WR

WISSENSCHAFTSRAT

2020

Zum Wandel in den Wissenschaften durch datenintensive Forschung

Positionspapier

Vorbemerkung

Der Wissenschaftsrat hat sich mit Aspekten der Digitalisierung im Wissenschaftssystem, die zu den Rahmenbedingungen datenintensiver Forschung gehören, bereits aus verschiedenen Perspektiven befasst. In der Vergangenheit standen dabei Aufgabenentwicklung und Organisation zentraler Hochschuleinrichtungen wie Bibliotheken und Zentren für Kommunikation und Information (Rechenzentren) sowie Organisations- und Finanzierungsstrukturen des Hoch- und Höchstleistungsrechnens im Vordergrund. Im letzten Jahrzehnt hinzugekommen ist die Betrachtung von Informationsinfrastrukturen als System, in deren Rahmen der Wissenschaftsrat eine Unterscheidung verschiedener Forschungsformen vorgeschlagen hat. Diese Forschungsformen – experimentierende, beobachtende, hermeneutisch-interpretierende, begrifflich-theoretische, gestaltende Forschungsform sowie Simulation – weisen erhebliche Unterschiede in der Nutzung von Informationsinfrastrukturen auf, die nicht in einem Schema von Rückständigkeit und Fortschrittlichkeit verstanden werden dürfen, sondern in der Unterschiedlichkeit der Formen des Wissens und der Erkenntnisprozesse in den verschiedenen Bereichen der Wissenschaft gründen. Entsprechend hat der Wissenschaftsrat 2012 differenzierte Empfehlungen zu Informationsinfrastrukturen ausgesprochen und die Einrichtung eines eigenen Rats für Informationsinfrastrukturen (RfII) unterstützt. | [1] Der 2014 gegründete RfII hat inzwischen mehrere Empfehlungen vorgelegt – zu Bedarf und Gestalt einer nationalen Forschungsdateninfrastruktur (NFDI), zu digitalen Kompetenzen für den Arbeitsmarkt Wissenschaft und zu Anforderungen an die Datenqualität in der Wissenschaft. | [2]

| [1] Wissenschaftsrat: Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020, Köln 2012, https://www.wissenschaftsrat.de/download/archiv/2359-12.pdf, zu den Forschungsformen dort S. 35 ff. Alle Weblinks in diesem Positionspapier wurden zuletzt am 23.10.2020 abgerufen.

| [2] Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen 2016, http://www.rfii.de/?p=1998; Rat für Informationsinfrastrukturen: Digitale Kompetenzen – dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft, Göttingen 2019, http://www.rfii.de/?p=3883; Rat für Informationsinfrastrukturen: Herausforderung Datenqualität. Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage, Göttingen 2019, http://www.rfii.de/?p=4043.

In seinem Aufgabenbereich hat der Wissenschaftsrat zuletzt einer einzelnen der von ihm 2012 skizzierten Forschungsformen, nämlich der sich rasch verbreitenden rechnerbasierten Simulation, eigene Empfehlungen gewidmet und auch Vorschläge zur Weiterentwicklung des Hoch- und Höchstleistungsrechnens gemacht. |[3] Ferner erfolgten jüngst eine fachbezogene Betrachtung der Chancen und Herausforderungen der Digitalisierung für die Medizin. |[4]

Mit den hier vorgelegten Überlegungen zum Wandel in den Wissenschaften durch datenintensive Forschung will der Wissenschaftsrat Aufmerksamkeit für Herausforderungen der Digitalisierung im Wissenschaftssystem schaffen, welche über die bislang im Vordergrund stehende Beschäftigung mit Infrastrukturen hinausgehen. In diesem Positionspapier sollen die für viele Forschungsfelder neuen oder doch in ihrem Tempo deutlich beschleunigten Entwicklungen aufgegriffen werden. Ziel ist, diese zu reflektieren und weitergehende Veränderungen im Wissenschaftssystem anzustoßen.

