch über den Datenlebenszyklus hinweg sind hier Mitspracherechte für Fachgesellschaften gegeben und die Erhebung und Weiterverarbeitung der Forschungsdaten ist vergleichsweise transparent dokumentiert. In vielen anderen Feldern sind Forschungsdatenmanagementprojekte in ihrem Status eher prekär. Dies gilt insbesondere für kleinere Projektzusammenhänge an Hochschulen und Universitäten sowie Universitätsbibliotheken, die nach dem Ende der üblichen befristeten Projektförderzeiten regelmäßig vor der Existenzfrage stehen. Dies ist insofern kein haltbarer Zustand, als gerade kleinere, lokale Informationsinfrastrukturprojekte in der Regel ganz unmittelbar in Forschungsprozesse eingebunden oder aus diesen heraus entstanden sind. Der RfII hat das

98 Dies ist Gegenstand der ersten Empfehlung in RfII (2016) – Leistung aus Vielfalt, S. 37 ff.

Problem prekärer Projekte 2016 deutlich angesprochen.[98] Er sieht nach wie vor einen großen Handlungsbedarf, den auch die NFDI nicht lösen kann, denn sie ist nicht zur Verstetigung existierender Projekte eingerichtet worden.

Gleichwohl können die künftigen Konsortien der NFDI Verantwortung für die Identifikation dessen übernehmen, was getan werden sollte. Der RfII empfiehlt, den Aufbau der NFDI zu nutzen, um die Verstetigung von Diensten auf dem Vernetzungs- und Kommunikationsweg langfristig zu fördern. Durch entsprechende Anbindungen an größere Organisationseinheiten der NFDI könnten kleinere lokale Projekte unter anderem an Instrumenten der Qualitätssicherung und des Best-Practice-Learning partizipieren. Entsprechende Zertifizierungsverfahren können dabei helfen, gemeinsame Standards in der Fläche zu etablieren.

4.5.2 INSTITUTIONELLE QUALITÄTSSICHERUNG

Leistungs- und Portfolioevaluationen organisieren

Der RfII empfiehlt jenen Forschungs- und Informationsinfrastruktur tragenden Einrichtungen, die sich noch nicht in Systemen einer institutionalisierten Leistungs- und Portfolioevaluation befinden, sich im Rahmen ihrer kontinuierlichen Verbesserungsprozesse in vertretbaren Abständen ebenfalls solchen Verfahren und Systemen zu unterziehen beziehungsweise sich im Rahmen des jeweils vertretbaren administrativen Aufwands an diesen zu orientieren. Zu nennen sind hier beispielsweise entsprechende Evaluationsverfahren in der Leibniz-Gemeinschaft oder die in Kapitel 1 beschriebenen Zertifizierungsverfahren. Hierbei sollten die institutionellen und personellen Strukturen so ausgerichtet werden, dass über die Zeit verlässlich auf Anforderungen technologischen und methodischen Wandels reagiert werden kann. „Evaluierung" meint hier: Hilfestellung, um Verbesserungen zu erzielen. Auch sollte der Erwerb von Qualitätssiegeln systematisch gefördert und gefordert werden. Sie machen Qualitätsanstrengungen im Sinne einer Leistung aller Organisationsmitglieder – Forschende wie Infrastrukturfachkräfte – sichtbar.

4.5.3 STANDARDS UND KRITERIEN

Trägerorganisationen in Verantwortung für die Durch- und Umsetzung von Standards

Träger von Forschungs- und Informationsinfrastrukturen betrachtet der RfII als zentrale Akteure zur Implementierung und (Mit-)Durchsetzung von Qualitätskriterien und Standards. Entsprechend sollten sie für die von ihnen aufgenommenen Daten grundsätzliche Festlegungen treffen und damit notwendige Vereinheitlichungen befördern, die technisch oder wissenschaftlich notwendig sind. Sie prüfen in diesem Zusammenhang die Vollständigkeit der Dokumentation und befördern eine Kultur der Explikation, dokumentieren selbst transparent und sorgfältig die vorgenommenen Veränderungen an Daten(-sätzen) und sorgen für technologische Anschlussfähigkeit über die Domänen- und Institutionengrenzen

hinaus. Sie haben auch gute Voraussetzungen, sich zu Kompetenzzentren für die Erstellung qualitätsgeprüfter Datenprodukte und-publikationen zu entwickeln. Die zahlreichen Forschungspublikationen, für die die Autorinnen und Autoren auf qualitätsgeprüfte Datensätze solcher Kompetenzzentren zurückgegriffen haben, belegen klar ihre Funktion als ermöglichende Einrichtungen.

4.5.4 TECHNISCHE INFRASTRUKTUR

Sicherung der materiellen Grundlagen der Forschungsinfrastruktur notwendig

Die Qualität von Forschungsdaten hängt mit der Qualität der technischen Infrastrukturen zusammen, auf denen Daten prozessiert und gespeichert werden (vgl. Kap. 2.1.5). Neben konkreten Hardware- und Softwarefehlern, die schwer zu entdecken sind, können Versionswechsel oder schlicht die Einstellung bestimmter Baureihen beziehungsweise Produktlinien sowie die Alterung oder umgebungsbedingte Schäden an Komponenten zu einer massiven Beeinträchtigung der Datenqualität führen. Auch die Replizierbarkeit wird hierdurch gegebenenfalls infrage gestellt. Um hier Abhilfe zu schaffen, müssen alle Forschungseinrichtungen und auch kleinere Forschungseinheiten an Hochschulen in die Lage versetzt werden, die Bearbeitung und Sicherung von Daten auf State-of-the-Art-Komponenten sowie mithilfe professionellen Personals durchführen zu können. Der RfII vertritt die Auffassung, dass öffentliche Forschungsförderung Hochschulen und Forschungsinstitute grundsätzlich in die Lage versetzen muss, dem Veralten von Infrastrukturkomponenten und Speichermedien begegnen zu können. Hierzu gehört auch, hinreichende finanzielle Mittel für das auf Dauer notwendige Personal und die gebotene bauliche Qualität von Forschungsbauten, in denen die Informationsinfrastrukturen untergebracht sind, bereitzustellen. Hochschulen und Forschungsinstitute sind ihrerseits dazu aufgefordert, der Pflege der technischen Dateninfrastruktur sowie dem hierfür nötigen Personalbedarf die entsprechende Priorität einzuräumen. Kleinere Forschungszusammenhänge können für die Datenspeicherung strategisch mit entsprechend ausgestatteten und professionalisierten Einrichtungen auf lokaler und regionaler Ebene, wie zum Beispiel Rechenzentren, großen Universitätsbibliotheken oder außeruniversitären Forschungseinrichtungen, zusammenarbeiten. So können Skaleneffekte erzielt werden.[99]

Gesprächsverbünde zwischen Anbietern und Nutzern initiieren

Darüber hinaus erscheint es geboten, stärker als bislang Kooperationen und vertragliche Vereinbarungen mit den kommerziellen Herstellern von Forschungsumgebungen zu suchen, um den Umgang mit eventuellen Versionswechseln

[99] Für weitere Empfehlungen, die die Langzeitarchivierung von Forschungsdaten sowie die technische Infrastruktur betreffen und Implikationen für die Aufgabenfelder der künftigen NFDI entfalten, vgl. RfII (2016) – Leistung aus Vielfalt, S. 45–48.

oder dem Auslaufen von Produktlinien einschließlich des Supports bereits im Vorfeld und im Interesse der Wissenschaft zu regeln. Der RfII empfiehlt in diesem Zusammenhang die Etablierung von Gesprächsverbünden oder Arbeitsgemeinschaften, in denen zwischen Forschung, Infrastrukturträgern und kommerziellen Anbietern auch weitere infrastrukturinduzierte Qualitätsaspekte für Daten – wie zum Beispiel die *Blackboxing*-Problematik – laufend verhandelt werden können (siehe auch zur Etablierung einer Clearing-Stelle Empfehlung 4.3.4). Auch hier ist mit Blick auf entsprechende Vertragsgestaltung auf Fairness zu achten, die der Wissenschaft die Souveränität über „ihre" Daten sichert (siehe Empfehlung 4.4.4).

4.6 DIGITALE KOMPETENZEN ALS BEDINGUNG FÜR GUTES DATENMANAGEMENT

Datenqualität in Studium, Aus- und Weiterbildung verankern

Der RfII sieht die Träger von Forschungs- und Informationsinfrastrukturen in der Pflicht, eine aktive Rolle in der Vermittlung und Weiterentwicklung von Daten- und Methodenkompetenz zu übernehmen und auch das eigene Personal stetig weiterzubilden. Dies beinhaltet, die wichtige Rolle von Datenqualität für den gesamten Forschungsprozess – von der Erhebung der Daten bis zu deren Verwertung in wissenschaftlichen Publikationen und Transferprozessen – als einen wesentlichen Inhalt in Ausbildungs- und Studienprogrammen sowie hierauf aufbauenden Fort- und Weiterbildungen zu verankern. So verstandene Studien-, Aus- und Weiterbildungsprogramme leisten nach Auffassung des RfII einen wesentlichen Beitrag zu einer gelebten guten wissenschaftlichen Praxis und helfen, gutes wissenschaftliches Arbeiten zu befördern sowie Fehlverhalten frühzeitig zu erkennen und zu vermeiden. Im Bereich der Personalentwicklung ist eine Investition in Datenkompetenz immer auch ein Beitrag zur Qualitätssteigerung der Wissenschaft und des Vertrauens in wissenschaftliche Erkenntnisse insgesamt.

In seinen Empfehlungen DIGITALE KOMPETENZEN – DRINGEND GESUCHT! hat der RfII zahlreiche Hinweise für eine Personalentwicklung in der Wissenschaft gegeben, die auch auf die Steigerung der Qualität der Forschungsdaten im gesamten Datenlebenszyklus abzielen.[100]

[100] Siehe RfII (2019) – Digitale Kompetenzen, Kapitel 4.

4.6.1 VERSÄULUNG AUFBRECHEN – AUFGABENBEZOGEN AUSBILDEN

Ausbildungsstätten miteinander verzahnen – Angebote aufgabenbezogen entwickeln

Ein Hemmnis sieht der RfII in Tendenzen zur institutionellen Versäulung der Datenproduktions- und -distributionskette – genauer: in der bislang wenig durchlässigen Trennung der Verantwortungssphären in einen technisch-administrativen, einen wissenschaftsnahen und einen im engeren Sinne wissenschaftlichen Bereich, mit jeweils unterschiedlichen Arbeitskulturen und tariflichen Regelungen, die die Arbeitsautonomie sowie die tariflichen Regulierungen zur Entlohnung und Dauerhaftigkeit der Beschäftigungsverhältnisse betreffen. Der RfII ist davon überzeugt, dass nur eine bessere Verzahnung der an der Datenproduktion und -bereitstellung beteiligten Einheiten in Hochschulen, außeruniversitären Forschungseinrichtungen, wissenschaftlichen Bibliotheken und Rechenzentren – und damit einhergehend: einer gesteigerten Durchlässigkeit der Personalkategorien – sowie massive Anstrengungen in der Aus- und Weiterbildung des Personals über die Grenzen der Personalkategorien hinweg zu einer Aufrechterhaltung und Verbesserung der Datenqualität in der Forschung beitragen können. In diesem Sinne hat der RfII empfohlen, Aus- und Weiterbildungsinhalte künftig aufgabenbezogen zu entwickeln und im Rahmen von Qualifizierungsallianzen umzusetzen. Bereits in Ausbildung und Studium sollten Perspektivenübernahmen durch Praktika und Hospitationen sowie später durch befristete Stellenrotationen über Organisationsgrenzen hinweg ermöglicht werden. Insbesondere aufseiten der Forschenden sieht der RfII den Bedarf, die Bedeutung von Infrastrukturen und infrastrukturellen Arbeiten für eine gute wissenschaftliche Praxis und hohe Qualität von Forschungsdaten zu betonen und ein Bewusstsein für die Wichtigkeit dieser Aufgaben zu schaffen.

4.6.2 INFORMATIONSWISSENSCHAFTLICHE KOMPETENZ AUSBAUEN

Vermittlung IT-basierter Kompetenzen flächendeckend intensivieren

Eine große Herausforderung für die Aufrechterhaltung einer hohen Datenqualität stellt der souveräne Umgang mit den Hardware-Umgebungen und Software-Komponenten dar, mittels derer Forschungsdaten erhoben, prozessiert, analysiert und für die Nachnutzung gespeichert werden (Stichwort unter anderem „Labor 4.0“). Zu den Anforderungen an fachwissenschaftliches beziehungsweise disziplinäres Wissen treten hohe Anforderungen an technische und technologische IT-basierte Kompetenzen, die in fachbezogenen Studien- und Ausbildungsgängen in der Regel nicht vermittelt werden. Hier ist neben einer notwendigen Qualifizierung des Personals mit informationswissenschaftlichem Grundwissen die Etablierung arbeitsteiliger Prozesse gefragt, in denen Personal mit entsprechendem IT-Know-how mit dem Fachpersonal, das digitale Verfahren anwendet oder Geräte bedient, zielgerichtet und projektbezogen interagieren kann. Auch für diesen Zweck müssen nach Auffassung des RfII Silobildungen zwischen Infrastruktureinrichtungen und den Forschungseinheiten aufgebrochen werden.

4.6.3 DATENMANAGEMENT IN DER FORSCHUNG SICHERSTELLEN

Datensicherung in der Forschung auch bei hoher personeller Fluktuation gewährleisten

Zu den wichtigen Aspekten einer Personalentwicklung unter dem Gesichtspunkt der Datenqualität gehört auch die Internationalisierung der Forschungsprozesse und damit einhergehend eine hohe Fluktuation des international rekrutierten Personals vor allem in den Hochschulen und außeruniversitären Forschungseinrichtungen. Der RfII empfiehlt an dieser Stelle, die Sicherung der Datengrundlagen, mit denen internationales wissenschaftliches Personal – aber selbstverständlich auch Junior und Senior Researcher aus dem nationalen Umfeld – während seines zumeist befristeten Aufenthalts an der jeweiligen Einrichtung gearbeitet hat, zur Leitungsaufgabe zu machen. Unkontrollierte Mitnahme von Primärdaten und Zwischenprodukten oder das Hinterlassen von „Datenfriedhöfen", die für weitere Nutzung und Verwertung nach dem Weggang der Forschenden unbrauchbar sind, sollten vermieden werden. Um dies zu gewährleisten, sollten alle Forschungsinstitutionen über Regeln für ein gutes wissenschaftliches Datenmanagement verfügen und zu Beginn auch einer kurzzeitigen Beschäftigung hierüber in einem Mitarbeitergespräch Datenverantwortung konkret zuweisen. Bei Beendigung des Beschäftigungsverhältnisses sollte überprüft werden, ob alle Beteiligten ihrer Datenverantwortung gerecht geworden sind.

4.6.4 DATENQUALITÄT UND KOMMUNIKATIONSKOMPETENZ VERBINDEN

Außenkommunikation verstärken

Inner- wie außerwissenschaftliches Vertrauen in die Qualität von Forschungsdaten entsteht nicht von selbst. In der heutigen Zeit einer digitalen „Datenflut" und teilweise durch Medienmanipulation geschaffener Pseudo-Evidenz kann die Wissenschaft nicht darauf hoffen, dass ihr Nimbus aus sich heraus gesellschaftliches Vertrauen erzeugt. Die mediale Skandalisierung von Fällen wissenschaftlichen Fehlverhaltens trägt dazu bei, dass bereits Einzelfälle die Wissenschaft als Ganze kompromittieren. Vertrauen in die Qualität der Daten, mit denen Forschung arbeitet, die sie erzeugt und auf die sie ihre Erkenntnisse stützt, muss deshalb aktiv gewonnen werden. Der RfII empfiehlt hier – über die vielen, bereits angesprochenen Punkte hinaus – für eine Professionalisierung der Fähigkeit zur Außenkommunikation über Aspekte der Datenqualität in den wissenschaftlichen genauso wie in den wissenschaftsunterstützenden Bereichen zu sorgen. Personen, die in den Datenlebenszyklus involviert sind, müssen nicht nur wissen, was sie tun, sondern ebenso dieses auch erklären können. Auch wissenschaftliche Infrastrukturleistungen wären daher als ein integraler Bestandteil der Öffentlichkeitsarbeit wissenschaftlicher Organisationen zu betrachten. Die Befähigung zu einer entsprechenden medialen Kompetenz sieht der RfII als einen wichtigen Baustein der Personalentwicklung, der frühzeitig in allen Ausbildungsstadien und „on the job" weiterentwickelt werden muss.

4.7 FÖRDERPOLITISCHE UND ORGANISATORISCHE VORAUSSETZUNGEN FÜR QUALITÄTSENTWICKLUNG

Der RfII rückt in diesem Positionspapier die Qualität von Forschungsdaten bewusst in den Mittelpunkt des wissenschaftlichen und wissenschaftspolitischen Interesses. Grundlegend hierfür ist die Annahme, dass gute Forschung und exzellente Forschungsleistungen auch im Zeitalter des digitalen Wandels in allen Disziplinen und Fächergruppen nur auf der Grundlage einer verlässlichen und qualitätsgesicherten Datenbasis gelingen können – von der Ersterhebung empirischer Forschungsdaten über Texteditionen bis hin zur Nachnutzung von Forschungsdaten zum Zwecke anschließender Forschungen oder zur Qualitätsprüfung bereits stattgefundener Forschung (Replikation). Vor diesem Hintergrund gibt der RfII auch Empfehlungen für die Förderpolitik, die Hochschulen und außeruniversitäre Forschungseinrichtungen sowie die Wissenschaftspolitik der Länder und des Bundes.

4.7.1 QUALITÄT VON FORSCHUNGSDATEN FÖRDERPOLITISCH UNTERSTÜTZEN

Förderpolitik weiterentwickeln – Ressourcen für Datenmanagement einplanen

Um Datenqualität in der Projektförderung stärker zu verankern, werden neben der DFG und dem BMBF als den beiden wichtigsten nationalen Forschungsförderern vor allem Stiftungen angesprochen. Sie könnten einen großen Beitrag zur Weiterentwicklung der Forschungsdatenqualität in Deutschland leisten. Stiftungen sollten – so wie sie es vielfach bereits auf dem Gebiet der Lehre praktizieren – mit neuen Förderformaten experimentieren, die geeignet sind, Erfahrungswerte zu etablieren, auf die öffentliche Forschungsförderung aufsetzen kann. Darüber hinaus betrachtet der RfII Datenqualität als ein Querschnittsthema, das alle Disziplinen und Forschungsthemen berührt. Passende Anforderungen sollten entsprechend von jeder Form öffentlicher Forschungsförderung im Rahmen der Antragsgestaltung gestellt und ihre Einhaltung während und nach Ablauf der Förderung auch nachgehalten werden.

a. Preise und Auszeichnungen für Beiträge zur Weiterentwicklung der Datenqualität

Im heutigen Forschungsprozess fehlen – über alle Disziplinen und Fächergruppen hinweg und insbesondere an den Universitäten – explizite Anreize, sich als Wissenschaftlerin und Wissenschaftler mit dem Thema Datenqualität und seiner Weiterentwicklung als Kernaufgabe zu befassen. Allenfalls ist sie ein Nebenprodukt der Arbeit von Professuren und Lehrstühlen mit einer methodologischen Denomination. Und auch dort ist es kein ausschlaggebendes Kriterium, das über die Besetzung einer solchen Position entscheiden würde. Stiftungen könnten hier entsprechende Anreize setzen, um durch Preise und Auszeichnungen die Beschäftigung mit Datenstandards, Datenmanagement

sowie fachadäquate IT-Lösungen zur Verbesserung der Datenqualität sichtbarer zu machen und zu belohnen.
Wenn es um Datenqualität geht, hat das Forschungsdatenmanagement den zweifelhaften Ruf einer technokratischen, wenig kreativen Angelegenheit, die zum Beispiel ein um seine Karriere bemühter vielversprechender Nachwuchswissenschaftler lieber nicht anfasst. Zuviel Management, zu wenig Reputation lautet hier die Kurzformel. Das für die Gewährleistung guter wissenschaftlicher Praxis über den gesamten Forschungsprozess hinweg elementar wichtige Aufgabenspektrum wird daher gern an wissenschaftsunterstützende Bereiche „wegdelegiert". Dabei ist gutes Forschungsdatenmanagement keine triviale Angelegenheit und sollte von technisch-administrativem und wissenschaftlichem Personal gemeinsam bewerkstelligt werden – wobei das Forschungsinteresse immer handlungsleitend sein muss. Auch hier gilt es, Best Practices und herausragende Leistungen durch öffentliche Prämierung sichtbarer zu machen, um die Attraktivität der Aufgabe auch für Forschende zu steigern. Der RfII sieht hier ein explizites Betätigungsfeld für wirtschaftsnahe Stiftungen – auch wegen der Nähe zu zertifizierten Qualitätsprozessen und Standardisierungsverfahren in der Wirtschaft.

b. Innovative Datenprodukte fördern
Datenpublikationen und andere Formen von Datenprodukten sind in Forschung und Verlagswesen noch nicht weit verbreitet. Auch hier gilt es, im Rahmen der Forschungsförderung Anreize zu setzen, um Datenaufbereitungen auf qualitativ hohem Niveau als legitimes und notwendiges Produkt des Forschungsprozesses und auf gleichem Niveau mit der Ergebnispublikation zu etablieren. Der RfII ermutigt insbesondere Stiftungen, mit Förderprogrammen zur Entwicklung von Datenprodukten in Vorleistung zu gehen, die neue Technologien der Darstellung einsetzen, aber auch langfristigen Erhalt und Nutzbarkeit dieser Produkte mitdenken.

c. Projektlaufzeiten für Datendokumentation verlängern
Mittelfristig sollten allerdings Datenprodukte in ihren verschiedenen Erscheinungsformen auch im regulären Fördergeschäft der öffentlich finanzierten Forschung zum Standard für die Rechenschaftslegung über ein erfolgreich abgeschlossenes Forschungsprojekt werden. Da dies im Rahmen der Individualförderung nicht in den üblichen Projektlaufzeiten von drei Jahren geleistet werden kann, müssten die Fördernden Möglichkeiten schaffen, zusätzliche Mittel für Edition beziehungsweise Kuratierung zu beantragen beziehungsweise die Projektförderzeiten insgesamt zu verlängern.

d. Qualität den Vorzug vor Quantität geben
Verfahren der Begutachtung von Forschungsförderanträgen sollten die Bedeutung qualitativer Parameter auch für die Datenarbeit explizit vorsehen. Zudem

sollte die Sensibilität dafür geschärft werden, dass die Güte von Publikationen auch von einer gut ausdokumentierten Datengrundlage abhängt. Einer kleineren Zahl gut dokumentiert und mit Daten hinterlegter Publikationen sollte in der Leistungsbilanz größerer Wert zugemessen werden, als einer großen Zahl von Publikationen, in denen die Datenbasis konventionell beziehungsweise oberflächlich dokumentiert ist.