Zur Vorbereitung dieses Positionspapiers hat der Wissenschaftsrat eine Arbeitsgruppe eingesetzt. In ihr haben auch Sachverständige mitgewirkt, die nicht Mitglieder des Wissenschaftsrats sind. Ihnen weiß sich der Wissenschaftsrat zu besonderem Dank verpflichtet. Ebenso dankt der Wissenschaftsrat weiteren Sachverständigen, die den Beratungsprozess im Rahmen von Expertengesprächen und mit Hintergrundinformationen konstruktiv unterstützt haben.

Der Entwurf des vorliegenden Papiers konnte wegen der COVID-19-Pandemie nicht wie geplant im April 2020 dem Wissenschaftsrat vorgelegt werden. Daraus ergab sich die Gelegenheit, erste Beobachtungen zur Rolle von Forschungsdaten im Umgang mit der Pandemie in einem Nachwort festzuhalten. Der Wissenschaftsrat hat das vorliegende Positionspapier am 23. Oktober 2020 verabschiedet.

| [3] Wissenschaftsrat: Bedeutung und Weiterentwicklung von Simulation in der Wissenschaft | Positionspapier (Drs. 4032-14), Dresden Juli 2014, https://www.wissenschaftsrat.de/download/archiv/4032-14.pdf; Wissenschaftsrat: Strategische Weiterentwicklung des Hoch- und Höchstleistungsrechnens in Deutschland | Positionspapier (Drs. 1838-12), Berlin Januar 2012, https://www.wissenschaftsrat.de/download/archiv/1838-12.pdf; Wissenschaftsrat: Empfehlungen zur Finanzierung des Nationalen Hoch- und Höchstleistungsrechnens in Deutschland (Drs. 4488-15), Stuttgart April 2015, https://www.wissenschaftsrat.de/download/archiv/4488-15.pdf. Ergänzend: Frühjahr 2019 nunmehr auch Programm zur Förderung des Nationalen Hochleistungsrechnens, siehe Leitfaden zur Begutachtung von Forschungsbauten – gültig ab Förderphase 2021 (Drs. 7653-19), Hamburg Mai 2019, https://www.wissenschaftsrat.de/download/2019/7653-19.pdf.

| [4] Wissenschaftsrat: Digitalisierung in der Medizin | Bericht der Vorsitzenden zu aktuellen Tendenzen im Wissenschaftssystem (Frühjahrssitzungen des Wissenschaftsrats, 8. bis 10. Mai 2019 in Hamburg), Hamburg 2019, https://www.wissenschaftsrat.de/download/2019/VS-Bericht_Mai_2019.pdf.

Kurzfassung

Datenintensive Forschung – verstanden sowohl als Forschung mit umfangreichen, hochdimensionalen oder neu kombinierten Datenbeständen als auch mit neuen oder weiterentwickelten Methoden – entwickelt sich außerordentlich dynamisch und ist in einer stark wachsenden Zahl von wissenschaftlichen Fachgemeinschaften zum Trend-Thema geworden. Bisher war die Aufmerksamkeit primär auf die Schaffung der Voraussetzungen datenintensiver Forschung einschließlich des systematischeren Auf- und Ausbaus von Informationsinfrastrukturen gerichtet. Mit dem Fortschreiten dieser Entwicklung werden damit verbundene Herausforderungen deutlicher, die von der Methodenausbildung, über Unterstützungsleistungen bei Datenkuratierung, -speicherung und -sicherheit, Kooperationsbedarfe über Fachgebiete und Fächergrenzen hinweg bis zu Fragen von Ethik, Recht und wissenschaftlicher Integrität reichen (dazu Teil A mit entsprechenden Fallbeispielen).

Der Wissenschaftsrat formuliert acht Leitlinien, um datenintensive Forschung im Wissenschaftssystem erfolgreich organisieren und betreiben sowie den Nutzen des Teilens von Daten und Software voll entfalten zu können (dazu Kapitel B.I). Sie sollen einen Kulturwandel in der Wissenschaft leiten und befördern helfen:

1 – Technische, rechtliche, organisatorische und soziale Voraussetzungen und Regeln der Forschung müssen so gestaltet werden, dass sie das Teilen von Daten und Software in der Wissenschaft und ihre kooperative Bearbeitung fördern. Dabei können Zugangsbeschränkungen zu in der Wissenschaft erzeugten oder bearbeiteten Daten nicht nur aus rechtlichen Gründen notwendig sein, bedürfen aber in allen anderen Fällen jeweils einer Begründung.

2 – In den Forschungsfeldern müssen sowohl methodisch verschiedene datenintensive Ansätze als auch andere Forschungsformen nebeneinander bestehen, weiterentwickelt werden und neu entstehen können. Dafür sind Freiräume zu erhalten und Angebote in den Organisations- und Förderbedingungen der Forschung auszubalancieren.

3 – Kompetenzen im Umgang mit Daten und Methoden ihrer Verarbeitung müssen für Tätigkeiten innerhalb und außerhalb der Wissenschaft systematisch vermittelt werden. Dabei sind unterschiedliche Spezialisierungsgrade und Karriere-

pfade zu berücksichtigen, die von allgemeiner Data Literacy über fachspezifische bis zu fachübergreifenden Kompetenzprofilen reichen.

4 – Hochwertige Beiträge zu kuratierten Datensammlungen und Datenpublikationen, zum Forschungsdatenmanagement wie auch zur Methoden- und Softwareentwicklung müssen als genuine Leistungen von Wissenschaftlerinnen und Wissenschaftlern anerkannt werden und Unterstützung erfahren.

5 – Forschungsdaten und die bei ihrer Erzeugung und Verarbeitung verwendeten digitalen Instrumente müssen im Interesse von Wissenschaft und Gesellschaft zuverlässig verfügbar bleiben, um die Reproduzierbarkeit von Ergebnissen zu sichern und weitere Nachnutzungen auch nach mittleren und langen Zeiträumen zu ermöglichen. Eine zentrale Voraussetzung dafür ist die systematische und nachvollziehbare Dokumentation der einzelnen Verarbeitungsschritte.

6 – Das Wissenschaftssystem muss sich den Herausforderungen der Schnelllebigkeit von Datenangeboten, Hard- und Software-Entwicklungen sowie dem immer schnelleren Aufkommen neuer Methoden stellen. Die Offenheit für teils radikale Neuerungen muss zugleich in ein ausgewogenes Verhältnis zum Interesse an Beständigkeit, Reproduzierbarkeit und nachhaltiger Nutzung bewährter Lösungen gebracht werden.

7 – Öffentlich finanzierte Wissenschaft soll neben eigenen Daten auch Forschung mit Daten der öffentlichen Hand und des privaten Sektors vorantreiben und deren Nutzungen in der Gesellschaft unterstützen. In Kooperationen mit nicht-wissenschaftlichen Partnern müssen jedoch wissenschaftliche Standards, die Einhaltung rechtlicher Regelungen und übergreifende Regeln guter wissenschaftlicher Praxis gewährleistet sein.

8 – Wissenschaft muss gerade in diesem Feld den Dialog mit der Gesellschaft suchen, um die Veränderungen ihrer Grundlagen, Fragestellungen und Ergebnisse durch datenintensive Forschung transparent zu machen. Sie muss in diesem Austausch Impulse aus der Gesellschaft zur eigenen Weiterentwicklung aufnehmen.