4.7.2 EMPFEHLUNGEN AN DIE HOCHSCHULEN

Datenqualität in Forschungsstrategien, bei Berufungen und in den Studienordnungen verankern

Die Hochschulen gelten der Wissenschaftspolitik als „Herzkammern" des Wissenschaftssystems. Entsprechend sollte ein für die Forschung zunehmend wichtiges Feld wie die Weiterentwicklung der Datenqualität auch an den Hochschulen offensiv aufgegriffen und verfolgt werden. Der RfII spricht sich dafür aus, das Thema Datenqualität auch und gerade im Hochschulkontext mit hoher Priorität anzugehen, da hiervon mittelfristig auch die Qualität der Lehre und damit die Ausbildungsfunktion der Hochschulen für die Gesellschaft direkt betroffen ist. Erfreut zeigt sich der RfII über das gemeinsame Bekenntnis der Hochschulen zur NFDI und den Anspruch, das nationale Netzwerk aktiv mitgestalten zu wollen.

Darüber hinaus erscheint es geboten, stärker als bislang Kooperationen und vertragliche Vereinbarungen mit den kommerziellen Herstellern von Forschungsumgebungen zu suchen, um den Umgang mit eventuellen Versionswechseln oder dem Auslaufen von Produktlinien einschließlich des Supports bereits im Vorfeld und im Interesse der Wissenschaft zu regeln. Der RfII empfiehlt in diesem Zusammenhang die Etablierung von Gesprächsverbünden oder

a. Datenqualität zum Gegenstand der Forschungsstrategie machen
Viele Hochschulen und außeruniversitäre Forschungseinrichtungen haben sich die Ermöglichung exzellenter Forschungsleistungen in profilbildenden Bereichen auf die Fahnen ihrer institutionellen Entwicklungsstrategie geschrieben. Auch hat die Diskussion über Datenmanagementpläne die Hochschulen längst erreicht; Richtlinien für das Forschungsdatenmanagement sind vielerorts beschlossen und auf dem Weg zur Implementierung. Gleichwohl ist das Thema „Datenqualität" damit noch nicht zum integralen Bestandteil der Forschungsstrategie (und ggf. auch der intern in besonderer Weise geförderten Forschungsschwerpunkte) der Hochschulen geworden.
Der RfII empfiehlt, Datenqualität unter institutionellen Gesichtspunkten (etwa: mittels gezielter Ausprägung von lokalen Datenkompetenzzentren im Rahmen von Forschungsschwerpunkten oder durch standortübergreifende Vernetzung) für eine Förderung von Methoden und Datenkultur zu sorgen. Dies geschieht häufig dort, wo infrastrukturtragende Einrichtungen bereits über gemeinsame Berufungen und eine Beteiligung des Personals an der Lehre in die Hochschulen hinein vernetzt sind, und wird vom RfII einhellig begrüßt.

b. Datenexpertise bei Berufungsverfahren berücksichtigen

Bei Neuberufungen von Wissenschaftlerinnen und Wissenschaftlern auf Professuren und Lehrstühle empfiehlt der RfII, Leistungen, die bei der Einrichtung, Pflege und forschungsnahen Vernetzung von Informationsinfrastrukturen erbracht wurden, explizit als wichtige Beiträge zur Forschung zu berücksichtigen. Bei der adäquaten Bewertung von Leistungen im Bereich der Entwicklung von Datenqualität und Informationsinfrastruktur sollte stets berücksichtigt werden: Forscherinnen und Forscher, die sich hier engagieren, machen den Standort attraktiv. Hochschullehrende, die nach ihnen kommen, und vor allem Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler werden ihre Forschungslinien dann in ein bereits gut entwickeltes Umfeld einbetten können.

c. Datenqualität in den Ordnungen der Studiengänge verankern

Um die Informations- und Medienkompetenz von Studierenden insgesamt zu steigern und insbesondere das Bewusstsein für Datenqualität nachhaltig im Methodenverständnis der Fachgemeinschaften zu verankern, ist es notwendig, die Thematik so früh wie möglich in die Studiengänge zu integrieren.[101] Wie Daten entstehen, wie sie prozessiert werden und welche Verwertungs- und Nutzungshorizonte mit bedacht werden müssen, welche rechtlichen, politischen und ethischen Rahmenbedingungen in diesem Zusammenhang eine Rolle spielen, muss zum Grundwissen eines jeden Studierenden gehören. Der RfII empfiehlt, neben einer fachbezogen-methodischen Komponente in den jeweiligen Studiengängen gerade die gesellschaftlichen Implikationen des Umgangs mit Daten zu vermitteln.

4.7.3 EMPFEHLUNGEN AN DIE AUSSERUNIVERSITÄREN FORSCHUNGSEINRICHTUNGEN EINSCHLIESSLICH DER RESSORTFORSCHUNG

Durchlässigkeit weiter steigern – enge Kooperation mit Hochschulen forcieren

Die außeruniversitären Forschungsinstitute und die Ressortforschung spielen in Deutschland eine wichtige Rolle für die Weiterentwicklung von Datenqualität. Viele von ihnen sind selbst infrastrukturtragende Einrichtungen und haben in den vergangenen Jahren die Forschung mit den eigenen Infrastrukturen sowie deren Öffnung für die Fachgemeinschaften massiv ausgebaut. Der RfII empfiehlt, diesen Weg beherzt weiter zu beschreiten, Datenqualität als ein zentrales Ziel auf allen Ebenen in die eigene Forschungsstrategie zu inkorporieren und in diesem Kontext insbesondere die Kooperation mit den Hochschulen noch weiter auszubauen. Dies sollte – im Interesse eines umfassenden Kompetenzaufbaus für die gesamte deutsche Wissenschaftslandschaft – auch und gerade im forschungsnahen Infrastrukturbereich einen regelmäßigen Austausch von

[101] Siehe RfII (2019) – Digitale Kompetenzen, S. 22 f., Empfehlung 4.2.

Personal sowie gemeinsames Engagement in datenbezogener Aus- und Weiterbildung einschließen. Gerade für die Etablierung gemeinsamer verbindlicher Standards und Verfahren (zum Beispiel Datenmanagementpläne) ergeben sich aus der hier vorgeschlagenen Durchlässigkeit nach Einschätzung des RfII große Potenziale für alle Seiten: Durch den stetigen Kontakt mit universitären Forschungspartnern können außeruniversitäre Einrichtungen ihre Informationsinfrastrukturen vor einer in der Vergangenheit zuweilen beobachteten Abkopplung von wissenschaftsweiten beziehungsweise an Universitäten früh erprobten Neuerungen bewahren. Gleichzeitig können individuelle Forscherinnen und Forscher von erfolgreichen (Daten-)Managementverfahren und -praktiken lernen, die sich in der oft stärker arbeitsteiligen und verbindlicheren Organisationsstruktur außeruniversitärer Institute besser entwickeln und umsetzen lassen.

4.7.4 EMPFEHLUNGEN AN BUND UND LÄNDER

NFDI als wichtger Startpunkt für gemeinschaftliches Handeln

Bund und Länder haben als institutionelle Förderer der Wissenschaft in Deutschland eine wichtige Rolle, was die Rahmenbedingungen betrifft, in denen Forscherinnen und Forscher qualitativ auf höchstem Niveau arbeiten können. Mit der Etablierung der NFDI haben Bund und Länder gezeigt, dass Fragen der Datenqualität für sie einen hohen Stellenwert in der aktuellen und künftigen Wissenschaftspolitik haben. Nach Auffassung des RfII gilt es nun, die NFDI rechtlich und organisatorisch so auszugestalten, dass auch aus diesem Netzwerk heraus Datenqualität in allen Forschungszusammenhängen mit langem Atem gefördert werden kann.

Verstetigungsperspektiven für projektförmige Infrastrukturen und Dienste weiter prüfen

Allerdings ist die NFDI keine Einrichtung, die den zahlreichen bislang prekär – das heißt über Projektförderung befristet – finanzierten Forschungs- und Informationsinfrastrukturen einen sicheren, verstetigenden Hafen bieten kann und soll (zumindest nicht aus eigenen Mitteln). Die Frage einer dauerhaften Entwicklungsperspektive für sich erfolgreich aus Projekten an Hochschulen und Universitätsbibliotheken entwickelnden Infrastrukturen ist damit nicht gelöst. Auch in vielen Sonderforschungsbereichen und von Clustern im Rahmen der Exzellenzförderung werden wertvolle Informationsinfrastrukturen aufgebaut, die nach Ende der Förderung ihren Wert nicht verlieren müssen. Für deren Weiterführung oder auch Überführung in andere Infrastrukturzusammenhänge an den Universitäten stehen häufig kaum Ressourcen bereit. Bund und Länder sollten deshalb prüfen, ob solche Dienste – bei nachgewiesener überregionaler Bedeutung und struktureller Relevanz – mit längerfristigen Entwicklungsperspektiven ausgestattet werden können.

4.8 WEITERFÜHRUNG DES FAIR-PROZESSES

FAIR als Ausgangs-, nicht als Endpunkt

Im europäischen Forschungsraum werden derzeit die FAIR-Prinzipien (Findable, Accessible, Interoperable, Re-usable) mit Nachdruck und auch einigem Erfolg als Maßstab für gutes Forschungsdatenmanagement etabliert. Die Implementierung der FAIR-Prinzipien verfolgt primär das Ziel der Herstellung von Nutzbarkeit und Intensivierung der Nutzung von Daten. Dabei steht die Herstellung von Zugänglichkeit durch Maschinenlesbarkeit als wesentliche Ermöglichungsbedingung für Datenqualität im Fokus. Die konkrete Operationalisierung der für Wissenschaftlichkeit wesentlichen Maxime im Kanon der FAIR-Prinzipien „(meta)data meet domain-relevant community standards" wird hingegen weniger beachtet (vgl. 1.2.5).

FAIR mit fachspezifischen Qualitätsdiskursen verbinden

Der RfII sieht seitens der Forschung einen dringenden Bedarf, auch dieses Leitmotiv mit Leben zu füllen und noch darüber hinaus zu gehen. Disziplin- und forschungsfeldspezifische Gütekriterien für Datenqualität sind erforderlich (vgl. 4.2.3), und diese müssen aktiv in die infrastrukturseitigen Kuratierungs- und Archivierungsprozesse eingebracht werden. Eine Rückbindung an fachspezifische Regeln zur (Meta-)Datendokumentation ist dringend erforderlich, denn eine gute Auffindbarkeit und Teilbarkeit fachlich nicht qualitätsgesicherter oder ungeprüfter Daten würde die FAIR-Intentionen nicht befördern. Die alleinige Individualisierung des Problems durch Verschiebung in den Verantwortungsbereich des einzelnen Wissenschaftlers – sowohl in der Bereitstellung als auch im Zuge der Verwertung – ist für eine FAIRe Datenkultur nicht zielführend. Wissenschaft und Infrastruktur sind nach Auffassung des RfII im FAIR-Kontext immer zusammen zu denken.

4.8.1 WISSENSCHAFTLICHE QUALITÄTSOFFENSIVE STARTEN

Bewusstsein für inhaltlich-fachliche Qualitätsstandards steigern

Aus diesem Grund empfiehlt der RfII, die Umsetzung der FAIR-Prinzipen durch eine wissenschaftliche Qualitätsoffensive zu ergänzen, die fachangemessene Beschreibungen von Daten für eine effektive Nachnutzung engagiert vorantreibt und damit die Daten für die Forschungspraxis sichtbar qualifiziert. Gefordert sind hier alle Akteure auf nationaler und europäischer Ebene, die den FAIR-Prozess maßgeblich mittragen und umsetzen. Gerade im Rahmen der Konstituierung eines neuen Akteurs wie der NFDI empfiehlt der RfII, die Wissenschaftlichkeit als Qualitätsdimension immer schon mitzudenken und fortan parallel zu den FAIR-Prinzipien umzusetzen. So sinnvoll es zu Beginn des FAIR-Prozesses war, sich auf Fragen des Datenzugangs zu fokussieren, um eine gemeinsame Grundlage zu schaffen, auf der diverse „Open"-Initiativen aufsetzen konnten, so wichtig ist es heute, um ein verschärftes Bewusstsein für gemeinsame Qualitätsstandards nicht nur formaler, sondern auch inhaltlicher Art zu ringen. Eine FAIR ergänzende Qualitätsinitiative kann diesem Ringen auch einen verbindlichen

Anspruch verleihen, alle Kraftanstrengungen jetzt und morgen von Zugang auf Verknüpfung und mit der Verknüpfung auf Anschlussfähigkeit und Übersetzbarkeit der Daten in verschiedene wissenschaftliche und gesellschaftliche Kontexte zu erweitern.

4.8.2 DATENQUALITÄT KOMMUNIZIEREN – VERTRAUEN SCHAFFEN

Mediale Strategie gemeinsam entwickeln

Wesentlicher Bestandteil der vorgeschlagenen Qualitätsoffensive ist die freimütige und in der Sache präzise Kommunikation. Der RfII empfiehlt, wissenschaftsweit eine gemeinsame mediale Strategie zu diskutieren, in die sich alle Akteure des Wissenschaftssystems einbringen können und die von den Wissenschaftsorganisationen engagiert umgesetzt wird (siehe hierzu auch Empfehlung 4.6.4). Der RfII regt insbesondere an, datenbezogene Tätigkeiten attraktiv journalistisch aufzubereiten und massenmedial stärker präsent zu machen: Nicht selten wird genau hier das „Eigentliche" gründlicher Forschung greifbar. Denn für zahlreiche wissenschaftliche Durchbrüche bilden qualitativ hochwertig aufbereitete und analysierte Daten die Grundlage – gerade auch im Bereich interdisziplinärer Querschnittsthemen, mit welchen die Wissenschaft Antworten auf große gesellschaftliche Herausforderungen liefert (demografischer Wandel, Klimaschutz, Volkskrankheiten etc.). Ebenso sind Stichworte wie „KI" oder „Industrie 4.0" ohne das Thema „gute Daten" journalistisch unvollständig. Denn Datenqualität ist elementare Voraussetzung solcher Zukunftsthemen.

Datenintensive Forschung in die Öffentlichkeit tragen

Wissenschaftliche Einrichtungen, aber auch die individuellen Forscherinnen und Forscher müssen der Infrastruktur- und Datendimension ihrer Forschung sowohl in der eigenen Wissenschaftskommunikation und Wissenschaftsberichterstattung, bei der Beurteilung und Rekrutierung von Personal als auch in der Lehre als einem relevanten und notwendigen Treiber des wissenschaftlichen Fortschritts mehr Gewicht verschaffen. Dass dies gelingen kann, zeigt die regelmäßige Berichterstattung über die Befunde zu Einkommens- und Vermögensungleichheiten oder der Entwicklung von Armutsrisiken für unterschiedliche Bevölkerungsgruppen oder der Ergebnisse der Astrophysik und der Erforschung der Elementarteilchen der Materie – also diejenige Forschung, bei denen der Infrastrukturbezug durch die Nutzung großer sozialwissenschaftlicher Bevölkerungsumfragen oder den naturwissenschaftlichen Großgeräten (Radioteleskope, Teilchenbeschleuniger) leicht erkennbar ist. Für andere wissenschaftliche Bereiche ist die mediale Repräsentation der Datenquellen zwar schwieriger herzustellen, aber bei gezielter Anstrengung durchaus möglich.

4.8.3 EINE KULTUR DES TRANSPARENTEN EXPLIZIERENS FÖRDERN

Kultur des Explizierens in Forschungspraxis und institutionellen Leitbildern verankern

Der RfII ermutigt im Zuge der Qualitätsoffensive alle Akteure im Wissenschaftssystem im Rahmen ihrer Verantwortlichkeiten zu einer Kultur des „transparenten Explizierens" – vom Betreuer einer Qualifikationsarbeit über das Management einer wissenschaftlichen Einrichtung bis zur Informationsinfrastruktur beziehungsweise zum Verlag und zur Forschungsförderung. Informationen zum Qualitätssicherungsprozess sollten ebenso selbstverständlich verfügbar gemacht werden, wie beispielsweise die wissenschaftliche Erhebungsmethode. Des Weiteren sollten wissenschaftliche Fachgemeinschaften eine angepasste Kultur des Referenzierens beziehungsweise Zitierens von Daten(-beständen) verabreden, die sowohl qualitätssichernd im Hinblick auf die Forschung als auch reputationsförderlich für die datenproduzierende Seite und die beteiligten Forschenden ist. Der RfII betrachtet eine so verstandene Kultur des transparenten Explizierens von Daten als eine wesentliche Ergänzung der FAIR-Prinzipien und einen wichtigen Beitrag zur guten wissenschaftlichen Praxis. Er empfiehlt, dies entsprechend in institutionellen Leitbildern für gutes wissenschaftliches Arbeiten festzuhalten und im Forschungsalltag der Hochschulen und außeruniversitären Forschungsinstitute aktiv dafür zu werben. Von der NFDI und ihren Konsortien erwartet er hierzu in der Zukunft wichtige Impulse, die nicht nur im nationalen Rahmen, sondern auch im europäischen Forschungsraum Resonanz auslösen sollten.

LITERATURVERZEICHNIS

Amann, Rudolf I. et al. (2019): Toward Unrestricted Use of Public Genomic Data, in: Science, Jg. 363, Nr. 6425, S. 350–352, DOI: 10.1126/science.aaw1280, zuletzt geprüft am: 30.08.2019.

Baker, Karen S./Duerr, Ruth E./Parsons, Mark A. (2015): Scientific Knowledge Mobilization. Co-evolution of Data Products and Designated Communities, in: International Journal of Digital Curation, Jg. 10, Nr. 2, S. 110–135, DOI: 10.2218/ijdc.v10i2.346, zuletzt geprüft am: 30.08.2019.

Carpenter, Todd (2017): What Constitutes Peer Review of Data – A survey of published peer review guidelines, 17 S., online verfügbar unter: https://arxiv.org/ftp/arxiv/papers/1704/1704.02236.pdf, zuletzt geprüft am: 30.08.2019.

CCSDS – Consultative Committee for Space Data Systems (2012): Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data System Practices. CCSDS 650.0-M-2, Washington, 135 S., online verfügbar unter: https://public.ccsds.org/pubs/650x0m2.pdf, zuletzt geprüft am: 30.08.2019.

Cousijn, Helena/Cruse, Patricia/Fenner, Martin (2018): Taking Discoverability to the Next Level: Datasets with DataCite DOIs Can Now Be Found Through Google Dataset Search, in: DataCite Blog, DOI: 0.5438/5AEP-2N86, zuletzt geprüft am: 30.08.2019.

Daston, Lorraine/Galison, Peter (2007): Objektivität, 1. Aufl., Frankfurt am Main: Suhrkamp, 530 S.

DBV – Deutscher Bibliotheksverband – Sektion 4 (2018): Wissenschaftliche Bibliotheken 2025, dbv, 22 S., online verfügbar unter: http://www.bibliotheksverband.de/fileadmin/user_upload/Sektionen/sektion4/Publikationen/WB2025_Endfassung_endg.pdf, zuletzt geprüft am: 30.08.2019.

DFG – Deutsche Forschungsgemeinschaft (2017): Replizierbarkeit von Forschungsergebnissen. Eine Stellungnahme der Deutschen Forschungsgemeinschaft, Bonn, 5 S., online verfügbar unter: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_de.pdf, zuletzt geprüft am: 30.08.2019.

DFG – Deutsche Forschungsgemeinschaft (2019): Leitlinien zur Sicherung guter wissenschaftlicher Praxis (Kodex), Bonn, 32 S., online verfügbar unter: https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf, zuletzt geprüft am: 30.08.2019.

DINI – Deutsche Initiative für Netzwerkinformationen e. V. (2018): Thesen zur Informations- und Kommunikationsinfrastruktur der Zukunft, 24 S., DOI: 10.18452/19126, zuletzt geprüft am: 30.08.2019.

Europäisches Parlament/Rat der Europäischen Union (2019): Richtlinie (EU) 2019/1024 des Europäischen Parlaments und des Rates vom 20. Juni 2019 über offene Daten und die Weiterverwendung von Informationen des öffentlichen Sektors (Neufassung), in: Amtsblatt der Europäischen Union, online verfügbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32019L1024&from=EN, zuletzt geprüft am: 30.08.2019.

Fecher, Benedikt/Friesike, Sascha/Hebing, Marcel (2015): What Drives Academic Data Sharing?, in: PLoS ONE, Jg. 10, Nr. 2, S. 1–25, DOI: 10.1371/journal.pone.0118053, zuletzt geprüft am: 30.08.2019.

Field, Laurence et al. (2013): Realising the Full Potential of Research Data. Common Challenges in Data Management, Sharing and Integration Across Scientific Disciplines. Version 3, 15 S., online verfügbar unter: http://orca.cf.ac.uk/66034/1/ESFRI_Common_Challenges_v1.pdf, zuletzt geprüft am: 30.08.2019.

Hodson, Simon et al. (2018): FAIR Data Action Plan. Interim Recommendations and Actions From The European Commission Expert Group On Fair Data, 21 S., DOI: 10.5281/ZENODO.1285290, zuletzt geprüft am: 30.08.2019.

KE – Knowledge Exchange (2014): Sowing the Seed. Incentives and Motivations for Sharing Research Data, a Researcher's Perspective, Kopenhagen, 48 S., online verfügbar unter: http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf, zuletzt geprüft am: 30.08.2019.

King, Gary/Persily, Nate (2019): A New Model for Industry-Academic Partnerships, 16 S., DOI: 10.1017/S1049096519001021, zuletzt geprüft am: 30.08.2019.

Kleiner, Matthias (2010): „Qualität statt Quantität" – Neue Regeln für Publikationsangaben in Förderanträgen und Abschlussberichten. Pressekonferenz, DFG- Deutsche Forschungsgemeinschaft, Berlin, 6 S., online verfügbar unter: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2010/statement_qualitaet_statt_quantitaet_mk_100223.pdf, zuletzt geprüft am: 30.08.2019.

Klimpel, Paul (2015): Eigentum an Metadaten? Urheberrechtliche Aspekte von Bestandsinformationen und ihre Freigabe, in: Euler, Ellen et al. (Hg.): Handbuch Kulturportale. Online-Angebote aus Kultur und Wissenschaft, Berlin, Boston, S. 57–64, online verfügbar unter: https://irights.info/wp-content/uploads/2016/01/Klimpel-2015-Eigentum-an-Metadaten.pdf, zuletzt geprüft am: 30.08.2019.

Koepler, Oliver et al. (2018): Thesenpapier Nationale Forschungsdateninfrastruktur für die Chemie (NFDI4Chem), 15 S., DOI: 10.5281/ZENODO.1404201, zuletzt geprüft am: 30.08.2019.

Lauber-Rönsberg, Anne/Krahn, Philipp/Baumann, Paul (2018): Gutachten zu den rechtlichen Rahmenbedingungen des Forschungsdatenmanagements, 23 S., online verfügbar unter: https://tu-dresden.de/gsw/jura/igetem/jfbimd13/ressourcen/dateien/publikationen/DataJus_Zusammenfassung_Gutachten_12-07-18.pdf?lang=de, zuletzt geprüft am: 30.08.2019.

Lazer, David/Kennedy, Ryan/Vespignani, Alessandro (2014): The Parable of Google Flu: Traps in Big Data Analysis, in: Science, Nr. 343, S. 1203–1205, online verfügbar unter: https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf, zuletzt geprüft am: 30.08.2019.

Liggesmeyer, Peter (2009): Software-Qualität. Testen, Analysieren und Verifizieren von Software, 2. Aufl., Heidelberg: Spektrum Akademischer Verlag, online verfügbar unter: http://dx.doi.org/10.1007/978-3-8274-2203-3, zuletzt geprüft am: 30.08.2019.

Lipton, Zachary C. (2016): The Mythos of Model Interpretability, 9 S., online verfügbar unter: https://arxiv.org/pdf/1606.03490v3.pdf, zuletzt geprüft am: 30.08.2019.

Neuroth, Heike et al. (2012): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme, Boizenburg/Göttingen: Hülsbusch/Universitätsverlag, 380 S., online verfügbar unter: https://publiscologne.th-koeln.de/files/428/Publikation_Osswald_Langzeitarchivierung_Bestandsaufnahme.pdf, zuletzt geprüft am: 30.08.2019.

Parsons, M. A./Fox, P. A. (2013): Is Data Publication the Right Metaphor?, in: Data Science Journal, Jg. 12, S. 32–46, DOI: 10.2481/dsj.WDS-042, zuletzt geprüft am: 30.08.2019.

Peer, Limor/Green, Ann/Stephenson, Elizabeth (2014): Committing to Data Quality Review, in: IJDC – International Journal of Digital Curation, Jg. 9, Nr. 1, S. 263–291, DOI: 10.2218/ijdc.v9i1.317, zuletzt geprüft am: 30.08.2019.

Pfaffenberger, Fabian (2016): Twitter als Basis wissenschaftlicher Studien: Eine Bewertung gängiger Erhebungs- und Analysemethoden der Twitter-Forschung, Wiesbaden: Springer, 146 S.

RatSWD (2018): Tätigkeitsbericht 2017 der vom RatSWD akkreditierten Forschungsdatenzentren (FDZ), Berlin, 44 S., DOI: 10.17620/02671.33, zuletzt geprüft am: 30.08.2019.

RfII – Rat für Informationsinfrastrukturen (2016): Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, 160 S., online verfügbar unter: http://d-nb.info/1104292440/34, zuletzt geprüft am: 30.08.2019.

RfII – Rat für Informationsinfrastrukturen (2017): Entwicklung von Forschungsdateninfrastrukturen im internationalen Vergleich. Bericht und Anregungen, Göttingen, 94 S., online verfügbar unter: http://d-nb.info/1143737180/34, zuletzt geprüft am: 30.08.2019.

RfII – Rat für Informationsinfrastrukturen (2019): Digitale Kompetenzen – dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft, Göttingen, 56 S., online verfügbar unter: http://d-nb.info/1192391217/34, zuletzt geprüft am: 30.08.2019.

RfII – Rat für Informationsinfrastrukturen (2019): Stellungnahme des RfII zu aktuellen Entwicklungen rund um Open Data und Open Access, Göttingen, 8 S., online verfügbar unter: http://d-nb.info/1186295503/34, zuletzt geprüft am: 30.08.2019.

Samek, Wojciech/Wiegand, Thomas/Müller, Klaus-Robert (2017): Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, 8 S., online verfügbar unter: https://arxiv.org/pdf/1708.08296.pdf, zuletzt geprüft am: 30.08.2019.

Strohschneider, Peter (2018): Rede anlässlich des Neujahrsempfangs der DFG, DFG – Deutsche Forschungsgemeinschaft, Berlin, 6 S., online verfügbar unter: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2018/180115_rede_strohschneider_neujahrsempfang_de.pdf, zuletzt geprüft am: 30.08.2019.

Stuart, David et al. (2018): Practical Challenges for Researchers in Data Sharing (Whitepaper), 17 S., DOI: 10.6084/M9.FIGSHARE.5975011, zuletzt geprüft am: 30.08.2019.

Swan, Alma/Brown, Sheridan (2008): The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs. Report to the Jisc, 34 S., online verfügbar unter: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf, zuletzt geprüft am: 30.08.2019.

Wang, Richard Y. (1998): A Product Perspective on Total Data Quality Management, in: CACM – Communications of the ACM, Jg. 41, Nr. 2, S. 58-65, DOI: 10.1145/269012.269022, zuletzt geprüft am: 30.08.2019.

Wang, Richard Y./Strong, Diane M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers, in: Journal of Management Information Systems, Jg. 12, Nr. 4, S. 5–33, online verfügbar unter: http://www.jstor.org/stable/40398176, zuletzt geprüft am: 30.08.2019.

Whyte, Angus/Pryor, Graham (2011): Open Science in Practice: Researcher Perspectives and Participation, in: IJDC – International Journal of Digital Curation, Jg. 6, Nr. 1, S. 199–213, DOI: 10.2218/ijdc.v6i1.182, zuletzt geprüft am: 30.08.2019.

Wilkinson, Mark D. et al. (2016): The FAIR Guiding Principles for Scientific Data Management and Stewardship, in: Scientific Data, Jg. 3, S. 1–9, DOI: 10.1038/sdata.2016.18, zuletzt geprüft am: 30.08.2019.

Wouters, Paul/ Wouter Haak (2017): Open Data. The Researcher Perspective. Hg. v. CWTS- Leiden University's Centre for Science and Technology Studies und Elsevier, 48 S., online verfügbar unter: https://www.elsevier.com/__data/assets/pdf_file/0004/281920/Open-data-report.pdf, zuletzt geprüft am: 30.08.2019.

WR – Wissenschaftsrat (2012): Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. Drs. 2359-12, Berlin, 90 S., online verfügbar unter: http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf, zuletzt geprüft am: 30.08.2019.

WR – Wissenschaftsrat (2015): Empfehlungen zu wissenschaftlicher Integrität. Positionspapier. Drs. 4609-15, Köln, 54 S., online verfügbar unter: http://www.wissenschaftsrat.de/download/archiv/4609-15.pdf, zuletzt geprüft am: 30.08.2019.

WR – Wissenschaftsrat (2018): Empfehlungen zur Internationalisierung von Hochschulen. Drs. 7118-18, WR – Wissenschaftsrat, München, 141 S., online verfügbar unter: https://www.wissenschaftsrat.de/download/archiv/7118-18.pdf, zuletzt geprüft am: 30.08.2019.

ONLINE-RESSOURCEN

Cochrane Deutschland
https://www.cochrane.de/de/cochrane

COPDESS – Coalition for Publishing Data in the Earth and Space Sciences
http://www.copdess.org

Deutscher Hochschulverband, Barometer
https://www.hochschulverband.de

Deutscher Museumsbund e. V.
https://www.museumsbund.de

Deutsches Textarchiv
http://www.deutschestextarchiv.de/

ESA Sentinel Online
https://sentinel.esa.int/web/sentinel/home

Force 11
https://www.force11.org

Forschungsdaten.org
https://www.forschungsdaten.org

GEO-Wiki
https://www.geo-wiki.org

GFBio – German Federation for Biological Data
https://www.gfbio.org/about

GFZ Data Services
http://dataservices.gfz-potsdam.de/portal/about.html

Go FAIR Initiative
https://www.go-fair.org

ICSU- World Data System
http://www.icsu-wds.org

Institut für Dokumentologie und Editorik
https://www.i-d-e.de

Schema.org – Community Website
https://schema.org

Social Science One
https://socialscience.one

SpringerNature
https://www.springernature.com

Statistische Ämter des Bundes und der Länder (Forschungsdatenzentren)
https://www.forschungsdatenzentrum.de/de

The Open Definition
https://opendefinition.org

UAG Datenmanagementpläne der DINI-nestor AG Forschungsdaten
https://www.forschungsdaten.org/index.php/UAG_Datenmanagementpläne

Visual6502-Projekt
http://www.visual6502.org

W3C Standards
https://www.w3.org

Wikidata
https://www.wikidata.org

ANHANG

A. ZUR GENESE VON KONZEPTEN DER DATENQUALITÄT UND IHRES EINSATZES IN DER WISSENSCHAFT

Ergebnisdokumentation der AG Datenqualität des RfII
(August 2017 bis März 2019, mit Aktualisierungen)

INHALT

ABKÜRZUNGSVERZEICHNIS

BMBF	Bundesministerium für Bildung und Forschung
CLARIN	Common Language Resources and Technology Infrastructure
CTS	Core Trust Seal
DCC	Digital Curation Center
DFG	Deutsche Forschungsgemeinschaft
DIN	Deutsches Institut für Normung
DSA	Data Seal of Approval
EOSC	European Open Science Cloud
EU	Europäische Union
FAIR	Findable, Accessible, Interoperable, Reusable
FDM	Forschungsdatenmanagement
HEFCE	Higher Education Funding Councils
ICSU	International Council for Science
INSPIRE	Infrastructure for Spatial Information in Europe
ISO	International Organization for Standardization
MARC	Machine-Readable Cataloging
NFDI	Nationale Forschungsdateninfrastruktur
NID	Normenausschuss Information und Dokumentation
OAIS	Open Archival Information System
RatSWD	Rat für Sozial- und Wirtschaftsdaten
RDA	Research Data Alliance
RfII	Rat für Informationsinfrastrukturen
TRAC	Trustworthy Repositories Audit & Certification
WR	Wissenschaftsrat
XML	Extensible Markup Language

1 EINLEITUNG

Der Rat für Informationsinfrastrukturen hat sich von August 2017 bis März 2019 intensiv mit existierenden Ansätzen der Qualitätssicherung und Qualitätssteigerung von Daten auseinandergesetzt. Dabei wurde umfangreiches Material gesammelt, um die Entstehung von Datenqualitätskonzepten aus Perspektive des Wissenschaftssystems nachzuvollziehen. Der vorliegende Anhang zum Positionspapier „Herausforderung Datenqualität" bietet hierzu eine strukturierte Übersicht. In die Ausarbeitung des Anhangs sind auch Ergebnisse von Fachgesprächen eingegangen, die die Arbeitsgruppe mit Expertinnen und Experten aus verschiedenen wissenschaftlichen Bereichen geführt hat. Der RfII dankt den Gästen der Arbeitsgruppe für zahlreiche wertvolle Hinweise und Einsichten. Er hofft, dass diese Ergebnisdokumentation für Wissenschaftlerinnen und Wissenschaftler sowie für Beschäftigte im wissenschaftsunterstützenden Infrastrukturbereich hilfreich ist, um sich einen Überblick über bestehende Konzepte der Datenqualität, ihre Herkunft, Limitierungen und Potenziale für die Wissenschaft zu verschaffen.

Aus dem vom RfII ausgewerteten Material sollte deutlich werden, dass in den zurückliegenden Jahrzehnten verschiedene, oftmals kaum aufeinander bezogene Ansätze für Qualitätssicherung mit Bezug zum Datenbereich entwickelt worden sind. Auch stammen frühe Datenqualitätskonzepte nicht aus der Wissenschaft selbst, was ihre Adaption in der Forschung erschwert, zumindest aber sehr voraussetzungsvoll macht. Denn sie orientieren sich nicht an der Logik von – stets auch ergebnisoffenen, revidierbaren und nachhaltig aufeinander aufbauenden – Forschungsprozessen, sondern an industriellen Produktentwicklungs- und Fertigungsprozessen. Somit setzt die aktuelle Diskussion über Datenqualität in der Wissenschaft nicht nur eine kritische Reflexion bestehender und sich entwickelnder Qualitätskonzepte und-ansätze voraus, sondern auch eine eigenständige, der Spezifik des Wissenschaftssystems entsprechende, Vorstellung von Qualität im Allgemeinen und Datenqualität im Besonderen.

2 QUALITÄTSBEGRIFF UND QUALITÄTSKONZEPTE

2.1 QUALITÄT UND WISSENSCHAFT

Das in der Wissenschaft verbreitete Qualitätsdenken beruht auf einem professionellen Selbstverständnis, das sich auf ein fachliches und auf hohem Niveau diskussionsoffen und kontrolliert fortentwickeltes Methodenwissen gründet. Demzufolge liegt es einerseits in der Verantwortung von wissenschaftlichen Communities/Fachgemeinschaften, eine hohe Qualität der Forschungsprozesse und -ergebnisse zu gewährleisten. Andererseits kommt der intrinsischen Motivation der Forschenden selbst, ihrer „Redlichkeit", eine wichtige Rolle zu.

Eine Reflexion über eine „Qualität" von Wissenschaft, die sich nicht nur auf Methoden, sondern auf vielfältige Gütekriterien und auch auf quantitativ messbare Indikatoren bezieht, hatte sich bereits vor dem massenhaften Einsatz elektronischer Datenverarbeitung intensiviert. Entsprechende Maßstäbe und Konzepte eines dezidierten „Qualitätsmanagements" stammten häufig aus der Managementtheorie, der Verfahrens- und Produktoptimierung in der industriellen Produktion oder bereits gelebter Praxis in und zwischen Wirtschaftsunternehmen. Erste Reflexionen zu „Datenqualität" setzten erst mit der Zunahme digitaler Angebote seit den 1990er Jahren ein. Vor allem ist seither die *Sicherung* von Qualität ein Thema, wobei der diesbezügliche Austausch sich oftmals in einem sehr eng gefassten Wortsinn auf „Infrastrukturfragen" beschränkt.

Hier stammen Anregungen nicht nur aus den Wirtschaftswissenschaften und der Wirtschaftsinformatik. Auch die „Wissensspeicher" der Wissenschaft, wie zum Beispiel Bibliotheken und Museen sowie internationale Fachverbände in den Naturwissenschaften sind in der Vereinbarung von Normen oder das Setzen von De-facto-Standards erfolgreich gewesen. Selten wird jedoch in diesen Ansätzen das für den Wissenschaftsprozess insgesamt wichtige Thema der fach- und domänenübergreifenden *Steuerung* einer Qualität von (digitalen) Daten diskutiert oder aber die Rolle von Qualität in einem umfassenderen Transformationsprozess von Wissenschaft und Gesellschaft insgesamt behandelt.

Ein für den Einsatz im Wissenschaftssystem operationalisierbarer Begriff von Datenqualität ist nach Auffassung des RfII prozessual angelegt. Er berücksichtigt die wechselseitige Steigerungsfähigkeit und Verknüpfung von technologischem Fortschritt in den Infrastrukturen und Forschungsumgebungen auf der einen und der Forschungsprozesse auf der anderen Seite (siehe die Begriffsdefinition des RfII).[1] Vor diesem Hintergrund eines offenen und dynamischen Qualitätsverständnisses in Sachen Forschungsdaten hat sich der RfII mit den im Folgenden beschriebenen Datenqualitätskonzepten auseinandergesetzt.

[1] Siehe Begriffsbestimmungen, Anhang A.2.

2.2 NORMIERUNG UND DAS SETZEN VON STANDARDS

In der Technologieentwicklung werden klassisch durch Normierung (Mindest-) Standards und damit Qualitätsmaße geschaffen: „Gut“ ist, was der Norm beziehungsweise dem Standard entspricht und damit in der Breite funktionsgerecht verwendbar ist. Für die Datenqualität in der Wissenschaft sind beispielsweise Normsetzungen und Standardisierungen im Bereich der Kommunikations- und Informationstechnologie wie auch im Bereich der Dokumentation und Erschließung wissenschaftlicher Information relevant. In beiden Bereichen kommen technische „de jure“ Standards, wie die DIN-Norm zur Anwendung, ebenso wie eine Vielzahl sogenannter De-facto-Standards,[2] die sich über Anwendungen und Akzeptanz verbreiten.

Das Instrument einer flächendeckenden, Kompatibilität und Qualität garantierenden Normierung entstammt historisch gesehen der Welt der maschinellen Bauteile, ist also eine Errungenschaft des industriellen Maschinenbaus.[3] Zügig durchgesetzt hat sich aber auch eine in gleicher Weise durchgeführte Normierung – und zwar verbindlich ausgehandelte und niedergelegte unmissverständliche Definitionen und detaillierte Durchführungsvorschriften – für Prozesse. Überwiegend sind dies Normierungen, die im Ursprung Regelungsbedarfen der Industrie oder anderer Wirtschaftszweige sowie der öffentlichen Grundversorgung entsprechen. Sie finden auch in Behörden und der Wissenschaft Anwendung. So begegnet die Wissenschaft auch in digitalen beziehungsweise digital unterstützten Forschungsprozessen vielfach Normen, etwa des *Deutschen Instituts für Normung* (DIN-Normen), vor allem aber internationalen Normen, zum Beispiel denen der drei *europäischen Kommissionen für Normung*[4] oder der *International Organization for Standardization* (ISO). Die Adaption von DIN/ISO-Prozessnormierungen kann sich in der Wissenschaft jedoch als beeinträchtigend erweisen; namentlich in der Grundlagenforschung steht sie Diversität und Innovation in der Methodenentwicklung unter Umständen sogar entgegen.

Gleichwohl finden sich unter dem Dach von DIN und ISO wissenschafts- und administrationsorientierte Interessensgemeinschaften, wie zum Beispiel der 1927 gegründete Normenausschuss Information und Dokumentation NID. Der NID ist für die nationale Normung unter anderem der Erstellung, Erhaltung und Publikation von Dokumenten und Daten im Bereich des Informationswesens, insbesondere auch im Archiv-, Bibliotheks-, Dokumentations-, Museums- und Verlagswesens zuständig. Er engagiert sich auch im gleichnamigen

---

2 Das Format doc ist ein Beispiel für proprietäre De-facto-Standards im Besitz einzelner Unternehmen; XML für einen offenen im Word Wide Web Consortium (W3C) entwickelten De-facto- bzw. inzwischen De-jure-Standard. Zur Unterscheidung von De-jure- und De-facto-Standards im Software-Bereich vgl. unter anderem Taylor et al. (2010) – Software Architecture, S. 622.

3 1917 entstand als erster seiner Art der privatwirtschaftlich initiierte Verein *Normenausschuss der Deutschen Industrie*, der Vorläufer des heutigen DIN – Deutsches Institut für Normung e. V. Die internationale ISA (Vorläuferin der *International Organization for Standardization* [ISO] mit Sitz in Genf) wurde 1926 gegründet. 1918 wurde mit DIN 1 die berühmte erste Norm für einen Kegelstift geschaffen, der als konisches Element Maschinenteile verbindet.

4 Europäisches Komitee für Normung (CEN), Europäisches Komitee für elektrotechnische Normung (CENELEC), Europäisches Institut für Telekommunikationsnormen (ETSI).

internationalen Pendant, dem Technical Committe „Information and Documentation“ bei der ISO.[5] Seit seiner Gründung hat sich das Aufgabenspektrum des NID stetig erweitert. In jüngerer Zeit befasste er sich auch immer wieder mit Fragen der Digitalisierung/Digitalität.

In der Definition des Deutschen Instituts für Normung ist eine Norm zunächst ein Dokument, das Anforderungen an Produkte, Dienstleistungen oder Verfahren festlegt.[6] Die Anwendung einer DIN-Norm wird als grundsätzlich freiwillig angesehen. Dennoch stellen Normen eine hohe, quasigesetzliche Verbindlichkeit her: Wenn Normen auch international zum Inhalt von Verträgen werden oder wenn der Gesetzgeber ihre Einhaltung zwingend vorschreibt, werden Normen bindend. Im Zuge der Globalisierung kooperieren nationale und internationale Einrichtungen bei Normsetzungsprozessen. Die ISO hat als Standardisierungsorganisation dabei ein besonderes Gewicht erhalten. Ihr gehören aktuell 161 nationale Einrichtungen (*national standards bodies*) an. Sowohl die ISO- als auch die DIN-Normierung setzen zur Festlegung eines ISO *„standards“* einen Expertenkonsens in einem partizipativen, quasi „parlamentarisch“ angelegten Gremiengang voraus.[7]

Neben der Norm hat das DIN die „DIN SPEC“ als eine etwas „weichere“ Form der Standardisierung eingeführt.[8] Anders als bei DIN-Normen besteht für die DIN SPEC keine Konsenspflicht. Sie kann innerhalb weniger Monate erarbeitet werden, vorausgesetzt, mindestens drei Parteien beziehungsweise Marktteilnehmer wirken mit. Das DIN überwacht, dass die DIN SPEC nicht mit bestehenden Normen kollidiert, und veröffentlicht sie. Damit reagiert die DIN auf Bedarfe in deutschen Qualitätsdiskursen, auch in der Wissenschaft, die oftmals keine starre Norm setzen, sondern in einer Community nur schnell anwendbare, praktikable Standards definieren will.

DIN-Normen finden sich heute für zahlreiche Bereiche der digitalen Forschung und erhalten auch im Zusammenhang mit der Frage nach der Datenqualität eine Bedeutung. So bilden Normen für Dateiformate, Datenträger und ihre Vernichtung (DIN 66399 „Büro- und Datentechnik – Vernichtung von Datenträgern“), standardisierte Referenzmodelle für die Langzeitarchivierung (Open Archival Information System OAIS, ISO 14721), aber auch Regelungen für „Thesauri and interoperability with other vocabularies“ (ISO 25964) sowie insbesondere Normen zur Modellierung und Codierung von Metadaten (etwa Dublin Core, ISO 15836) ein allmählich wachsendes Rahmenwerk, das als Basis für Digitalisierungsprozesse dienen kann.[9]

5 https://www.din.de/de/mitwirken/normenausschuesse/nid (zuletzt geprüft am: 30.08.2019).

6 https://www.din.de/de/ueber-normen-und-standards/basiswissen (zuletzt geprüft am: 30.08.2019).

7 „Developing ISO standards is a consensus-based approach and comments from all stakeholders are taken into account“; https://www.iso.org/developing-standards.html (zuletzt geprüft am: 30.08.2019).

8 https://www.din.de/de/ueber-normen-und-standards/basiswissen (zuletzt geprüft am: 30.08.2019).

9 Für Einzelnachweise vergleiche das Verzeichnis der Normen und Standards am Ende des Dokuments.

Die ISO-Norm 25964 („Thesaurus-Norm“) verdeutlicht exemplarisch, wie Digitalität und technologischer Fortschritt in der Standardisierung ihren Niederschlag finden. In der Dokumentationswissenschaft werden Thesauri, also kontrollierte Vokabulare, als Hilfsmittel für die Sacherschließung genutzt (indexieren, speichern, finden). Die digitale Informationsverarbeitung bietet mittels Thesauri insbesondere die Möglichkeit für dynamische Verknüpfungen zwischen verwandten Begriffen. Diese neuen technischen Möglichkeiten sind in der 2011 und 2013 veröffentlichten, zweiteiligen und maßgeblich durch die Wissenschaft ausgestalteten ISO-Norm 25964 „Thesauri and interoperability with other vocabularies“ abgebildet, die auch Richtlinien zur Thesaurusföderation („Mappings“) enthält. Diese neue, bereits auf Belange wie die Nutzung semantischer Webtechnologien abstellende Norm ersetzt seitdem zwei ältere ISO-Normen für ein- und mehrsprachige Thesauri aus den 1970er und 1980er Jahren.[10]

Zu den prominentesten Vorgaben für wissenschaftsweit, fachübergreifend, interoperable digitale Informationssysteme zählen Normen und Standards für Metadaten im Bibliothekswesen. Das Konzept der Metadaten (also von beziehungsweise in Datensätzen fixierten Daten über Daten) entstand aus Überlegungen der Library of Congress in Washington, die in den späten 1950er Jahren damit begonnen hatte, nach Möglichkeiten der Verwendung von Automatisierungstechniken bei internen Arbeitsabläufen zu forschen. In den 1960er Jahren wurde das bibliografische Datenformat MARC (MAchine-Readable Cataloging) entwickelt, das nun in der Version MARC 21 weltweit genutzt wird. Die Entwicklung der Metadatenstandards beschleunigte sich in den 1990er Jahren, folgte aber häufig keinem international und über Domänengrenzen hinweg verbindlichen Pfad. Es gab keine Begrenzung für den Typ oder die Menge der Ressourcen, die durch Metadaten beschrieben werden sollten, wie auch keine Begrenzung für die Anzahl der sich überschneidenden Metadatenstandards für jede Art von Ressourcen beziehungsweise Subjektdomäne oder für die Arten von Berufen beziehungsweise Themenbereichen, die an der Entwicklung und Anwendung von Metadatenstandards hätten beteiligt sein sollen.[11] Breit und fächerübergreifend akzeptiert ist heute lediglich der Dublin Core Standard als Basis für die Beschreibung jeglicher Art von Dokumenten. Die 15 Kernelemente sind seit 2009 als ISO-Standard 15836 anerkannt. Er wurde Mitte der 1990er Jahre erstmals entworfen und wird heute gemeinschaftlich unter dem Dach der non-profit Dublin Core Metadata Initiative weiterentwickelt und verbreitet.[12] Im deutschsprachigen Raum organisiert der Standardisierungsausschuss an der Deutschen Nationalbibliothek den Einsatz einheitlicher Standards für die Erschließung, Formate und Schnittstellen in Bibliotheken und koordiniert auch die Internationalisierung der eingesetzten Standards.[13]

Auch in den wissenschaftlichen Fächern und Disziplinen finden sich eigene, teils sehr starke Standardisierungsinitiativen. Oftmals bestehen Querbezüge zum staatlich oder hoheitlich

10 ISO 2788 und ISO 5964, für Einzelnachweise vgl. Verzeichnis der Normen und Standards.

11 http://www.metadataetc.org/metadatabasics/overview.htm (zuletzt geprüft am: 30.08.2019).

12 http://dublincore.org/about/ (zuletzt geprüft am: 30.08.2019).

13 https://www.dnb.de/DE/Professionell/Standardisierung/Standardisierungsausschuss/standardisierungsausschuss_node.html (zuletzt geprüft am: 30.08.2019).

organisierten Normierungswesen, zum Beispiel wenn wissenschaftseigene Normen zur Anerkennung als DIN- oder ISO-Norm vorgeschlagen werden. Dies geschieht teils aus pragmatischen Gründen, um den Prozess der Weiterentwicklung zu stabilisieren, teils wegen einer durch die Normierung erhofften hohen Schlagkraft und Akzeptanz „im System". Die folgenden Schlaglichter verdeutlichen das Ineinander von Normierung und normartig „gesetzten" Standards der Wissenschaft, wie auch ihre teils durch die technische Entwicklung beeinflusste Genese:

- Das *Referenzmodell CIDOC-CRM* legt für den Bereich der materiellen Kultur, besonders für Museen, Definitionen und formale Strukturen zur Beschreibung der impliziten und expliziten Konzepte und Beziehungen fest. Das Datenmodell wird in der Dokumentation des Kulturerbes verwendet, mit besonderem Schwerpunkt auf Informationsaustausch. Die Initiative ging vom International Council of Museums (ICOM) aus. Das vollständige Referenzmodell wurde 1999 erstmals publiziert, 2006 in einen ISO-Standard überführt.[14]

- Im Bereich der Langzeitarchivierung ist das *Reference Model for an Open Archival Information System – OAIS* eine breit akzeptierte Grundlage für die Gestaltung von Diensten. Das „open" steht hier für die Offenheit des Entwicklungsprozesses. Die erste Version von 2002 wurde vom Consultative Committee for Space Data Systems vorgelegt (einer multinationalen Organisation der Raumfahrtagenturen) und sehr schnell in einen ISO-Standard überführt.[15] Aus den Bibliotheken und Archiven folgte kurze Zeit darauf die erste Norm zur Zertifizierung nach dem OAIS-Standard (zur weiteren Entwicklung der prozessbezogenen Normen vgl. 2.4).

- Mit den *Guidelines for Electronic Text Encoding and Interchange* wurde vor knapp 30 Jahren in der Literaturwissenschaft ein heute viel verwendeter De-facto-Standard für digitale Repräsentationen verschiedener Textformen geschaffen.[16] Die *Text Encoding Initiative* (heute: TEI Consortium) startete 2005 eine Initiative zusammen mit der ISO, um die Merkmalstrukturen der ISO-Norm 24610 „Sprachressourcen – Merkmalstrukturen" in die eigenen Richtlinien zu übernehmen.[17] Hier wurden also auf Betreiben einer Forschungscommunity Teile einer amtlichen Norm in die wissenschaftliche Standardsetzung zurücküberführt.

- In den Sozialwissenschaften bietet die *Data Documentation Initiative* seit den frühen 1990er Jahren Regeln für die Beschreibung von Daten und Protokollen entlang des Datenlebenszyklus. Ähnlich wie die *Text Encoding Initiative* werden mit dem Aufkommen der Webtechnologie Dokumentenformate zur Codierung und zum Austausch von Texten basierend auf XML spezifiziert. Treiber für die Entwicklung des *Codebook* waren über die Jahrzehnte vor allem die sozialwissenschaftlichen Datenarchive.[18]

---

14 http://www.cidoc-crm.org/home (zuletzt geprüft am: 30.08.2019).

15 http://www.oais.info/ (zuletzt geprüft am: 30.08.2019).

16 https://tei-c.org/about/history/ (zuletzt geprüft am: 30.08.2019).

17 http://www.tei-c.org/activities/council/reports/tcr03-report-of-the-tei-council-to-the-members-meeting-2005/ (zuletzt geprüft am: 30.08.2019).

18 Zur Entwicklungsgeschichte vgl. https://www.ddialliance.org/what/history.html (zuletzt geprüft am: 30.08.2019).

- Das *internationale Klassifikationssystem für ikonographische Forschung ICONCLASS* wurde zur Dokumentation von bildlichen Darstellungen seit den frühen 1950er Jahren von einem Mitglied der Königlichen Niederländischen Akademie der Wissenschaften (KNAW) entwickelt, 1972 publiziert und zwischen 1990 und 2001 als computergestützte Edition von der Akademie herausgegeben.[19] Seit 2006 verfügt ICONCLASS über eine Geschäftsstelle im Niederländischen Rijksbureau voor Kunsthistorische Documentatie/Netherlands Institute for Art History (RKD), die das wissenschaftliche Netzwerk koordiniert und zum Beispiel die Online-Relaunches der Web-Editionen operativ unterstützt.

- Taxonomien beziehungsweise Nomenklaturen sind auch in der Biologie und Chemie bedeutsam. Diese entwickeln sich durchaus dynamisch. Während naturkundliche Taxonomien sich teils auf Jahrhunderte alter Basis fortentwickeln, werden Bakterien beispielsweise erst seit circa vierzig Jahren einheitlich benannt. Der aktuell geltende *International Code of Nomenclature of Bacteria* (Bacteriological Code oder ICNB) wird von der International Union of Microbiological Unions herausgegeben. Vorläufer-Codes aus den 1940er und 1950er Jahren hatten keine große Verbreitung. 1980 wurde daher ein Neuanfang zur einheitlichen Benennung aller Bakterien vereinbart und umgesetzt.[20] Mit der zunehmenden Verfügbarkeit von Sequenzierungen spielen Genomdaten eine wichtige Rolle in der Taxonomie.

- Eng mit den staatlich mandatierten Normierungsorganisationen verbunden ist die Arbeit des *IEEE – Institute of Electrical and Electronics Engineers*, ein weltweiter Berufsverband, der auch Aufgaben einer Standardisierungsorganisation wahrnimmt.[21] Das IEEE ist Herausgeber wichtiger einschlägiger Standards im Fach, die teils auch über die ISO anerkannt werden. Zur Datenqualität finden sich in der IEEE-Bibliothek zahlreiche technische Standards (unter anderem zu Telekommunikation/Datenübertragung), wie auch sehr spezifische „Empfehlungen", zum Beispiel für Prüf- oder Qualitätssicherungsverfahren für Daten und Software.

Seit den 1990er Jahren gibt es also ein allmählich wachsendes Rahmenwerk, das teils fachlichen Initiativen entspringt, teils durch das wissenschaftsweite Bibliothekswesen, durch Gedächtniseinrichtungen (wissenschaftliche Archive und Sammlungen) und durch informationstechnologische Expertise vorangetrieben wird. Standardsetzungen schaffen eine Grundlage, um Daten in bereits normierter Weise zu erstellen beziehungsweise in normierte Systeme zu migrieren, oder bieten auch Übersetzungsregeln zwischen gewachsenen Wissensorganisationssystemen. Dabei sind einzelne Bereiche durch ein hohes Maß an internationalen Konsensbildungsprozessen geprägt.

---

19 http://www.iconclass.nl/home (zuletzt geprüft am: 30.08.2019).

20 Zur Historie des ICNB vgl. Website des International Committee on Systematics of Prokaryotes, http://www.the-icsp.org/. Vorläufer des heute gültigen internationalen Codes für Algen, Pilze und Pflanzen reichen hingegen ins 18. Jahrhundert zurück, https://www.iapt-taxon.org/nomen/main.php (beide zuletzt geprüft am: 30.08.2019).

21 https://standards.ieee.org/ (zuletzt geprüft am: 30.08.2019).

Allerdings fehlt es in der Forschungspraxis in vielen Bereichen an der Umsetzung. Dies liegt auch daran, dass die Rückkopplung zwischen Standardisierungsausschüssen von Infrastrukturträgern, Expertenkomitees und den wissenschaftlichen Fachgemeinschaften noch ausbaufähig ist, ebenso die Einigung auf international gültige Standards der Dokumentation.

Speziell für den Bereich des Forschungsdatenmanagements will die Research Data Alliance (RDA) dieses Defizit beheben. Das global agierende Expertennetzwerk RDA wurde 2013 „bottom-up" aus der Wissenschaft heraus gegründet und finanziert sich über verschiedene internationale Geldgeber wie die EU. Sie verfolgt die Mission, den offenen Austausch und die Wiederverwendung von Daten über Technologien, Disziplinen und Länder hinweg zu ermöglichen. Mittels international agierender Arbeitsgruppen hat die RDA unter anderem eigene technische Spezifikationen hervorgebracht,die als „Information and Communication Technology Specifications" (ICT-Standards) in der Europäischen Union (EU) anerkannt werden sollen.[22]

Mit der Konsolidierung konkurrierender Standards befasst sich auch die globale Datenplattform GEOSS.[23] Sie soll weltweit Erdsystemdaten auch aus sehr disparaten Datenquellen verschiedener Nationen integrieren. Um die erforderliche Interoperabilität zu gewährleisten, wurde ein partizipativer Prozess aufgesetzt, um gemeinsam mit den Datenproduzenten ein Register anwendbarer Standards aufzubauen.[24] Einen ähnlichen Ansatz verfolgte auch die europäische Geodateninfrastruktur INSPIRE: Ihrem Aufbau ging ein langjähriger, partizipativer Konsensbildungsprozess zu den anwendbaren Datenstandards voraus, die heute in einer europäischen Richtlinie verbindlich festgelegt sind.[25]

Außerwissenschaftlich ist es vor allem das World Wide Web, dessen Regeln und Standards Einfluss auf die wissenschaftliche Dokumentationspraxis haben.[26] Auch Suchmaschinenoptimierung kann ein Treiber sein. So hat Google im November 2018 ergänzend zu seinem bereits viel beachteten Dienst *Google Scholar* eine „Dataset Search"-Funktion eingeführt. Um hier gefunden zu werden, ist die Beigabe von Metadaten nach einem definierten Schema erforderlich.[27] Akteure wie die wissenschaftsweite Initiative DataCite reagieren bereits mit Anpassungen ihrer Dienste.[28]

Normsetz ungen und Standardisierungen folgen einem Leitbild von Steuerung, das als „juridisch" beschrieben werden kann: Qualitätssicherung soll durch Regeln oder verbindlich gesetzte Standards geleistet werden, die repräsentativ besetzte Gremien aushandeln und stabil fixieren. Nachdem

22 Siehe https://www.rd-alliance.org/recommendations-outputs/standards (zuletzt geprüft am: 30.08.2019).

23 GEOSS hat einen hohen Grad an Verbindlichkeit – weltweit treten Länder der Plattform per Regierungserklärung bei.

24 https://www.earthobservations.org/gci_sr.shtml (zuletzt geprüft am: 30.08.2019). Das GEOSS Standards Registry listet 194 Einträge (Stand Mai 2019).

25 EC/EU (2007) – Schaffung einer Geodateninfrastruktur Richtlinie 2007/2/EG.

26 Zu den umfangreichen Standards und Tools vgl. Website des World Wide Web Consortium (W3C) – https://www.w3.org/standards/ (zuletzt geprüft am: 30.08.2019).

27 Zunächst als Beta-Version, vgl. auch Morphy (2018) – What is Google Dataset Search. Zu den Anforderungen an Vokabulare für strukturierte Daten im Internet vgl. schema.org (https://schema.org/, zuletzt geprüft am: 30.08.2019)

28 Cousijn/Cruse/Fenner (2018) – Taking Discoverability.

sie beschlossen sind, sollen sie – bezogen auf die Forschung – wissenschaftsweit, mindestens aber in der Breite einer Community oder Disziplin übernommen werden. Zwar wird diese mit Normen und Standards verknüpfte Erwartung der Vielstimmigkeit, der hohen Binnendynamik und auch der Heterogenität von Forschungsprozessen nur begrenzt gerecht. Gleichwohl haben Bemühungen um Standardisierung auch in der Wissenschaft bereits wichtige globale Regelwerke geschaffen, die für den internationalen Austausch von Daten unerlässlich sind.

2.3 STANDARDISIERUNG VON DATENQUALITÄT

Die Diskussion um Fragen der Datenqualität ist in der Wirtschaft im Rahmen eines betrieblichen Qualitätsmanagements entstanden, das wiederum auf Produktqualität abzielt. Dabei stehen hinter dem Begriff „Qualitätssicherung“ (synonym verwendet: „Qualitätskontrolle“) Ansätze und Maßnahmen zur Sicherstellung zuvor festgelegter Qualitätsanforderungen. Im Zusammenhang mit der Zunahme digitalisierter Prozesse in Unternehmen (zunächst wohl vor allem hinsichtlich des Aufbaus und der Nutzung von elektronischen Datenbanken) setzte sich die Betriebswirtschaftslehre der 1990er Jahre auch mit Qualitätsfragen EDV-überwachter Produktionsketten auseinander. Dabei wird der Begriff „Data Quality“ (Datenqualität) erstmals explizit formuliert. Bis heute prägend sind die 1996 publizierten Ansätze der Datenqualitätsforscher Richard Wang und Diane Strong. Sie definieren Datenqualität bedarfsorientiert, also aus der Perspektive der Datennutzung beziehungsweise des „Datenkonsums“.[29] Wang/Strong unterscheiden vier generelle Merkmale (s. a. Tabelle 1):

- *inhaltsbezogene* Datenqualität (intrinsic): Daten besitzen eine Qualität aus sich heraus, indem sie beispielsweise fehlerfrei, glaubwürdig und objektiv sind.
- *kontextbezogene* Datenqualität (contextual): Eine Qualität der Date ergibt sich aus deren Eignung für einen kontextspezifischen Zweck, aber auch beispielsweise durch ihre Relevanz, Aktualität, ihren Mehrwert durch Verknüpfung.
- *darstellungsbezogene* Datenqualität (representational): Eine Qualität von Daten entsteht, indem diese im Hinblick auf Darstellungsformate konzise und konsistent und im Hinblick auf ihre Bedeutung interpretierbar und leicht verständlich sind.
- *zugangsbezogene* Datenqualität (accessability): Eine Qualität gewinnen Daten, sofern sie zugänglich und bearbeitbar sind und der Zugang zu ihnen sicher ist und bleibt.

Daraus entwickelt Wang 1998 einen umfassenden Ansatz für das Qualitätsmanagement von Daten, das Total Data Quality Management. Dabei werden über den Lebenszyklus von Daten hinweg eine kontinuierliche Qualitätsdefinition, Qualitätsmessung und Qualitätsanalyse durchgeführt.[30]

[29] Vgl. Wang/Strong (1996) – What Data Quality Means to Data Consumers, S. 6 ff. sowie S. 18 ff. Für einen Überblick der Forschung zur Datenqualität siehe auch Shankaranarayanan/Blake (2017) – From Content to Context. Wang (1998) – Total
[30] Data Quality Management.

Tabelle 1: Klassifikationsansatz nach Wang/Strong.[31]

Category	Dimension	Definition: the extent to which…
Intrinsic	Believability	data are accepted or regarded as true, real and credible
	Accuracy	data are corrected, reliable and certified free of error
	Objectivity	data are unbiased and impartial
	Reputation	data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	data are beneficial and provide advantages for their use
	Relevancy	data are applicable and useful for the task at hand
	Timeliness	the age of the data is appropriate for the task and hand
	Completeness	data are of sufficient depth, breadth and scope for the task at hand
	Appropriate amount of data	the quantity or volume of available data is appropriate
Representational	Interpretability	data are in appropriate language and unit and the data definitions are clear
	Ease of understanding	data are clear without ambiguity and easily comprehended
	Representational consistency	data are always presented in the same format and are compatible with the previous data
	Concise representation	data are compactly represented without being overwhelming
Accessibility	Accessibility	data are available or easily and quickly retrieved
	Access security	access to data can be restricted and hence kept secure

Die 2009 erstmals veröffentlichte ISO-Norm 8000 „Datenqualität und Stammdatenqualität" wurde in Anlehnung an die bekannte ISO 9001 zum Qualitätsmanagement entwickelt und spezifiziert vier Prinzipien für die Datenqualität (s. Tabelle 2). Vorläufer-Standards stammen aus dem Bereich Kreditkartenabrechnung/E-Commerce.[32] Die ISO 8000 stellt vor allem auf drei Qualitätsmerkmale ab: *Provenance, Accuracy* und *Completeness*.

Tabelle 2: Definition von Datenqualität in der ISO 8000.

Principles of data quality The following principles of data quality underlie ISO 8000:	
a.	Data quality involves data being fit for purpose; i. e., the decision it is used in.
b.	Data quality involves having the right data, in the right place, at the right time.
c.	Data quality involves meeting agreed customer data requirements.
d.	Data quality involves preventing the recurrence of data defects by improving processes to prevent repetition and eliminate waste.

[31] Wang/Strong (1996) – What Data Quality Means to Data Consumers zitiert nach Batini/Scannapieca (2006) – Data Quality, S. 38.

[32] Zur Historie: Electronic Commerce Code Management Association (ECCMA), nach eigenen Angaben „project leader for ISO 8000", https://eccma.org/about_eccma/ (zuletzt geprüft am: 30.08.2019).

Eine weitere ISO-Norm speziell zur Datenqualität findet sich in der sogenannten SquaRE-Serie zur Softwareentwicklung.[33] Die Serie beinhaltet ein Datenqualitätsmodell (ISO/IEC 25012) und eine Norm „Measurement of data quality" (ISO/IEC 25024).[34] Die dort spezifizierten 15 Qualitätsmerkmale weisen mit den in der Matrix von Wang/Strong aufgeführten Kategorien von Datenqualität einige Übereinstimmungen auf (s. Tabelle 3). Die Norm selbst scheint jedoch zumindest im deutschsprachigen Raum nicht sehr verbreitet zu sein; zumindest konnten keine Zertifizierungen nach dieser Norm ermittelt werden.

Tabelle 3: Qualitätsmerkmale nach ISO/IEC 25012 (SquaRE).[35]

Characteristics	Data Quantity	
	Inherent	System Dependent
Accuracy	x	
Completeness	x	
Consistency	x	
Credibility	x	
Currentness	x	
Accessibility	x	x
Compliance	x	x
Confidentiality	x	x
Efficiency	x	x
Precision	x	x
Traceability		x
Understandability		x
Availability		x
Portability		x
Recoverability		x

In Diskursen über die im Wissenschaftssystem beziehungsweise in Forschungsprozessen geforderte Qualität von Daten finden sich insgesamt kaum Referenzen auf die Datenqualitätskonzepte von Wang/Strong oder auch auf die oben genannten ISO-Normen. Man wird sagen können, dass hier Qualitätsmaße nicht nur von der Wissenschaft unabhängig, sondern zunächst einmal auch abseits ihrer Bedarfe formuliert worden sind. Entsprechend gering fiel die Resonanz aus, die diese Normierung von Datenqualität abseits der Wirtschaftsinformatik auslöste. Die Tatsache, dass wissenschaftliche Aufbereitungen oftmals nicht abgeschlossen, sondern weiteren Nutzungen und Transformationen unterliegen, erschwert die Anwendung insbesondere solcher Standards, die auf die Optimierung einer bestimmten „Produktqualität" abzielen.

[33] Systems and software Quality Requirements and Evaluation (SQuaRE).
[34] Für Einzelnachweis vgl. Verzeichnis der Standards und Normen.
[35] Rafique et al. (2012) – Information Quality Evaluation Framework, S. 570.

Am ehesten finden sich ähnlich abgeleitete Ansätze in den Ingenieurwissenschaften oder der Qualitätssicherung von Kohortenstudien in der Medizin.[36] Bezüge zum Datenqualitätsdiskurs der 1990er Jahre sind auch in der Genese der FAIR-Prinzipien erkennbar (vgl. 2.8).

Im Umfeld von datenintensiv arbeitenden Wirtschaftsunternehmen haben sich zudem Ansätze für eine explizite *Data Governance* entwickelt. Datenqualität soll durch die Festlegung von Verantwortlichkeiten und formalisierten Prozessen für den Umgang mit Daten gesteuert werden. Solche Konzepte finden sich früh zum Beispiel in der Finanz- und Kreditwirtschaft, die Daten(-produkte) vermarktet.[37] In der Wissenschaft haben sich formale Steuerungsverfahren für das Datenmanagement am ehesten im Rahmen großer Längsschnittstudien herausgebildet, durch die unter anderem umfangreiche Datenschutzauflagen erfüllt werden.[38] Eine andere Spielart von *Data Governance* findet sich unter dem Schlagwort *Good Clinical Data Management* in der medizinischen Forschung.

2.4 VALIDIERUNG UND ZERTIFIZIERUNG

Zertifizierungsverfahren richten sich auf die organisatorische oder institutionelle Umsetzung von Qualitätsnormen beziehungsweise Standards. Sie schreiben aufseiten derer, die sich für deren Einhaltung zertifizieren lassen, Verantwortung fest. Damit steuern sie Qualität nicht durch bloße Normierung, sondern durch einen positiven Verstärker, indem sie Akteure durch einen Anreiz – das Zertifikat als Gütesiegel – zur Qualitätssicherung verpflichten. Aufseiten derer, die einen Dienst oder ein Produkt nutzen wollen, stellt das Gütesiegel Vertrauen her oder bietet in einem Umfeld konkurrierender Angebote eine Orientierung.

Auch die Vergabe von Gütesiegeln ist ein Instrument, das nicht in der Wissenschaft, sondern im öffentlichen Bereich und im Wirtschaftsleben entstanden ist. Das britische Parlament beschloss 1887, dass auf allen Produkten das Herkunftsland angegeben sein müsse, als Schutz vor vermeintlich billiger und minderwertiger Importware. Die zunächst in diesem Zusammenhang zur Abgrenzung benutzte Bezeichnung „*Made in Germany*“ entwickelte sich über die Zeit dann zu einem faktischen Gütezeichen. In Deutschland war die Gründung des Reichs-Ausschusses für Lieferbedingungen (RAL) am 23. April 1925 von großer Bedeutung. 1954 wurden die „RAL-Grundsätze für Gütezeichen“ veröffentlicht, das heißt Regeln für die Gütesicherung festgelegt. 1985 wurden die „Grundsätze für Gütezeichen“ durch das Bundesministerium für Wirtschaft im Bundesanzeiger veröffentlicht.[39]

---

[36] Nonnemacher et al. (2014) – Datenqualität in der medizinischen Forschung.

[37] Für einen Überblick zur Data-Governance-Forschung vgl. Al-Ruithe et al. (2018) – Data Governance and Cloud Data Governance.

[38] Vgl. NAKO (2015) – Datenschutz- und IT-Sicherheitskonzept. Ähnliche Beispiele lassen sich auch in anderen sozialwissenschaftlichen Langzeitstudien finden.

[39] https://www.ral-guetezeichen.de/ueber-uns/ral-guetezeichen-historie/ (zuletzt geprüft am: 30.08.2019).

Aber auch im Rahmen der großen Standardisierungsorganisationen (DIN, ISO usw.) werden Normen zur Konformitätsbewertung entwickelt. Tabelle 4 zeigt an einem Beispiel die Bezüge zwischen der Spezifizierung einer bestimmten Produkt- oder Prozessqualität und dem Schritt der Konformitätsprüfung (Normierung des Zertifizierungsverfahrens und Normierung der Anforderungen an die zertifizierende Stelle).

Tabelle 4: Zusammenhängende ISO-Normen für Standardisierung und Zertifizierung am Beispiel des OAIS-Referenzmodells.

Standard Nr.	Bezeichnung	Art
ISO 14721:2012	Open archival information system (OAIS) -- Reference model	Reference Model
ISO 16363:2012	Audit and certification of trustworthy digital repositories	Audit and certification standard
ISO 16919:2014	Requirements for bodies providing audit and certification of candidate trustworthy digital repositories	Product and company certification

Quelle: eigene Darstellung, Bezeichnungen sind der ISO-Website entnommen. Vgl. auch das Verzeichnis zu Standards und Normen im Anhang.

Für Datenarchive hat sich die Bezeichnung „vertrauenswürdige Repositorien" (trustworthy repositories) zu einer Art Gütesiegel entwickelt, die Zertifizierung nach dem in Tabelle 4 dargestellten OAIS-Referenzmodell ist dabei so etwas wie ein Goldstandard. Aktuell existieren insgesamt drei von Aufwand und Tiefe her gestufte Konformitätsprüfungen, die sich auf den OAIS-Standard beziehen. Die Verfahren unterscheiden sich mit Blick auf die angewandten Prüfkataloge und Anforderungen:

- Die *Trustworthy Repositories Audit & Certification* (TRAC) wurde ab 2003 von einer gemeinsamen Task Force der US-amerikanischen Research Libraries Group und der National Archives and Records Administration (NARA) entwickelt. Der Prüfkatalog enthält umfassende Prüfungs- und Zertifizierungskriterien für OAIS-kompatible Repositorien, von der Organisation, der Personalausstattung und Finanzierung bis hin zu standardisierten Abläufen der Datenarchivierung und des Managements von Verträgen und Lizenzen. Die Spezifikation wurde später in einen ISO-Standard überführt (s. a. Tabelle 4).

- In Deutschland entwickelt wurde das nestor-Siegel für *vertrauenswürdige digitale Langzeitarchive* (ab 2011).[40] Aufgeschlüsselt werden Anforderungen, die eng am Datenlebenszyklus orientiert sind. So muss beispielsweise differenziert dargelegt werden, wie das Repositorium Transfer-, Archivierungs- und Nutzungspakete eines Informationsobjektes handhabt. Das Siegel kann auch international vergeben werden. Der Prozess zum Erwerb des nestor-Siegels ist offenbar aufwendiger als derjenige für das jüngere internationale Pendant Core Trust Seal (siehe unten). Dies ist möglicherweise ein Erklärungsansatz für

[40] Nestor (2019) – Erläuterungen zum nestor-Siegel.

die stark unterschiedliche Zahl der in beiden Verfahren zertifizierten Repositorien (vgl. Tabelle 5). Die „Kriterien für vertrauenswürdige Langzeitarchive“ des nestor-Siegels sind als DIN-Norm anerkannt.

- Das jüngste Zertifizierungsverfahren wurde 2016 unter dem Dach der Research Data Alliance entwickelt.[41] Das sogenannte Core Trust Seal (CTS) ist wie auch das nestor-Siegel ein peer-review-gestütztes Selbstevaluierungsverfahren, in dem Einrichtungen ihre Konzepte und Leitlinien für die Datenarchivierung bewerten. Der 16-Punkte-Katalog ist, wie bereits erwähnt, etwas weniger umfangreich als der des nestor-Siegels.

Für das nestor-Siegel ist darzulegen, wie das Repositorium Datenintegrität und Authentizität sicherstellt. Bei der Authentizität wird allerdings explizit darauf hingewiesen, dass die Anforderung nach dem gegenwärtigen Stand der Technik (2019) nicht vollständig erfüllt werden kann. Als Begründung wird eine Limitierung bei Anforderung K13 angegeben. Dort heißt es wörtlich: „Mit Hilfe von Kriterium 13 ‚Signifikante Eigenschaften‘ soll geprüft werden, ob die Definition und Beschreibung signifikanter Eigenschaften von Informationsobjekten in der Systemarchitektur, im Datenmodell und den Workflows hinreichend berücksichtigt wurde. […] Zurzeit kann eine vollständige Erfüllung dieses Kriteriums nicht erwartet werden, da das Verständnis und die Umsetzung des Konzepts der signifikanten Eigenschaften in der Community noch reifen müssen.“[42] Unterstützung kann der „nestor-Leitfaden Bestandserhaltung“ bieten.[43]

In Deutschland hat seit einigen Jahren neben den oben genannten Siegeln die Akkreditierung sozial- und wirtschaftswissenschaftlicher Datenzentren durch den RatSWD[44] an Bedeutung gewonnen. Dem Mandat der Initiative entsprechend stellt das Verfahren vor allem auf die Zugänglichkeit der Ressourcen ab, da es sich oftmals um sensible Daten handelt, die besonderem Schutz unterliegen. Der RatSWD hat hierfür Zugangsmodelle und Standards entwickelt.[45] Ursprünglich für die Erschließung von Daten aus Behörden und öffentlichen Stellen gedacht, schließen sich auch immer mehr wissenschaftliche Datenzentren dem Verfahren an. Mit der Akkreditierung verbunden ist die Aufnahme in das Netzwerk der Forschungsdatenzentren beim RatSWD, das unter anderem Erfahrungsaustausch und die Weiterentwicklung gemeinsamer Standards organisiert.

---

41 Urheber ist eine gemeinsame Arbeitsgruppe des World Data System und des Data Seal of Approval, die 2016 unter dem Dach der Research Data Alliance die Kriterien des Data Seal of Approval und das Zertifikat für Mitglieder des ICSU World Data System zusammengeführt haben. Für Erhalt und Weiterentwicklung des Siegels wurde jedoch eine eigene Initiative gegründet.

42 Nestor (2019) – Erläuterungen zum nestor-Siegel, S. 25.

43 Nestor (2012) – Leitfaden zur digitalen Bestandserhaltung.

44 Der Rat für Sozial- und Wirtschaftsdaten ist ein Gremium von in der Regel quantitativ-empirisch arbeitenden Wissenschaftlern in den Sozial- und Wirtschaftswissenschaften sowie Vertretern öffentlicher Einrichtungen, das an der Verbesserung der Datennutzung und des Datenzugangs für die empirische Forschung arbeitet – unter anderem durch die Einrichtung von Forschungsdatenzentren. Der RatSWD wurde 2004 vom Bundesministerium für Bildung und Forschung eingerichtet; www.ratswd.de.

45 RatSWD (2017) – Qualitätssicherung der FDZ.

Die Zahl der in vergleichbarer Weise zertifizierten oder akkreditierten Dateninfrastrukturen ist weltweit insgesamt bislang überschaubar: Von über 2300 registrierten wissenschaftlichen Datenrepositorien in der internationalen Datenbank re3data.org ist nur ein Bruchteil zertifiziert (s. Tabelle 5).

Verbindliche Anreize für die Zertifizierung von Diensten fehlen bislang. Communities wie die der kollaborativ betriebenen europäischen Forschungsinfrastruktur CLARIN[46] (s. Tabelle 5) verlangen von den angeschlossenen sogenannten Servicezentren ein Zertifikat und fördern damit – ähnlich wie der RatSWD – aktiv eine Standardisierung bestimmter Qualitäten. Die Wirksamkeit ist jedoch auf einen fachlich einschlägigen Kreis beschränkt. Die Einführung einer Zertifizierung/Akkreditierung für FAIR Services in der European Open Science Cloud (EOSC) dürfte eine vergleichsweise breitere Wirkung entfalten. Die Zertifizierung soll auf Konformität mit den FAIR-Prinzipien (vgl. 2.8) abstellen. Die zur Vorbereitung eingesetzte EU Expert Group on FAIR Data hat in ihrem Abschlussbericht empfohlen, eine Zertifizierung von FAIR Services für die EOSC auf der Grundlage des Core Trust Seal aufzubauen (und zum Erwerb des CTS aufzufordern). Ein erster Prototyp für die Zertifizierung von FAIR Services soll nach dem Willen des EOSC Executive Boards bis 2020 vorliegen.

Dennoch ist zu erwarten, dass eine gewisse Vielfalt an Akkreditierungs- beziehungsweise Zertifizierungsverfahren erhalten bleibt und Repositorien gegebenenfalls mehrere Zertifikate zugleich erwerben. Die Wahl des jeweiligen Verfahrens hängt einerseits mit den damit verbundenen spezifischen Anforderungen beziehungsweise Bedarfen der jeweiligen Community als auch mit dem damit einhergehenden Aufwand zusammen. Die niedrigschwelligen, primär auf Selbstevaluation beruhenden Verfahren wie CTS sind gerade für kleinere Dienste(-betreiber) besser leistbar.

[46] CLARIN – European Research Infrastructure for Language Resources and Technology, zum Assessment der Servicezentren vgl. https://www.clarin.eu/content/clarin-centres (zuletzt geprüft am: 30.08.2019).

Tabelle 5: Zertifikate von wissenschaftlichen Datenrepositorien, geordnet nach Häufigkeit.
Im Jahresvergleich 2018/2019 ist insbesondere ein Anstieg für das neu eingeführte Core Trust Seal zu verzeichnen. Das parallele Auslaufen des World Data System Certificate und des Data Seal of Approval, die nicht mehr vergeben werden, zeigt sich ebenfalls in den Zahlen.

	Name	Zahl zertifizierter Datenrepositorien[47]		Anbieter
		Stand 28.6.2018	30.7.2019	
1.	World Data System Certificate Vergabe bis 2017, jetzt Core Trust Seal (s. Nr. 4)	69	55	ICSU World Data System (international)
2.	Data Seal of Approval (DSA) Vergabe bis 2017, jetzt Core Trust Seal (s. Nr. 4)	48	31	DANS (NL) bzw. Data Seal of Approval Board & General Assembly (international)
3.	RatSWD Akkreditierung	32	32	Rat für Sozial- und Wirtschaftsdaten (DE)
4.	Core Trust Seal (CTS) Seit 2017	28	62	Core Trust Seal Board, Zusammenlegung von WDS und DSA unter dem Dach der Research Data Alliance (international)
5.	CLARIN Certificate	24	27	CLARIN ERIC (EU), eine Forschungsinfrastruktur im ESFRI-Programm
6.	Nestor-Siegel DIN 31644	5	1	Nestor Kompetenznetzwerk Langzeitarchivierung (DE)
7.	DINI Zertifikat „Open-Access-Repositorien und Publikationsdienste“	5	6	Deutsche Initiative Netzwerkinformation (DE)
8.	Trustworthy Repositories Audit & Certification (TRAC) ISO 16363	1	1	Consultative Committee for Space Data Systems (Ursprung), aktuell: ISO/TC 20/SC 13 Space data and information transfer systems (technical committee)

Quelle: eigene Darstellung basierend auf einer eigenen Auswertung der Datenbank unter re3data.org.

Die oben geschilderten Konformitätsbewertungen sollen primär Vertrauen in die *Prozessqualität* der Erzeugung, Verarbeitung und Speicherung von Daten schaffen. Zum Erwerb der jeweiligen Zertifikate muss ein Betreiber unter anderem darlegen, welche Maßnahmen zur Sicherung von Integrität und Authentizität der Daten sowie zum langfristigen Erhalt der Interpretierbarkeit der Daten ergriffen werden; welche Metadatenstandards eingesetzt werden (differenziert nach beschreibenden, strukturellen und technischen Metadaten) und in welchem Umfang Daten durch das Datenarchiv kuratiert werden.[48] Damit besteht ein klarer Bezug zur Qualität der Inhaltsdaten und der Metadaten, auch wenn Datenqualität nicht direkt Gegenstand der Zertifizierung ist.

[47] Datenrepositorien können mehrere Siegel haben, daher ergibt die Summe der einzelnen Zeilen nicht die Gesamtzahl zertifizierter Repositorien (könnte vom Betreiber ausgewertet werden, Anfrage bei Bedarf möglich).

[48] Detaillierte Informationen zu den Kriterien siehe nächster Abschnitt.

Konformitätsbewertungen vielerlei Art existieren jedoch auch für die Produktqualität oder intrinsische Qualität von Daten, wobei hier aus Nutzersicht das Format „Gütesiegel“ weniger eine Rolle zu spielen scheint. Von Interesse sind vielmehr automatisierte Werkzeuge, die im Einzelfall eine niedrigschwellige automatisierte (Selbst-)Überprüfung ermöglichen.

Ein Beispiel ist der INSPIRE-Validator der europäischen Infrastruktur für Geodaten.[49] Dieser ermöglicht Nutzern beziehungsweise Datenerzeugern, eine automatisierte Prüfung ihrer Daten auf Konformität mit den gemeinsam festgelegten technischen Standards für Geodaten der europäischen Mitgliedstaaten (geregelt in der INSPIRE-Direktive).[50] Automatische oder semiautomatische Werkzeuge zur Prüfung bestimmter Qualitätsmerkmale kommen zudem in verschiedenen Softwaresystemen zum Einsatz, die für das Datenmanagement oder für Analysen genutzt werden. Beispiele sind logische Prüfungen zur Einhaltung bestimmter Werte-Intervalle in Tabellen oder die sogenannten „Missing-Data-Techniken“. Darunter versteht man zum Beispiel das automatische Auffüllen fehlender Werte mit dem statistisch häufigsten Wert, um Verzerrungen vorzubeugen, die durch fehlende Werte entstehen würden. Ein solches „Auffüllen“ kann aus wissenschaftlicher Sicht allerdings auch zu „unerwünschten Nebenwirkungen“ führen, die die Aussagekraft einer Auswertung beeinträchtigen.[51] Für die Praxis der wissenschaftlichen Informationsbereitstellung ebenfalls relevant sind technische Tools zur Konformitätsbewertung im Internet. Beispiele sind hier unter anderem die Optimierung für die Google Dataset Search[52] oder die automatische Prüfung von Webseiten auf Konformität mit W3C-Standards.[53]

2.5 DATENPOLICIES UND DATENMANAGEMENTPLÄNE

Die Setzung von Normen und die Definition von Standards erreichen nicht ohne Weiteres die konkrete Ebene der wissenschaftlichen Praxis. Hier greifen vielmehr Leitlinien oder Policies, die einzelne wissenschaftliche Projekte, Einrichtungen oder Organisationen als Maßgaben für ihr eigenes Forschungsdatenmanagement beziehungsweise damit zusammenhängende Datenmanagementpläne formulieren. Auch für solche Selbstverpflichtungen und Leitlinien liegen die Anfänge in den 1990er Jahren. Ein frühes Beispiel sind die „Bermuda Principles“ für die DNA-Sequenzierung aus dem Jahre 1996. Hier einigten sich die Beteiligten des Humanen Genomprojekts darauf, dass genomische Sequenzinformationen, die von den großen Zentren generiert werden, innerhalb von 24 Stunden nach der Generierung frei verfügbar und öffentlich zugänglich sein sollten. Diese Vereinbarung trug der Open-Access-Debatte Rechnung, zu

[49] http://inspire-sandbox.jrc.ec.europa.eu/validator/ (zuletzt geprüft am: 30.08.2019).

[50] EC/EU (2007) – Schaffung einer Geodateninfrastruktur Richtlinie 2007/2/EG.

[51] Zum Beispiel kann die Übernahme von Standardeinstellungen für das Feld „gender“ in einer Datenbank oder einem Excel-Sheet zu ungewollten Verzerrungen führen.

[52] https://developers.google.com/search/docs/guides/prototype (zuletzt geprüft am: 30.08.2019).

[53] W3C steht für World Wide Web Consortium. Zu den Standards vgl. https://www.w3.org/standards/, zum Validator siehe: https://validator.w3.org/ (zuletzt geprüft am: 30.08.2019).

deren Prinzipien sich die Wissenschaft Anfang der 2000er Jahre in unterschiedlichen Erklärungen verpflichtet hatte.[54] Die 1998 in einer ersten Fassung verabschiedeten Regeln für „gute wissenschaftliche Praxis" der DFG gaben durch die Empfehlung Nr. 7 zur „Sicherung und Aufbewahrung von Forschungsprimärdaten" einen weiteren Impuls in der Debatte zum Umgang mit Daten. Der Fokus verlagerte sich in den kommenden Jahren von der Formulierung grundsätzlicher Prinzipien zu Open Access und Fragen der langfristigen Sicherung hin zu detaillierteren Konzepten für ein Forschungsdatenmanagement.[55] So verabschiedeten zum Beispiel in Großbritannien seit 2009 zahlreiche Universitäten eine Forschungsdaten-Policy oder Leitlinie für das Forschungsdatenmanagement.[56] Im Juli 2016 haben die Universitäten im Verbund mit den Higher Education Funding Councils (HEFCEs), den Research Councils UK und dem Wellcome Trust ein gemeinsames *Concordat on Open Research Data* formuliert.[57]

Ab etwa 2014 haben vergleichbare Anstrengungen auch bei Förderern, Universitäten und Forschungseinrichtungen in Deutschland eingesetzt.[58] Die 2015 verabschiedete „Leitlinie zum Umgang mit Forschungsdaten" der DFG spezifiziert unter anderem, dass der Umgang mit den zu erhebenden Forschungsdaten möglichst schon in der Projektplanung bedacht werden soll,und dass sie möglichst zeitnah zu veröffentlichen sind.[59] Die DFG bezieht sich in den Leitlinien auf Grundsätze zum Umgang mit Forschungsdaten, die in der Allianz der Wissenschaftsorganisationen 2010 verabschiedet wurden.[60] In der Folge sind zum Beispiel im Rahmen der DFG-Fachkollegien zahlreiche Handreichungen für Forschungsdatenmanagement der jeweiligen Fachgemeinschaften ausgearbeitet worden.

Dass eine Zunahme der Verabschiedung der Forschungsdaten-Policies in den vergangenen zehn Jahren zu beobachten ist, hängt zum einen damit zusammen, dass die nationalen Forschungsförderer relativ früh – und zwar durchaus parallel – das Open-Access-/Open-Science-Paradigma aufgegriffen haben[61] und die Vorlage einer Forschungsdaten-Policy oder von Datenmanagementplänen mehr und mehr verpflichtend machten. So sollen Projekte, die im Rahmen von Horizont 2020 aus europäischen Mitteln gefördert werden, die erhobenen

54 Wie beispielsweise der Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen; Allianz der Wissenschaftsorganisationen (2003) – Berliner Erklärung.

55 Zu den Entwicklungsschritten vgl. https://www.forschungsdaten.info/themen/aufbereiten-und-veroeffentlichen/leitlinien-und-policys/ (zuletzt geprüft am: 30.08.2019).

56 http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies (zuletzt geprüft am: 30.08.2019), vgl. auch die Darstellung zur Entstehung von Forschungsdatenpolicies in UK und anderen Ländern in RfII (2017) – Fachbericht Länderanalysen.

57 Zur ausführlichen Darstellung vgl. RfII (2017) – Fachbericht Länderanalysen, S. 59–67 sowie HEFCE et al. (2016) – Concordat on Open Research Data.

58 Vgl. hierzu auch https://www.forschungsdaten.info/praxis-kompakt/fdm-in-den-bundeslaendern/ (zuletzt geprüft am: 30.08.2019).

59 DFG (2015) – Leitlinien Forschungsdaten.

60 Allianz der Wissenschaftsorganisationen (2010) – Grundsätze Forschungsdaten. Die Grundsätze stellen auf qualitätsgesicherte Forschungsdaten, Zugänglichkeit und Sicherung, Verwendung von Standards sowie Aufbau von Infrastrukturen ab.

61 Zu Open Science als Treiber für die Entstehung nationaler und subnationaler Regelungen zu Forschungsdaten und Dateninfrastrukturen vgl. RfII (2017) – Fachbericht Länderanalysen, S. 9 f.

Forschungsdaten für die Öffentlichkeit zugänglich machen.[62] Und auch einige Förderprogramme der DFG oder des BMBF fordern Datenmanagementpläne oder die Offenlegung der erhobenen Forschungsdaten explizit ein.[63]

Freilich fällt bezüglich der in den vergangenen zehn Jahren veröffentlichten Leitlinien eine Vielfalt in der Regelungstiefe, der Begriffsklärung und hinsichtlich der Verbindlichkeit auf. Vertreter des Open-Access-Gedankens bemängeln zudem, dass Forschungsdaten-Leitlinien nur die Offenlegung der Daten betreffen, nicht aber einen fairen Umgang mit den Daten nach der Veröffentlichung. Die Angst vor einer unredlichen Nutzung durch Konkurrenten („data parasitism") hindere viele Forschende daran, ihre Daten offenzulegen.[64]

In jüngster Zeit werden auch Ratgeber zur Erstellung von FDM-Leitlinien und für das Aufsetzen von Datenmanagementplänen publiziert. Das britische Digital Curation Centre (DCC) hat zum Beispiel 2014 einen 5-Schritte-Ratgeber zur Policy-Gestaltung verabschiedet.[65] Zugleich wurden konkrete Ratgeber zur Erstellung von Datenmanagementplänen und auch Forschungsdaten-Policies verabschiedet.[66] Das Committee on Data of the International Science Council (CODATA) verabschiedete 2014 „*Current Best Practice for Research Data Management Policies*". 2017 wurde das LEARN Toolkit of Best Practice for Research Data Management veröffentlicht.[67] Das vom BMBF geförderte Projekt FDMentor entwickelte zwischen 2017 und 2019 ein Forschungsdaten-Policy-Kit.[68] Im Rahmen eines DFG-geförderten Projekts wurde ab 2015 die freie Software RDMO – Research Data Management Organiser entwickelt.[69] Die Dynamik der Entwicklung lässt erkennen, welche Bedeutung der Erstellung einer Forschungsdaten-Policy und oftmals als Teil davon den Datenmanagementplänen beigemessen wird. Für die Umsetzung an Institutionen und in einzelnen Forschungsprojekten werden zunehmend umfangreiche Handreichungen, Checklisten und auch digitale Tools bereitgestellt, auch werden von den Institutionen mehr und mehr Anlaufstellen für FDM-Beratung eingerichtet.

62 EC (2017) – H2020 Programme MGA.

63 Siehe hierzu DFG (2015) – Leitlinien Forschungsdaten sowie BMBF (2012) – Bekanntmachung „Sprachliche Bildung und Mehrsprachigkeit".

64 Zum Beispiel Amann et al. (2019) – Toward Unrestricted Use of Data.

65 DCC (2014) – Developing a Research Data Policy.

66 Vgl. DCC – Digital Curation Centre, http://www.dcc.ac.uk/resources/data-management-plans (zuletzt geprüft am: 30.08.2019).

67 LEARN (2017) – Toolkit of Best Practice for RDM. LEARN steht für LEaders Activating Research Networks, das Projekt geht zurück auf die 2013 verabschiedete „Roadmap for Research Data" der League of Research Universities (LERU).

68 http://www.forschungsdaten.org/index.php/FDMentor (zuletzt geprüft am: 30.08.2019).

69 https://rdmorganiser.github.io/ (zuletzt geprüft am: 30.08.2019).

Die zunehmende Implementierung von Regeln für das Forschungsdatenmanagement umfasst somit mehrere Faktoren und Ebenen:

1. Die wissenschaftspolitische Forderung, dass wissenschaftliche Einrichtungen eine Forschungsdaten-Policy verabschieden.
2. Die Ebene der Forschungsförderer, die bei der Antragstellung die Vorlage eines Datenmanagementplans fordern.
3. Die auch durch Ausschreibungen motivierte Entwicklung von Tools zur Erstellung von Datenmanagementplänen.
4. Die Ebene des umsetzenden Forschers, der in seinem Projekt entlang der Vorgaben seiner Forschungseinrichtung und unter Anwendung der verschiedenen Tools einen Datenmanagementplan erstellen soll.

FDM-Leitlinien und Datenmanagementpläne stellen vertragsartige Instrumente dar, deren erhoffte Steuerungswirkung, was Qualität angeht, auf der Idee der Selbstverpflichtung aufbaut. Im besten Fall sorgen sie dafür, dass Forschende Verantwortung übernehmen und die (gerade in digitalen Forschungsprozessen unter Umständen aufwendige und ressourcenintensive) Auseinandersetzung mit Datenqualitätsfragen gut bewältigen können. Idealerweise „leiten" Leitlinien jeden Einzelnen zu vorausschauender Planung und gezieltem Ressourceneinsatz an: Die konkrete Umsetzung der Datenmanagementpläne und anderer in Forschungsdaten-Policies festgeschriebener Regeln – zum Beispiel die Zugänglichmachung beziehungsweise Veröffentlichung der Daten – liegt beim jeweiligen Forschungsprojekt, mit mehr oder weniger umfassender Betreuung durch qualifiziertes Personal. Die Kehrseite dieser Art von Qualitätssteuerung ist allerdings, dass Normierungsinstrumente für eine „Feinsteuerung" zum Einsatz kommen, die den Forschungsprozess engmaschig einfassen und auch binden. Die Berichtspflichten zur Umsetzung der Anforderungen werden demzufolge lose gehandhabt, nicht zuletzt, weil seitens der Förderer anerkannt wird, dass es oft noch an geeigneten Forschungsumgebungen fehlt.

2.6 OPTIMIERUNG VON PROZESSKETTEN: DER DATENLEBENSZYKLUS

Auf die Einsicht, dass Forschung auf Forschung aufbaut, dass Wissenschaft also auf selbstgeschaffenes Wissen permanent zurückgreift sowie dieses dadurch auch erneuert, haben die Wissenschaftstheorie und die Informationswissenschaft mit Kreislaufmodellen reagiert.[70] Um den Umgang mit Forschungsdaten – auch unter Qualitätsaspekten – darzustellen, wurden ab Mitte der 2000er Jahre Modelle sogenannter Datenlebenszyklen entwickelt.[71] Die vitalistische Rede vom „Lebenszyklus" (statt nur „Zyklus") orientiert sich an einem betriebswirtschaftlichen Leitbild, das einerseits eine befristete Produktlebenszeit (mit festem Ende), andererseits aber

[70] Diese können an Hermeneutik („Sinn"), Ökologie („Stoff"), Kybernetik („Information") oder Entscheidungstheorie („Wissen") anknüpfen, und damit an ganz unterschiedliche Leitvorstellungen.

[71] Chang (2012) – Data Life Cycle.

auch Recycling-Optionen (Neuanfang/Wiederbelebung) vorsieht.[72] Datenlebenszyklen zu betrachten beziehungsweise zu modellieren, wird auch als Strukturierungshilfe für die Planung des Datenmanagements empfohlen.[73]

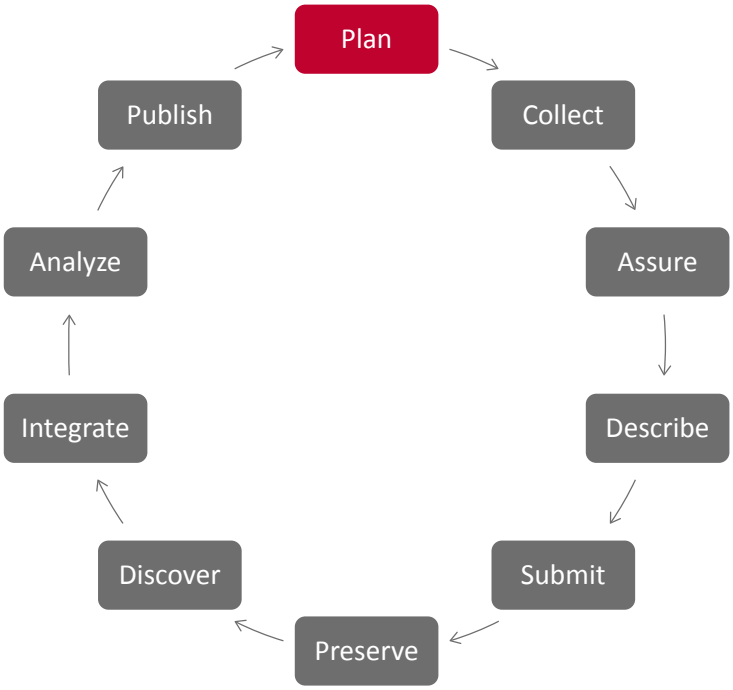

Abbildung 1: Modell eines Datenlebenszyklus.
Quelle: German Federation for Biological Data (2019).[74]

Schon 2012 hat eine einschlägige Überblicksseite zahlreiche Datenlebenszyklus-Modelle vorgestellt, die sich hinsichtlich Detaillierungsgrad und Schwerpunktsetzung (unter anderem Fachspezifik, aber auch operativer Zielstellung) unterscheiden.[75] Beim Curation-Lifecycle-Model des britischen Digital Curation Centre (DCC) von 2007/08 liegt der Schwerpunkt auf der langfristigen Archivierung,[76] wohingegen der Datenlebenszyklus der internationalen Data Documentation Initiative (DDI), der 2008 publiziert worden ist, einen besonderen Fokus auf die Datensammlung und-auswertung legt.[77] Das oben dargestellte Modell der German Federation for Biological Data (GFBio) stellt auf Konstellationen ab, bei denen selbst erhobene Daten im weiteren Forschungsverlauf mit Daten aus anderen Quellen integriert werden (vgl. Abbildung 1).

[72] Das absehbare Ende einer Nutzung (eine „Entwertung", der Verbrauch) von Daten wird somit aus der Wirtschaft auch auf die Welt der Forschungsprozesse und das Forschungsdatenmanagement übertragen. Der „Projektsicht" auf Forscherhandeln kommt dies entgegen. Wissenschaft insgesamt kann Daten indes nicht allein als Verbrauchsgüter, sondern muss sie – jedenfalls die relevanten unter ihnen – als potenziell von „ewigem" Interesse betrachten.

[73] https://www.forschungsdaten.info/themen/planen-und-strukturieren/datenlebenszyklus/ (zuletzt geprüft am: 30.08.2019).

[74] Vgl. https://www.gfbio.org/de/training/materials/data-lifecycle (zuletzt geprüft am: 30.08.2019).

[75] Siehe Working Group on Information Systems and Services (2012) – Data Life Cycle Models sowie Chang (2012) – Data Life Cycle.

[76] http://www.dcc.ac.uk/resources/curation-lifecycle-model (zuletzt geprüft am: 30.08.2019).

[77] http://www.ddialliance.org/Specification/DDI-Lifecycle/ (zuletzt geprüft am: 30.08.2019).

Nutzung beziehungsweise „Nutzungsfelder“ stehen auch im Fokus des Domänenmodells von Treloar und Harboe-Ree aus dem Jahr 2008. Unterschieden werden exklusive Nutzung im Umfeld des Forschungsprozesses selbst, das Teilen von Daten mit Forschern eines weiteren Umfeldes und der öffentliche/offene Zugang.[78] Dieses Modell entstand im Kontext der Einführung eines lokalen Datenrepositoriums an der australischen Monash University und ist erkennbar dazu gedacht, Forschenden die Optionen auf den verschiedenen Handlungsebenen zu verdeutlichen. Es belegt den infrastrukturgetriebenen Entstehungskontext von Datenlebenszyklen. Diskutiert werden anhand von Zyklusmodellen ebenso die professionellen Akteure des Forschungsdatenmanagements wie auch die Arbeitsteilung zwischen ihnen. Hier sind zum Beispiel die Definitionen von Swan und Brown aus dem Jahr 2008 („data creators, data managers, data librarians and data scientists“)[79] ebenso zu nennen wie die Definitionen der Kernkompetenzen von Donnelly im Jahr 2008.[80]

Die Steuerungsidee zur Datenqualitätsverbesserung, die sich hinter Datenlebenszyklen verbirgt, ist verfahrenstechnischer Art und trägt insbesondere dem Gedanken einer nachhaltigen (Wieder-)Nutzung von Daten Rechnung. Daten werden analog zu der – nicht zum Verkauf oder Verbrauch, sondern zum kontinuierlichen Gebrauch – bereitzuhaltenden Ressource Wissen modelliert. Die einerseits praxisorientierten, andererseits der Wissenschaft als Wissensproduktion und „Wissensarbeit“ zuträglichen Konzepte wählen weder den Weg einer (punktuellen) Normierung, noch orientieren sie sich lediglich am Produkt oder „Output“. Es zählt vielmehr – der traditionellen Rolle wissenschaftlicher Methoden ähnlich – die Qualität des Prozesses. Datenlebenszyklus-Modelle tragen überdies dazu bei, wissenschaftliche Infrastrukturen mit ihren Nutzerinnen und Nutzern in einem kontinuierlichen Interaktionsprozess zu sehen.

2.7 DAS KONZEPT „FIT FOR PURPOSE”

Älter als der Datenlebenszyklus ist die Vorstellung der die Qualität bemessenden Eignung von Daten für einen Zweck. Dieser Leitgedanke entstammt Konzepten der industriellen Qualitätskontrolle und ist über Jahrzehnte fortgeschrieben und weiterentwickelt worden.[81] So definiert das Quality Handbook des Wirtschaftsingenieurs Juran Qualität mit Blick auf Fertigungsprozesse als „fitness for use“. 1999 wurde das Konzept auf das Feld der digitalen Datenqualität übertragen: „high-quality data are data that are fit for use in their intended operational, decision-making, planning, and strategic roles.“ [82] Juran bezieht sich in seinem Werk wiederum auf die Arbeiten von Wang und Strong (vgl. 2.3).

[78] Treloar/Harboe-Ree (2008) – Data Management and the Curation Continuum.
[79] Swan/Brown (2008) – Skills, Role and Career Structure of Data Scientists.
[80] Donnelly (2008) – RDMF2. Core Skills Diagram.
[81] Juran (1951) – Quality-Control Handbook (1); seit 1951 zahlreiche Neuauflagen.
[82] Redman (1999) – Second-Generation Data Quality Systems, hier S. 8 in Sektion 34.

Ab 2010 wird das Konzept in die neue Formel des „fit for purpose“ transferiert. Anstatt „fit for use“ sollte nun „fit for purpose“ als Formel verwendet werden, um die Qualität von Produkten zu definieren. Das „Produkt“ – ob es sich um Güter, Dienstleistungen oder Informationen handelt – so der Ansatz, müsse aus Sicht des Nutzers für den jeweiligen Nutzungszweck geeignet sein. Unter Nutzer sei hierbei nicht nur der erste Abnehmer des Produkts zu verstehen, sondern beispielsweise auch die späteren Käufer sowie zum Beispiel auch Lieferanten und Aufsichtsbehörden. „Fit for purpose“ sei ein Produkt folglich dann, wenn es den Bedürfnissen und Ansprüchen eines erweiterten Kundenkreises gerecht werde, geringe Mängel aufweise und für die Exzellenz des gesamten Produkt- und Geschäftszyklus einstehen könne.[83] Auch die internationale Standardisierungsorganisation ISO verwendet den Begriff: ISO „creates documents that provide requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose.“ [84]

Mit der Wendung „fitness for purpose“ wurde ein Begriff gewählt, der eine Terminologie des US-amerikanischen und britischen Rechtsraums übernimmt. So stellt das Prinzip den Verkäufer einer Ware in die Pflicht, sich bei einem Geschäft an bestimmte Standards zu halten.[85] Der Verkäufer trägt also die Verantwortung/Haftung dafür, dass der mit dem Käufer verabredete Zweck einer Ware erfüllt wird, eine Haftung, die sich über viele Jahre oder gar Jahrzehnte erstrecken kann. Der britische *Sales of Goods Act* von 1979 fordert, Güter sollten eine „satisfactory quality“ (Mängelfreiheit) vorweisen und „fit for purpose“ sein.[86]

Die hier in ihrer Verankerung im rechtlichen und auch wirtschaftlichen Kontext deutlich hervortretende Kategorie des „fit for purpose“ ist auch in vielen aktuellen Ansätzen in der Wissenschaft, zumal als pragmatische Kurzdefinition für Datenqualität zu finden.[87] So verwenden viele internationale Akteure die Formel „fit for purpose“. Sie wird dabei auf verschiedene Gegenstände bezogen: Daten, Infrastrukturen, Open Educational Resources oder die European Open Science Cloud (EOSC). Datenqualität wird auf diese Weise umfassend, aber bestimmungsoffen definiert als: Gesamtheit von Eigenschaften und Merkmalen von Daten bezüglich ihrer Eignung, einen bestimmten Zweck zu erfüllen.[88]

---

83 Juran (2010) – Attaining Superior Results Through Quality, S. 71; Juran (2010) – The Universal Methods to Manage Quality.

84 https://www.iso.org/standards.html (zuletzt geprüft am: 30.08.2019).

85 „Generally, when a buyer makes known to a seller the particular purpose for which the goods are bought, there is an implied condition that the goods are reasonably fit for that purpose (customer‘s requirements, needs, or desires)“; Loomis Bros. Corp. v. Queen, 1958 Pa. Dist. & Cnty. Dec. LEXIS 269, 4–5 (Pa. C.P. 1958).

86 „Fit for purpose means both for their everyday purpose, and also any specific purpose that you agreed with the seller (for example, if you specifically asked for a printer that would be compatible with your computer, or wall tiles that would be suitable for use in a bathroom)“; https://www.which.co.uk/consumer-rights/regulation/sale-of-goods-act (zuletzt geprüft am: 30.08.2019).

87 Anschaulich erläutert in RIN (2008) – To Share or not to Share, 49 ff., zu den Verwendern der Formel gehören das UK Data Archive, die OECD, European Science Foundation, European Commission und viele mehr.

88 Vgl. Präsentation Schmalzl im ersten Fachgespräch der AG Datenqualität. Archive unterscheiden zudem zwischen „Primärzweck“ und „Sekundärzweck“.

Attraktiv an der pauschalen Orientierung am „Zweck“ (beziehungsweise der Absicht des Verwenders) ist auch, dass der Kontext (guter) Wissenschaft zwar mitgedacht werden kann, aber nicht näher spezifiziert werden muss. Denn: Sowohl die technische Beschaffenheit der Daten („technical aspects“) als auch ihr wissenschaftlicher Wert („scholarly method and content“) machen Forschungsdaten „fit for purpose“.[89] Die Kurzformel, Datenqualität ergebe sich als „fitness for purpose“, suggeriert zum einen, das Konzept lasse sich leicht operationalisieren. Zum anderen ist es als relationales Konzept tatsächlich maximal flexibel, da die „fitness for purpose“ durch die Nutzung entsteht, also nahezu beliebig variieren kann. Zu bedenken ist zudem, dass der Begriff „fit for purpose“ zumindest im ursprünglichen Sinne eine haftungsrechtliche Dimension und einen verbrieften Anspruch des Kunden oder Nutzers impliziert.

In der Praxis wird nicht immer die Herrichtung von Daten für einen gänzlich offenen Zweck gefordert oder gelebt. Bei der Zertifizierung von Repositorien legen sowohl das internationale Core Trust Seal als auch das nestor-Siegel nahe, dass Repositorien Zielgruppen („designated communities“) definieren, für die sie die Daten vorhalten, und mit denen sie Anforderungen an die Dokumentation festlegen.[90] Der Begriff geht zurück auf das OAIS-Modell: „A designated community is characterized as having members who have an established common ground that minimizes miscommunication. Traditionally, designated communities of data users are domain literate and have some familiarity with the scientific context, data generation, or intended data use.[91] Für die Nachnutzung der Daten wird also so etwas wie eine gemeinsame Wissensbasis vorausgesetzt.

Der Nutzungsbegriff ist aber ebenfalls noch in Gebrauch. Neben der Formel „fit for use“ findet sich ein pragmatisches „informed use“: „In practice, this means that data must undergo quality review, a process whereby data and associated files are assessed, and required actions are taken, to ensure files are independently understandable for informed reuse“.[92]

2.8 DIE FAIR-PRINZIPIEN

Der Gedanke einer „reinen“ Nutzungsorientierung, die Orientierung am möglichen Gebrauch bei zugleich offen gehaltenen Zwecken also, liegt implizit oder explizit auch Programmen zugrunde, die pragmatische „Prinzipien“ für den Umgang mit Forschungsdaten fordern. Sie werden im Sinne einer Selbstverpflichtung seitens wissenschaftlicher Einrichtungen oder auch von Forschungsprojekten zur „Nachnutzbarkeit“ von Forschungsdaten adaptiert. In der Vorstellung von Nachnutzung spielt die erforderliche gemeinsame Wissensbasis oftmals eine untergeordnete Rolle, es geht vor allem um Zugänglichkeit der Daten.

89 RIN (2008) – To Share or not to Share, S. 48 ff.

90 Vgl. DSA/ICSU-WDS (2016) – Core Trustworthy Data Repositories Requirements bzw. Nestor (2019) – Erläuterungen zum nestor-Siegel.

91 Baker et al. (2015) – Scientific Knowledge Mobilization, S. 113.

92 Peer/Green et al. (2014) – Committing to Data Quality Review, S. 263.

Ein prominentes Beispiel sind die 2016 veröffentlichten FAIR-Data-Prinzipien (s. Tabelle 6). Im Gegensatz zu den Konzepten „fit for purpose“ beziehungsweise „fit for use“ handelt es sich um ein Modell, das stärker auf die prozedurale Steuerung von Qualitätsentwicklungen und auch auf die Einbindung beziehungsweise allmähliche Gewinnung von gewissen Standards setzt.

Die FAIR-Data-Initiative als Programm, das der datengetriebenen Wissenschaft „Prinzipien“ an die Hand gibt, geht auf einen mehrtägigen Workshop in den Niederlanden im Januar 2014 zurück. Die Begründer hatten sich das Ziel gesetzt, den rasch fortschreitenden Übergang hin zu datengestützter Wissenschaft durch ein komprimiertes, aber dennoch umfassendes Rahmenwerk zu begleiten. Dazu gehörte von Beginn an eine Kampagne, die Forschende, Verlage und die Politik mit einbezog.[93]

Tabelle 6: FAIR Data Principles 2016.

TO BE FINDABLE:	
F 1	(meta)data are assigned a globally unique and eternally persistent identifier.
F 2	data are described with rich metadata
F 3	(meta)data are registered or indesed in a searchable resource.
F 4	metadata specify the data identifier.
TO BE ACCESSIBLE:	
A 1	(meta)data are retrievable by their identifier using a standardized communications protocol.
A 1.1	the protocol ist open, free and universally implementable.
A 1.2	the protocol allows for an authentication and authorization procedure, where necessary.
A 2	metadata are accessible, even when the data are no longer available.
TO BE INTEROPERABLE:	
I 1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I 2	(meta)data use vocabularies that follow FAIR principles.
I 3	(meta)data include qualified references to other (meta)data.
TO BE RE-USABLE:	
R 1	meta(data) have a plurality of accurate and relevant attributes.
R 1.1	(meta)data are released with a clear and accessible data usage license.
R 1.2	(meta)data are associated with their provenance.
R 1.3	(meta)data meet domain-relevant standards.

Quelle: Force 11.[94]

[93] Lorentz Center (2014) – Jointly Designing a Data Fairport.

[94] https://www.force11.org/group/fairgroup/fairprinciples (zuletzt geprüft am: 30.08.2019).

Die Grundprinzipien Findable – Accessible – Interoperable – Re-usable erinnern auf den ersten Blick an Grundwerte oder ethische Prinzipien, die auch in anderen Feldern verschiedentlich zur Ordnung kollektiven Handelns entworfen wurden.[95] Die vier Begriffe werden ergänzt durch eine Reihe von Anforderungen an die Daten beziehungsweise Metadaten und Protokolle für den Datenabruf.

Akteure wie die Europäische Kommission und die Research Data Alliance haben die FAIR-Prinzipien schnell aufgegriffen. Für den europäischen Datenpolitiker Jean-Claude Burgelman ist „FAIR data" gar die „DNS der EOSC".[96] Zur Weiterverbreitung und operativen Umsetzung der FAIR-Prinzipien hat sich 2017 aus der Wissenschaft und infrastrukturtragenden Einrichtungen heraus die niederländisch-deutsch-französische GO-FAIR-Initiative gebildet. GO-FAIR hat sich vorgenommen, einen Beitrag zu frühen Phasen der Entwicklung der European Open Science Cloud (EOSC) und zur internationalen Bottom-up-Entwicklung eines Internet of FAIR Data and Services zu leisten.[97]

Die im FAIR-Akronym zusammengefassten Grundsätze haben insbesondere die Verbesserung von Maschine-Maschine-Interaktionen bei der Datennutzung zum Ziel.[98] Ebenso fördern sie unter dem Stichwort einer „Offenheit" für möglichst freie Nutzung die Organisation eines möglichst breiten Datenzugangs nicht nur für die Wissenschaft allein (representational data quality/accessibility).

Festzustellen ist allerdings, dass die FAIR-Programmatik primär der Intensivierung der Datennutzung, nicht aber der wissenschaftlichen Qualitätsverbesserung dient. Aus diesem Grund wird in einem 2016 veröffentlichten Grundsatzpapier von „Qualität" auch nur an zwei Stellen explizit gesprochen. Unter der Überschrift „Supporting discovery through good data management" heißt es: „The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies." An einer weiteren Stelle wird der Qualitätsgesichtspunkt allein mit Blick auf Datenpublikationen genannt: „The goal is for scholarly digital objects of all kinds to become 'first class citizens' in the scientific publication ecosystem, where the quality of the publication — and more importantly, the impact of the publication — is a

[95] So hat die sog. „Prinzipienethik" von Tom Beauchamp und James Childress anhand der vier Prinzipien „Respect for Autonomy", „Nonmaleficence", „Beneficence" und „Justice" ab 1977 mit großem Erfolg den Bereich der Biomedizin neu zu ordnen vermocht. Eine entsprechende Vorgehensweise wurde verschiedentlich auf andere Felder übertragen (zuletzt mit der Formulierung von „Prinzipien" für den Bereich des Einsatzes von Künstlicher Intelligenz); Beauchamp, Childress (1983) – Principles of Biomedical Ethics.

[96] Vgl. Burgelman (2017) – From Vision to Action, Folie 12.

[97] https://www.go-fair.org/ (zuletzt geprüft am: 30.08.2019).

[98] „Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.", Wilkinson et al. (2016) – The FAIR Guiding Principles, S. 1. Das einprägsam gefasste Modell für bessere Maschinenlesbarkeit erinnert an eine frühere Initiative von Tim Berners-Lee. Der Begründer des Internets hatte mit dem Aufkommen des „semantic web" ein sogenanntes „5-Sterne-Modell" für offene Daten initiiert, https://5stardata.info/de/ (zuletzt geprüft am: 30.08.2019).

function of its ability to be accurately and appropriately found, reused, and cited over time, by all stakeholders, both human and mechanical."[99]

In den FAIR-Prinzipien wird wissenschaftliche (Daten-)Qualität weder direkt gefordert, noch wird definiert, was Datenqualität ist. Als generelles Ziel wird lediglich die hohe Qualität von Datenpublikationen genannt: Diese sollen für Individuen und vor allem auch maschinell „nachnutzbar" sein. Nutzung beziehungsweise Nachnutzung überlagert damit im Rahmen von FAIR die mögliche (gesonderte) Frage nach der wissenschaftlichen Qualität von Daten. Tatsächlich scheint diese in der FAIR-Programmatik zeitweise sogar fallengelassen worden zu sein. Eine indirekte, aber wichtige Referenz findet sich zwar in der Originalpublikation der FAIR-Formel aus dem Jahr 2014. Unter dem Prinzip „Re-usable" wird postuliert: „(meta) data meet domain-relevant community standards" (s. Tabelle 6). In späteren Darstellungen und Papieren zu den FAIR-Prinzipien findet sich dieser Punkt hingegen nicht immer. Ohne Einhaltung solcher Datenstandards bliebe freilich der wissenschaftliche Wert FAIRer Daten unklar.

Eine Fusion des pragmatischen Prinzipien-Ansatzes sowohl mit Normierungen als auch Management und Datenlebenszyklus findet sich in den 2016 publizierten „H2020 Programme Guidelines on FAIR Data Management in Horizon 2020".[100] Dort sind existierende Konzepte insbesondere zugunsten der Einführung von obligatorischen Datenmanagementplänen eng miteinander verwoben. Auch der zunächst nicht genutzte Begriff des Datenlebenszyklus wird inzwischen punktuell verwendet.[101] In der *EOSC Implementation Roadmap* von 2018 spielt die Implementierung der FAIR-Prinzipien eine zentrale Rolle. Im europäischen Forschungsrahmenprogramm Horizont 2020 waren für die Operationalisierung der FAIR-Data-Prinzipien 2018/2019 erhebliche Mittel vorgesehen, zudem sollen Dienste in der EOSC nach den FAIR-Prinzipien akkreditiert/zertifiziert werden. Als Rahmenwerk für die breite Anwendung der FAIR-Prinzipien in der Praxis soll ein *FAIR Data Action Plan* dienen. Zu den grundlegenden Empfehlungen gehört eine Erweiterung des Begriffsverständnisses von FAIR, so heißt es wörtlich: „FAIR is not limited to its four constituent elements: it must also comprise appropriate openness, the assessability of data, long-term stewardship, and other relevant features." [102] Allerdings rät die Expertengruppe dazu, die etablierte „Marke" FAIR nicht um weitere Buchstaben zu erweitern. Von „Qualität" ist auch in dem aktuellen Entwurf (nur) zweimal die Rede, bezogen auf „quality of services", nicht aber auf Methoden, Ergebnisse oder Daten.

99 Wilkinson et al. (2016) – The FAIR Guiding Principles.

100 „Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/ or generated by a Horizon 2020 project.", EC (2018) – H2020 Programme, S. 4 (im PDF-Dokument).

101 So zur Erläuterung der Unterscheidung von „open data" und „FAIR data" auf der Homepage der GO-FAIR-Initiative, vgl. https://www.go-fair.org/faq/ask-question-difference-fair-data-open-data/ (zuletzt geprüft am: 30.08.2019).

102 Vgl. Hodson et al. (2018) – Fair Data Action Plan. Interim Recommendations, S. 3; Burgelman (2017) – From Vision to Action.

Die FAIR-Prinzipien bieten somit ein Set an sehr grundsätzlichen Kriterien, die geeignet sind, einen Prozess mit der akzeptierten Zielrichtung einer breiten Datennutzbarkeit und Datennutzung anzutreiben. Ob FAIR unter dem Gesichtspunkt von qualitativ hochwertiger oder methodengerechter Wissenschaft einen Fortschritt darstellt (oder auch nur anstrebt), lassen die einschlägigen Papiere offen. Insbesondere springt der Verzicht auf fachwissenschaftliche (Qualitäts-)Standards als Maßgröße ins Auge. Derzeit klafft somit eine wahrnehmbare Lücke zwischen den grundsätzlichen Prinzipien, den FDM-Policies, den Zertifizierungsansätzen für Datenrepositorien etc. und ihrer konkreten Operationalisierung und Verbindlichkeit in der Wissenschaft („community standards“).

3 FAZIT

Die dargelegten Initiativen, Konzepte und Entwicklungen versuchen Datenqualität zu modellieren und zu sichern. Sie stellen aber auch – wird der Qualitätsgedanke auf die Wissenschaft übertragen – Ansätze für eine Steuerung innerwissenschaftlicher Qualitätssicherung und Qualitätsverbesserung von Daten dar. Wege und Begriffe hierfür müssen gerade in der rasch voranschreitenden digitalen Welt teilweise noch ins Verhältnis gesetzt werden mit dem „gelebten" strengen Qualitätsverständnis, das wissenschaftlichen Methodendiskursen zugrunde liegt.

Die obigen Beschreibungen sollten gezeigt haben, dass im gelebten Qualitätsverständnis des heutigen wissenschaftlichen Umgangs mit Daten viele Ansätze und die mit ihnen verknüpften Steuerungsmodi ineinandergreifen, ohne dass sie über fachliche Grenzen und Domänengrenzen hinweg ohne weiteres im Sinne von *one size fits all* Lösungen einsetzbar wären:

Im weitesten Sinne juridische Ansätze wie Normsetzungen und Standardisierungen (zum Beispiel ISO-Normen und fachliche Standards), haben sich im Bereich der „Wissensspeicher", den Ingenieurwissenschaften sowie in einigen Natur- und Geisteswissenschaften als Ordnungssysteme bewährt. Sie sind aber verhältnismäßig unbeweglich und eignen sich jenseits formaler Nomenklaturen und Klassifizierungen eher wenig zur Lösung der Datenqualitätsprobleme in dynamischen Forschungsfeldern.

Datenvalidierung und organisations- beziehungsweise prozessbezogene Operationalisierungen von Datenqualität durch Zertifikate oder Siegel setzen auf Aufwertung vor allem von Datenrepositorien. Indem sie Vertrauen stiften, setzen sie positive Anreize für Forschende, Daten an entsprechend konformitätsbewertete Einrichtungen abzugeben. Ansonsten haben sie unter Umständen wenig unmittelbaren Bezug zu den konkreten Forschungsprozessen, in denen die Daten entstehen.

Leitlinien und Policies auf der Ebene von Organisationen und Förderern steuern primär das Management des Umgangs mit Forschungsdaten. Als vertragsartige beziehungsweise verständigungsorientierte Instrumente zur Steuerung von Datenqualität erreichen sie zwar das *Floor Level* konkreter Forschungsprozesse, werden dort aber häufig nur vordergründig bedient. Ein direkter Zusammenhang mit den Grundprinzipien guter wissenschaftlicher Praxis wird seitens der Forschenden eher nicht hergestellt.

Ein primär verfahrenstechnischer Ansatz zur Steuerung von Datenqualität sind idealtypische Beschreibungen zu optimierender Prozesse, der Datenlebenszyklus ist hierfür ein Beispiel. Er bietet eine sehr gute Orientierungshilfe, um sich als Wissenschaftler Fragen der geeigneten Dokumentation von Daten – sowie der datenverarbeitenden Geräte – an den verschiedenen Schnittstellen von Zyklus und Forschungsprozess zu stellen. Über die genauere Ausgestaltung

der Schnittstellen mit Blick auf spezifische Anforderungen von Disziplinen- oder Forschungsfelder sagt er allerdings noch nichts aus. Auf verschiedene Forschungsformen und Forschungsbereiche abgestimmte Zyklen könnten ein Ansatz sein.

Prinzipien und Leitlinien wie sie aus Überlegungen des Total Data Quality Management, aus „fit for use“ und „fit for purpose“ abgeleitet worden sind, stellen pragmatische und primär prozedurale Ideen für die Steuerung von Qualitätsentwicklungen zur Verfügung. Hier lassen sich auch die FAIR-Prinzipien einordnen, die in der Wissenschaft zuletzt die größte Resonanz ausgelöst haben und zwischenzeitlich nicht nur in vielfältige Leitlinien und Policies auf organisatorischer Ebene umgesetzt worden sind, sondern auch die datenbezogene Wissenschaftspolitik in Europa maßgeblich orientieren. Allerdings ist auch hier zu konstatieren, dass ein unmittelbarer Rückbezug der Prinzipien auf fachlich-disziplinäre Gütekriterien, die aus der Wissenschaft selbst kommen müssen, noch geleistet werden muss.

Die Arbeitsgruppe „Datenqualität“ des RfII hat auf Grundlage der Erkenntnisse, die aus der Sichtung der hier dargestellten Ansätze, Konzepte und Modelle resultieren, Vorschläge für Empfehlungen zur Weiterentwicklung der Datenqualität in der Wissenschaft entwickelt.

LITERATURVERZEICHNIS

Allianz der Wissenschaftsorganisationen (2003): Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen, Berlin, 4 S., online verfügbar unter: http://openaccess.mpg.de/68053/Berliner_Erklaerung_dt_Version_07-2006.pdf, zuletzt geprüft am: 30.08.2019.

Allianz der Wissenschaftsorganisationen (2010): Grundsätze zum Umgang mit Forschungsdaten, 2 S., DOI: 10.2312/ALLIANZOA.019, zuletzt geprüft am: 30.08.2019.

Allianz-Initiative Digitale Information (2012): Schwerpunktinitiative „Digitale Information" der Allianz der deutschen Wissenschaftsorganisationen – Fortsetzung der Zusammenarbeit in den Jahren 2013 bis 2017. Leitbild 2013–2017, 16 S., DOI: 10.2312/ALLIANZOA.018, zuletzt geprüft am: 30.08.2019.

Allianz-Initiative Digitale Information- AG Forschungsdaten (2015): Positionspapier „Research data at your fingertips" der Arbeitsgruppe Forschungsdaten, Potsdam, 4 S., DOI: 10.2312/allianzfd.001, zuletzt geprüft am: 30.08.2019.

Al-Ruithe, Majid/Benkhelifa, Elhadj/Hameed, Khawar (2018): A Systematic Literature Review of Data Governance and Cloud Data Governance, in: Pers Ubiquit Comput, Jg. 1, Nr. 1, 1–21, DOI: 10.1007/s00779-017-1104-3, zuletzt geprüft am: 30.08.2019.

Amann, Rudolf I. et al. (2019): Toward Unrestricted Use of Public Genomic Data, in: Science 363, Nr. 6425, S. 350–352, DOI: 10.1126/science.aaw1280, zuletzt geprüft am: 30.08.2019.

Baker, Karen S./Duerr, Ruth E./Parsons, Mark A. (2015): Scientific Knowledge Mobilization. Co-evolution of Data Products and Designated Communities, in: International Journal of Digital Curation, Jg. 10, Nr. 2, S. 110–135, DOI: 10.2218/ijdc.v10i2.346, zuletzt geprüft am: 30.08.2019.

Batini, Carlo/Scannapieca, Monica (2006): Data Quality. Concepts, Methodologies and Techniques, Berlin, Heidelberg: Springer, 262 S.

Beauchamp, Tom L./Childress, James F. (1983): Principles of Biomedical Ethics, 2. Auflage, New York: Oxford Univ. Press, 364 S.

BMBF- Bundesministerium für Bildung und Forschung (2012): Bekanntmachung des Bundesministeriums für Bildung und Forschung von Richtlinien zur Förderung von Forschung im Bereich „Sprachliche Bildung und Mehrsprachigkeit", 11 S., online verfügbar unter: https://www.bmbf.de/foerderungen/bekanntmachung.php?B=774, zuletzt geprüft am: 30.08.2019.

Burgelman, Jean-Claude (2017): From Vision to Action. From Open to FAIR Data. 8th Open-AIRE workshop, 2 S., online verfügbar unter: https://de.slideshare.net/OpenAIRE_eu/horizon-2020-open-research-data-pilot-jeanclaude-burgelman-dg-rtd-european-commission-8th-openaire-workshop, zuletzt geprüft am: 30.08.2019.

Chang, Chung (2012): Data Life Cycle, online verfügbar unter: https://blogs.princeton.edu/onpopdata/2012/03/12/data-life-cycle/, zuletzt geprüft am: 30.08.2019.

Cousijn, Helena/Cruse, Patricia/Fenner, Martin (2018): Taking Discoverability to the Next Level: Datasets with DataCite DOIs Can Now Be Found Through Google Dataset Search, in: DataCite Blog, DOI: 10.5438/5AEP-2N86, zuletzt geprüft am: 30.08.2019.

DCC – Digital Curation Center (2014): Five Steps to Developing a Research Data Policy, 2 S., online verfügbar unter: http://www.dcc.ac.uk/sites/default/files/documents/publications/DCC-FiveStepsToDevelopingAnRDMpolicy.pdf, zuletzt geprüft am: 30.08.2019.

DFG – Deutsche Forschungsgemeinschaft (2013): Sicherung guter wissenschaftlicher Praxis. Denkschrift. Empfehlungen der Kommission zur Selbstkontrolle in der Wissenschaft, Bonn, 112 S., online verfügbar unter: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf, zuletzt geprüft am: 30.08.2019.

DFG – Deutsche Forschungsgemeinschaft (2015): Leitlinien zum Umgang mit Forschungsdaten, Bonn, 2 S., online verfügbar unter: http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf, zuletzt geprüft am: 30.08.2019.

DFG – Deutsche Forschungsgemeinschaft (2019): Leitlinien zur Sicherung guter wissenschaftlicher Praxis (Kodex), Bonn, 32 S., online verfügbar unter: https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf, zuletzt geprüft am: 30.08.2019.

Donnelly, Martin (2008): RDMF2. Core Skills Diagram, in: Research Data Management Forum, online verfügbar unter: http://data-forum.blogspot.de/2008/12/rdmf2-core-skills-diagram.html, zuletzt geprüft am: 30.08.2019.

DSA – Data Seal of Approval/ ICSU-WDS- International Council for Science-World Data System (2016): Core Trustworthy Data Repositories Requirements, 14 S., online verfügbar unter: https://drive.google.com/file/d/0B4qnUFYMgSc-eDRSTE53bDUwd28/view, zuletzt geprüft am: 30.08.2019.

EC – European Commission (2017): H2020 Programme. Multi-Beneficiary General Model Grant Agreement. Version 5.0, 167 S., online verfügbar unter: http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf, zuletzt geprüft am: 30.08.2019.

EC – European Commission (2018): H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020. Version 3.0, Brüssel, 12 S., online verfügbar unter: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, zuletzt geprüft am: 30.08.2019.

EC – European Commission/EU- European Union (2007): Richtlinie 2007/2/EG des Europäischen Parlaments und des Rates vom 14. März 2007 zur Schaffung einer Geodateninfrastruktur in der Europäischen Gemeinschaft (INSPIRE), Amtsblatt der Europäischen Union, 14 S., online verfügbar unter: http://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32007L0002&from=EN, zuletzt geprüft am: 30.08.2019.

EC – European Commission (2013): Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Version 1.0. The EU Framework Programme for Research and Innovation, Brüssel, 14 S., online verfügbar unter: http://www.gsrt.gr/EOX/files/h2020-hi-oa-pilot-guide_en.pdf, zuletzt geprüft am: 30.08.2019.

HEFCE – Higher Education Funding Council for England et al. (2016): Concordat on Open Research Data, 24 S., online verfügbar unter: https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/, zuletzt geprüft am: 30.08.2019.

Hodson, Simon et al. (2018): Fair Data Action Plan. Interim Recommendations and Actions from the European Commission Expert Group on Fair Data, 21 S., DOI: 10.5281/ZENODO.1285290, zuletzt geprüft am: 30.08.2019.

HRK – Hochschulrektorenkonferenz (2012): Hochschule im digitalen Zeitalter. Informationskompetenz neu begreifen – Prozesse anders steuern. Entschließung der 13. Mitgliederversammlung der HRK am 20. November 2012 in Göttingen, Bonn, 24 S., online verfügbar unter: http://www.hrk.de/uploads/media/Entschliessung_Informationskompetenz_20112012_01.pdf, zuletzt geprüft am: 30.08.2019.

HRK – Hochschulrektorenkonferenz (2014): Management von Forschungsdaten. Eine zentrale strategische Herausforderung für Hochschulleitungen. Empfehlung der 16. Mitgliederversammlung der HRK am 13. Mai 2014 in Frankfurt am Main, Bonn, 6 S., online verfügbar unter: https://www.hrk.de/uploads/tx_szconvention/HRK_Empfehlung_Forschungsdaten_ 13052014_01.pdf, zuletzt geprüft am: 30.08.2019.

Juran, Joseph M. (1951): Quality-Control Handbook (1), 1.Aufl., New York [u.a.]: McGraw-Hill, 800 S.

Juran, Joseph M. (2010): Attaining Superior Results Through Quality, in: Juran, Joseph M./Feo, Joseph A. de (Hg.): Juran's Quality Handbook (6), 6. Aufl., New York: McGraw-Hill, S. 3–40.

Juran, Joseph M. (2010): The Universal Methods to Manage for Quality, in: Juran, Joseph M./Feo, Joseph A. de (Hg.): Juran's Quality Handbook (6), 6. Aufl., New York: McGraw-Hill, S. 69–82.

LEARN (2017): Toolkit of Best Practice for Research Data Management, 185 S., DOI: 10.14324/000.learn.00, zuletzt geprüft am: 30.08.2019.

Lorentz Center (2014): Jointly Designing a Data Fairport. Conference Report. 13–16 of January 2014, 13 S., online verfügbar unter: https://www.lorentzcenter.nl/lc/web/2014/602/extra.pdf, zuletzt geprüft am: 30.08.2019.

Morphy, Erika (2018): What is Google Dataset Search?, online verfügbar unter: https://www.cmswire.com/big-data/what-is- google-dataset-search/, zuletzt geprüft am: 30.08.2019.

NAKO – Der Vorstand des Nationale Kohorte e. V. (2015): Datenschutz- und IT-Sicherheitskonzept der Gesundheitsstudie NAKO, Heidelberg, 105 S., online verfügbar unter: https://nako.de/wp-content/uploads/2015/09/Datenschutzkonzept-NAKO-Gesundheitsstudie_v2.37_2015-12-03.pdf, zuletzt geprüft am: 30.08.2019.

Nestor – Arbeitsgruppe Digitale Bestandserhaltung (2012): Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung. Version 2.0 (nestor-Materialien, 15), 88 S., online verfügbar unter: https://d-nb.info/1047612364/34, zuletzt geprüft am: 30.08.2019.

Nestor – Arbeitsgruppe Zertifizierung (2019): Erläuterungen zum nestor-Siegel für vertrauenswürdige digitale Langzeitarchive. Version 2.1 (nestor-Materialien, 17), 61 S., online verfügbar unter: https://d-nb.info/1189191830/34, zuletzt geprüft am: 30.08.2019.

Nonnemacher, Michael/Nasseh, Daniel/Stausberg, Jürgen (2014): Datenqualität in der medizinischen Forschung. Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Registern, 2. aktualisierte und erweiterte Auflage, Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.

OECD – Organisation for Economic Co-operation and Development (2007): Principles and Guidelines for Access to Research Data from Public Funding, Paris, 24 S., online verfügbar unter: http://www.oecd.org/sti/sci-tech/38500813.pdf, zuletzt geprüft am: 30.08.2019.

Peer, Limor/Green, Ann/Stephenson, Elizabeth (2014): Committing to Data Quality Review, in: IJDC – International Journal of Digital Curation, Jg. 9, Nr. 1, S. 263–291, DOI: 10.2218/ijdc.v9i1.317, zuletzt geprüft am: 30.08.2019.

Rafique, Irfan et al. (2012): Information Quality Evaluation Framework: Extending ISO 25012 Data Quality Model, in: International Journal of Computer, Electrical, Automation, Control and Information Engineering, Jg. 6, Nr. 5, S. 568–573, online verfügbar unter: https://waset.org/publications/9538/information-quality-evaluation-framework-extending-iso-25012-data-quality- model, zuletzt geprüft am: 30.08.2019.

RatSWD – Rat für Sozial- und Wirtschaftsdaten (2010): Kriterien des RatSWD für die Einrichtung der Forschungsdaten-Infrastruktur, Berlin, 6 S., online verfügbar unter: https://www.ianus-fdz.de/attachments/download/333/RatSWD_Kriterien-Einrichtung-FDZ_09-2010.pdf, zuletzt geprüft am: 30.08.2019.

RatSWD – Rat für Sozial- und Wirtschaftsdaten (2017): Qualitätssicherung der vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) akkreditierten Forschungsdatenzentren (FDZ), DOI: 10.17620/02671.4, zuletzt geprüft am: 30.08.2019.

Redman, Thomas C. (1999): Second-Generation Data Quality Systems, in: Juran, Joseph M./Godfrey, A. Blanton (Hg.): Juran's Quality Handbook (5), 5. Aufl., New York [u. a.]: McGraw-Hill, S. 34.1–34.14.

RfII – Rat für Informationsinfrastrukturen (2016): Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, 160 S., online verfügbar unter: http://d-nb.info/1104292440/34, zuletzt geprüft am: 30.08.2019.

RfII – Rat für Informationsinfrastrukturen (2017): Entwicklung von Forschungsdateninfrastrukturen im internationalen Vergleich. Bericht und Anregungen, Göttingen, 94 S., online verfügbar unter: http://d-nb.info/1143737180/34, zuletzt geprüft am: 30.08.2019.

RIN – Research Information Network (2008): To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. A Report Commissioned by the Research Information Network. Main Report, 56 S., online verfügbar unter: https://eprints.soton.ac.uk/266742/1/Published_report_-_main_-_final.pdf, zuletzt geprüft am: 30.08.2019.

Shankaranarayanan, G./Blake, Roger (2017): From Content to Context: The Evolution and Growth of Data Quality Research, in: Journal of Data and Information Quality, Jg. 8, Nr. 2, S. 1–28, DOI: 10.1145/2996198, zuletzt geprüft am: 30.08.2019.

Swan, Alma/Brown, Sheridan (2008): The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs. Report to the Jisc, 34 S., online verfügbar unter: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf, zuletzt geprüft am: 30.08.2019.

Taylor, Richard N./Medvidovic, Nenad/Dashofy, Eric M. (2010): Software Architecture. Foundations, Theory, and Practice, Safari Tech Books Online, Hoboken, NJ: Wiley, 750 S., online verfügbar unter: http://proquest.tech.safaribooksonline.de/9780470167748, zuletzt geprüft am: 30.08.2019.

Treloar, Andrew/Harboe-Ree, Cathrine (2008): Data Management and the Curation Continuum: How the Monash Experience is Informing Repository Relationships, 13 S., online verfügbar unter: https://ndownloader.figshare.com/files/9799513, zuletzt geprüft am: 30.08.2019.

Wang, Richard Y. (1998): A Product Perspective on Total Data Quality Management, in: CACM – Communications of the ACM, Jg. 41, Nr. 2, S. 58–65, DOI: 10.1145/269012.269022, zuletzt geprüft am: 30.08.2019.

Wang, Richard Y./Strong, Diane M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers, in: Journal of Management Information Systems 12, Nr. 4, S. 5–33, online verfügbar unter: http://www.jstor.org/stable/40398176, zuletzt geprüft am: 30.08.2019.

Wilkinson, Mark D. et al. (2016): The FAIR Guiding Principles for Scientific Data Management and Stewardship, in: Scientific Data, Jg. 3, S. 1–9, DOI: 10.1038/sdata.2016.18, zuletzt geprüft am: 30.08.2019.

Working Group on Information Systems and Services (2012): Data Life Cycle Models and Concepts. CEOS Version 1.2, 90 S., online verfügbar unter: https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts%20v13.docx?api=v2, zuletzt geprüft am: 30.08.2019.

WR – Wissenschaftsrat (2012): Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. Drs. 2359-12, Berlin, 90 S., online verfügbar unter: http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf, zuletzt geprüft am: 30.08.2019.

ONLINE-RESSOURCEN

CIDOC CRM – International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM), Informationsseite
http://www.cidoc-crm.org/home

CMI – The Dublin Core Metadata Initiative
http://dublincore.org/about/

DCC – Digital Curation Centre
http://www.dcc.ac.uk/

DDI – Data Documentation Initiative
https://www.ddialliance.org

Deutsche Nationalbibliothek, Standardisierungsausschuss
https://www.dnb.de/DE/Professionell/Standardisierung/Standardisierungsausschuss/standardisierungsausschuss_node.html

DIN-Normenausschuss Information und Dokumentation (NID)
https://www.din.de/de/mitwirken/normenausschuesse/nid

ECCMA – The Electronic Commerce Code Management Association
https://eccma.org/about_eccma/

forschungsdaten.info – Forschungsdatenmanagement in den Bundesländern
https://www.forschungsdaten.info/praxis-kompakt/fdm-in-den-bundeslaendern/

forschungsdaten.org – Projekt FDMentor
http://www.forschungsdaten.org/index.php/FDMentor

GEOSS Standards and Interoperability Registry
https://www.earthobservations.org/gci_sr.shtml

GFBio – German Federation for Biological Data
https://www.gfbio.org/

GO FAIR
https://www.go-fair.org/

IEEE-SA – Institute of Electrical and Electronics Engineers Standards Association
https://standards.ieee.org/

INSPIRE Validator
http://inspire-sandbox.jrc.ec.europa.eu/validator/

International Code of Nomenclature for algae, fungi, and plants
https://www.iapt-taxon.org/nomen/main.php

International Code of Nomenclature of Bacteria
http://www.the-icsp.org/

ISO Technical Committee 20/SC 13 – Space Data and Information Transfer Systems
https://www.iso.org/committee/46612.html

Metadata Basics
http://www.metadataetc.org/metadatabasics/overview.htm

OAIS – Open Archival Information System Reference Model
http://www.oais.info/

RAL Gütezeichen
https://www.ral-guetezeichen.de/ueber-uns/ral-guetezeichen-historie/

RDA – Research Data Alliance, ICT Standards
https://www.rd-alliance.org/recommendations-outputs/standards

RDMO – Research Data Management Organiser
https://rdmorganiser.github.io/

Schema.org Community für strukturierte Daten im Internet
https://schema.org/

TEI – Text Encoding Initiative
https://tei-c.org/about/

The World Wide Web Consortium
https://www.w3.org

W3C – Markup Validation Service
https://validator.w3.org/

VERZEICHNIS DER STANDARDS UND NORMEN

DIN 31644 – Kriterien für vertrauenswürdige digitale Langzeitarchive
https://www.din.de/de/mitwirken/normenausschuesse/nid/normen/wdc-beuth:din21:147058907

DIN 66399 – Büro- und Datentechnik – Vernichten von Datenträgern
https://www.din.de/de/mitwirken/normenausschuesse/nia/normen/wdc-beuth:din21:155420083

ISO 14721:2012 – Open Archival Information System (OAIS) Reference Model
https://www.iso.org/standard/57284.html

ISO 15836-1:2017 – Information and Documentation – The Dublin Core Metadata Element Set- Part 1: Core Clements
https://www.iso.org/standard/71339.html

ISO 16363:2012 – Space Data and Information Transfer Systems-- Audit and Certification of Trustworthy Digital Repositories
https://www.iso.org/standard/56510.html

ISO 24610-1:2006 – Language Resource Management-- Feature Structures-- Part 1: Feature Structure Representation
https://www.iso.org/standard/37324.html

ISO 25964 – The International Standard for Thesauri and Interoperability with other Vocabularies
https://www.niso.org/schemas/iso25964

ISO 2788:1986 – Documentation-- Guidelines for the Establishment and Development of Monolingual Thesauri
https://www.iso.org/standard/7776.html

ISO 5964:1985 – Documentation-- Guidelines for the Establishment and Development of Multilingual Thesauri
https://www.iso.org/standard/12159.html

ISO/IEC 25012:2008- Software Engineering-- Software Product Quality Requirements and Evaluation (SQuaRE) -- Data Quality Model
https://www.iso.org/standard/35736.html

ISO/IEC 25024:2015(en) – Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — Measurement of Data Quality
https://www.iso.org/obp/ui/#iso:std:iso-iec:25024:ed-1:v1:en

ISO/TS 8000-1:2011 – Data Quality-- Part 1: Overview
https://www.iso.org/standard/50798.html
https://eccma.org/iso-8000/

ISO 9001:2015 – Quality Management Systems – Requirements
https://www.iso.org/standard/62085.html

ISO 16919:2014 – Space Data and Information Transfer Systems-- Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories
https://www.iso.org/standard/57950.html

B. BEGRIFFSBESTIMMUNGEN

B. BEGRIFFSBESTIMMUNGEN

Der vorbereitende Ausschuss Datenqualität hat im Zuge seiner Arbeit auch zwei der Begriffsdefinitionen des RfII überarbeitet, die 2016 im Positionspapier LEISTUNG AUS VIELFALT veröffentlicht wurden. Das Plenum hat die nachstehende Neuformulierung in der 10. Ratssitzung vom November 2017 verabschiedet.

DATENQUALITÄT
[data quality]

Der Begriff Datenqualität bezeichnet sowohl allgemeine, unter anderem unter Methodengesichtspunkten geforderte, typische Eigenschaften der Daten selbst als auch deren durch qualitätssichernde Maßnahmen gegebenenfalls zusätzlich geschaffene Eignung für eine weitere Nutzung.

Die Bewertung von Datenqualität richtet sich zum einen nach den abhängig von der jeweiligen Forschungsfrage und damit von der Verwendung zur Erarbeitung eines Forschungsergebnisses zu definierenden Ansprüchen an die Daten. Diese umfassen etwa die Genauigkeit von Messwerten, die Zuverlässigkeit eines empirisch gewonnenen Ergebnisses, die Vollständigkeit oder Aktualität von Daten und die Dokumentation der Datengewinnung und der Datenspeicherung.

Darüber hinaus sind aber auch Nachhaltigkeitsgesichtspunkte von der Bewertung der spezifischen Qualität von Daten nicht zu trennen. Solche Gesichtspunkte umfassen die Datenbeschaffenheit, etwa die Transferierbarkeit der Daten, die Haltbarkeit von Datenträgern etc. Sie betreffen insbesondere die vorausschauende Haltung von Forschungsdaten für spätere, idealerweise viele und gegebenenfalls auch noch unbekannte, Formen der wissenschaftlichen, wirtschaftlichen und gesellschaftlichen Nutzung.

Unter dem Gesichtspunkt einer weiteren Nutzung („Nachnutzung“) wird Datenqualität dadurch bestimmt, dass Datensätze und-sammlungen leicht zu recherchieren/auffindbar sind und dass sie ausreichend Zusatzinformationen beinhalten. Diese sollten in Form von möglichst standardisierten technischen und fachlichen Metadaten zu Qualitätsaspekten vorliegen und Auskunft über die Datengenerierung, Weiterverarbeitung und die verwendeten Instrumente und Methoden geben. Voraussetzung für die Nachvollziehbarkeit und wenn möglich Nachnutzung von digitalen Forschungsergebnissen ist, dass die enthaltenen Daten bezüglich der ihnen zugrunde liegenden Datenmodelle (verwendete Vokabulare, Formate, etc.) und der verwendeten Methoden (etwa Messgeräte, Befragungen, Algorithmen etc.) umfassend dokumentiert sind. Wo immer möglich, sollten nicht nur Metadaten, sondern auch weitergehende, gegebenenfalls spezielle Dokumentationen, anerkannten, verfügbaren Standards folgen.

Die – auch langfristige – Verfügbarkeit, Zugänglichkeit und Zitierbarkeit von Forschungsdaten einschließlich ihrer Metadaten sind wiederum Aspekte der Qualität von Informationsinfrastrukturen und- services, welche die sichere Speicherung, das zielgenaue Auffinden (*Retrieval*), den Zugriff auf die Daten und ihre Nachnutzung (auch im Kontext der Langzeitarchivierung) ermöglichen. Die Klärung von rechtlichen Rahmenbedingungen einer möglichen Datennutzung gehört, im Zusammenhang mit informationsinfrastrukturellen Diensten, ebenfalls zur Datenqualität.

- Quellen, Positionspapiere

Zu guter wissenschaftliche Praxis: Allianz der Wissenschaftsorganisationen (2003) – Berliner Erklärung und DFG (2013) – Sicherung guter wissenschaftlicher Praxis, S. 21–22; DFG (2019) – Leitlinien zur Sicherung guter wissenschaftlicher Praxis; zu Datenqualität: OECD (2007) – Access to Research Data.

FORSCHUNGSDATEN, FORSCHUNGSDATENMANAGEMENT
[research data, research data management]

Forschungsdaten sind nicht allein die (End-)Ergebnisse von Forschung. Es handelt sich vielmehr um jegliche Daten, die im Zuge wissenschaftlichen Arbeitens entstehen, zum Beispiel durch Beobachtungen, Experimente, Simulationsrechnungen, Erhebungen, Befragungen, Quellenforschungen, Aufzeichnungen, Digitalisierung, Auswertungen. Zu Forschungsdaten werden auch solche, nicht selbst gewonnenen Daten, auf die die Wissenschaft zu Forschungszwecken zugreift, um sie für den konkreten Forschungsprozess als methodisch erforderliche Grundlage zu nutzen. Dies ist zum Beispiel gegeben, wenn amtliche Statistiken oder andere Behördendaten oder Produkte nicht wissenschaftlicher Dienstleister wissenschaftlich verarbeitet werden. Dass auch die verwendeten Forschungswerkzeuge sowie die mitlaufend entstehenden Spuren wissenschaftlicher Arbeit – also Prozessdaten, wie sie namentlich die digitale Forschung vielfach automatisch hervorbringt – zu den Forschungsdaten zählen, ist überall dort von Bedeutung, wo die Dokumentation und Archivierung von Forschungsprozessen und Forschungsdaten zu deren Qualitätssicherung gehört oder aber unter Nachhaltigkeitsgesichtspunkten sowie für die historische Forschung geboten ist. Pragmatisch, wenn auch nicht immer trennscharf, lassen sich Forschungsdaten von-metadaten unterscheiden. Metadaten dokumentieren und kontextualisieren den Entstehungsprozess von Forschungsdaten. Im Forschungsprozess können Metadaten selbst wieder Gegenstand von Forschung werden, was unter anderem für den Lebenszyklus von Forschungsdaten von Bedeutung ist.

Das Forschungsdatenmanagement umfasst alle – über das Forscherhandeln im engeren Sinne hinaus auch organisationsbezogenen – Maßnahmen, die getroffen werden müssen, um qualitätsvolle Daten zu gewinnen, um die gute wissenschaftliche Praxis im Datenlebenszyklus einzuhalten, um Ergebnisse reproduzierbar zu machen und um ggf. bestehenden Dokumentationsverpflichtungen Rechnung zu tragen (zum Beispiel im Gesundheitswesen). Auch ist die (ggf. domänenübergreifende) Verfügbarkeit von Daten zur Nachnutzung ein wichtiger Punkt. Zunehmend finden Datenmanagementpläne in wissenschaftlichen Institutionen Anwendung. Datenmanagementpläne, die zu Anfang eines Projekts entwickelt und niedergelegt werden oder das Ergebnis einer Forschungsarbeit sind, sollen die zu nutzenden und zu generierenden Daten und die notwendigen Dokumentationen, Metadaten und Standards beschreiben, mögliche rechtliche Einschränkungen (zum Beispiel Datenschutz) rechtzeitig benennen, benötigte Speicherressourcen einplanen sowie Kriterien festlegen, welche Daten Externen in welcher Form verfügbar gemacht werden und wie die Daten langfristig zu sichern sind. Auf der Organisationsebene müssen forschende Einrichtungen (zum Beispiel Hochschulen) den Zugang zu entsprechenden Infrastrukturdiensten innerhalb der Einrichtung (zum Beispiel durch Auf- und Ausbau passender Kapazitäten) oder in Zusammenarbeit mit externen Partnern (durch Kooperationsverträge etc.) ermöglichen. Dabei sollte aktiv auch auf das übergeordnete Ziel einer domänenübergreifenden, wissenschaftsweiten Datennutzung hingearbeitet werden.

- Quellen, Positionspapiere

Allianz-Initiative Digitale Information – AG Forschungsdaten (2015) – Research data at your fingertips; zu Forschungsdaten = Grundlage von Argumentation und Rechnen: EC (2013) – Guidelines on Open Access, S. 3; zu Forschungsdaten = Primärquelle wissenschaftlicher Aktivität: OECD (2007) – Access to Research Data, S. 13; zu Forschungsdaten aus Perspektive der Sozialwissenschaften: RatSWD (2010) – Kriterien Forschungsdaten-Infrastruktur, S. 4; zu Forschungsdaten als Daten aus dem Forschungsprozess: Allianz-Initiative (2012) – Leitbild 2013–2017, S. 7; WR (2012) – Empfehlungen zu Informationsinfrastrukturen, S. 53–57; DCC – Data Management Plans, http://www.dcc.ac.uk/resources/data-management-plans (zuletzt geprüft am: 30.08.2019); DFG (2015) – Leitlinien Forschungsdaten; HRK (2014) – Management von Forschungsdaten; HRK (2012) – Hochschule im digitalen Zeitalter.

C. MITWIRKENDE

C. MITWIRKENDE

C.1 RATSMITGLIEDER

Vertreter der wissenschaftlichen Nutzer

Prof. Dr. Marion Albers
Fakultät für Rechtswissenschaften, Universität Hamburg

Prof. Dr. Lars Bernard
Fakultät für Umweltwissenschaften, Technische Universität Dresden

Prof. Dr. Stefan Decker
FIT – Fraunhofer-Institut für Angewandte Informationstechnik und RWTH Aachen

Prof. Dr. Petra Gehring (Vorsitzende)
Fachbereich Gesellschafts- und Geschichtswissenschaften, Technische Universität Darmstadt

Prof. Dr. Kurt Kremer
MPI – Max-Planck-Institut für Polymerforschung

Prof. Dr. Wolfgang Marquardt
Forschungszentrum Jülich GmbH

Prof. Dr. Joachim Wambsganß
ZAH – Zentrum für Astronomie der Universität Heidelberg

Prof. Dr. Doris Wedlich
KIT – Karlsruher Institut für Technologie – Bereich I: Biologie, Chemie und Verfahrenstechnik

Vertreter von Bund und Ländern

Rüdiger Eichel
Niedersächsisches Ministerium für Wissenschaft und Kultur

Dr. Hans-Josef Linkens
Bundesministerium für Bildung und Forschung

Dr. Dietrich Nelle
Bundesministerium für Bildung und Forschung

Annette Storsberg
Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

Vertreter der Einrichtungen

Sabine Brünger-Weilandt
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH

Prof. Dr. Dr. h.c. Friederike Fless
DAI – Deutsches Archäologisches Institut und Freie Universität Berlin

Prof. Dr. Michael Jäckel
Universität Trier

Prof. Dr. Stefan Liebig (stellv. Vorsitzender)
DIW – Deutsches Institut für Wirtschaftsforschung

Prof. Dr. Sandra Richter
Deutsches Literaturarchiv Marbach

Katrin Stump
Universitätsbibliothek Braunschweig

Prof. Dr. Klaus Tochtermann
ZBW – Leibniz-Informationszentrum Wirtschaft und Christian-Albrechts-Universität zu Kiel

Prof. Dr. Ramin Yahyapour
GWDG – Gesellschaft für Wissenschaftliche Datenverarbeitung mbH und Georg-August-Universität Göttingen

Vertreter des öffentlichen Lebens

Dr. Anke Beck
IntechOpen Verlag

Marit Hansen
Landesbeauftragte für Datenschutz Schleswig-Holstein

Dr. Nicola Jentzsch
SNV – Stiftung Neue Verantwortung (bis 03/2019)

Dr. Harald Schöning
Software AG

C.2 PROJEKT DATENQUALITÄT

Ausschuss Datenqualität

Prof. Dr. Frank Oliver Glöckner (Moderation), Prof. Dr. Lars Bernard, Prof. Dr. Petra Gehring, Dr. Margit Ksoll-Marcon, Prof. Dr. Stefan Liebig

Arbeitsgruppe Datenqualität

Prof. Dr. Dr. h. c. Friederike Fless (Leitung), Prof. Dr. Marion Albers, Prof. Dr. Lars Bernard, Prof. Dr. Petra Gehring, Prof. Dr. Frank Oliver Glöckner (Gast), Prof. Dr. Stefan Liebig, Dr. Hans-Josef Linkens (vertreten durch Dr. Lena Maerten), Prof. Dr. Doris Wedlich

Redaktionsgruppe

Prof. Dr. Petra Gehring (Leitung), Prof. Dr. Lars Bernard, Prof. Dr. Dr. h. c. Friederike Fless, Prof. Dr. Kurt Kremer, Prof. Dr. Stefan Liebig, Dr. Hans-Josef Linkens (vertreten durch Dr. Lena Maerten)

Die Gremien wurden seitens der RfII-Geschäftsstelle inhaltlich und organisatorisch begleitet durch Dr. Stefan Lange, Dr. Kirsten Gerland, Dr. Ilja Zeitlin, Dr. Sven Rank.

C.3 DANK

Der RfII bedankt sich bei allen Expertinnen und Experten, die sich an der Arbeit der Arbeitsgruppe Datenqualität beteiligt haben.[103]

Dr. Michael Diepenbroek
Prof. Dr. Gerd Gigerenzer
Prof. Dr. Uwe Hasebrink
Prof. Dr. Richard Lenz
Prof. Dr. Thomas Ludwig
Prof. Dr. Eva Schlotheuber
Dr. Markus Schmalzl
Dr. Nico Siegel
Dr. Cornelia Weber

[103] Namentlich aufgeführt sind die Teilnehmerinnen und Teilnehmer der Expertenkonsultationen der Arbeitsgruppe Datenqualität im Mai 2018 und September 2018.

IMPRESSUM

Verabschiedet im September 2019

Rat für Informationsinfrastrukturen (RfII)
Geschäftsstelle
Papendiek 16
37073 Göttingen

Tel. 0551-392 70 50
E-Mail info@rfii.de
Web www.rfii.de

GESTALTUNG, SATZ UND DRUCK
NEFFO DESIGN (Buchholz), Klartext GmbH (Göttingen)

ZITIERVORSCHLAG
RfII – Rat für Informationsinfrastrukturen: Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, Göttingen 2019, 172 S.

Der RfII bevorzugt eine gendergerechte Sprache. In Einzelfällen werden Kollektivbezeichnungen gebraucht, die jeweils Personen aller Geschlechter einbeziehen.

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