Zur Konkretisierung dieser Leitlinien werden verschiedene Empfehlungen an zwei zentrale Akteursgruppen im Wissenschaftssystem formuliert (dazu Kapitel B.II): an die Hochschulen und Forschungseinrichtungen sowie an Forschungsförderer mit Bund und Ländern. Aufbauend auf dem Diskurs in den Fachgemeinschaften können diese beiden Akteursgruppen den Kulturwandel im Zusammenhang mit datenintensiver Forschung wesentlich befördern und unterstützen. Wissenschaftliche Einrichtungen sollen sich mit Blick auf datenintensive Forschung nach außen strategisch positionieren und nach innen einen verbindlichen Rahmen für die datenintensiv Forschenden definieren. Sie müssen sich außerdem mit besonderem Ressourcenbedarf auseinandersetzen, lokale Beratungsangebote schaffen, Aus- und Weiterbildungsangebote verändern, An-

reize für qualitätsvolle Datenarbeit schaffen, Forschungs- und Infrastrukturkompetenzen intern verbinden und den Bedarf an neuen Forschungskapazitäten und -strukturen prüfen. Forschungsförderer sowie Bund und Länder müssen Beratungsstrukturen vernetzen, Datenkuratierung finanziell unterstützen, eine ganze Reihe verschiedener neuer Förderbedarfe und -formen prüfen, Begutachtungsprozesse anpassen und auch eine Intensivierung der Wissenschaftskommunikation unterstützen.

A. Datenintensive Forschung

A.I ZUM VERSTÄNDNIS DATENINTENSIVER FORSCHUNG

Wissenschaft kann heute auf immer mehr Daten zurückgreifen und produziert selbst immer mehr Daten. Die massive Zunahme verfügbarer Daten, Erleichterungen bei ihrem Austausch sowie neue Methoden ihrer Analyse eröffnen neue Möglichkeiten für die Forschung, die Auswirkungen auf das Wissenschaftssystem haben. Sehr große Datenbestände, die in hoher Geschwindigkeit oder nahezu in Echtzeit generiert werden, in ihrer Beschaffenheit und Qualität hochgradig divers, in ihrer Abdeckung oft umfassend, hoch aufgelöst, relational, flexibel und skalierbar sind, werden häufig unter dem unbestimmten Begriff Big Data geführt. Angesichts der großen Aufmerksamkeit für Big Data, deren Anwendungsmöglichkeiten seit Anfang des letzten Jahrzehnts Wirtschaft, Politik und Wissenschaft gleichermaßen faszinieren wie beunruhigen, wird vielfach jedoch übersehen, wie sich in den letzten Jahren auch die Erschließung, Verknüpfung, Auswertung und Nachnutzung von kleineren Datenbeständen (Small Data) verändert haben. Daraus ergeben sich ebenfalls vielfältige neue Möglichkeiten und Herausforderungen für die Wissenschaft. Daher umfasst die Betrachtung datenintensiver Forschung, ihrer Voraussetzungen und Auswirkungen im Folgenden Forschung mit sehr unterschiedlich umfangreichen Datenbeständen. | [5]

Daten sind zwar immer schon zentrales Element von Forschung, aber nicht einfach zu definieren. Lange Zeit wurden Daten ausschließlich analog verarbeitet und selbst mit Einsetzen der Digitalisierung seit den 1960er Jahren erfolgte die digitale Datenverarbeitung zunächst nur in engen Grenzen. Neben Daten, die speziell für Forschungszwecke erhoben werden, nutzt Wissenschaft auch solche aus anderen Quellen, etwa Daten aus öffentlichen Verwaltungseinrichtungen, Daten von Unternehmen oder Daten, die bei der Nutzung von digitalen Geräten

| [5] Zu Big Data und Small Data: Kitchin, R.; Lauriault, T. P.: *Small data in the era of big data*, in: GeoJournal, 80 (2015) 4, S. 463–475, DOI: 10.1007/s10708-014-9601-7. Zum Potenzial von Small Data: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag: *Beyond Big Data*, November 2019, http://www.tab-beim-bundestag.de/de/pdf/publikationen/themenprofile/Themenkurzprofil-034.pdf.

mit oder ohne Internetanbindung anfallen (z. B. Maschinenprotokolle, Kontaktdaten, physiologische Werte, Kaufverhalten). Daten können unterschiedlich erzeugt werden und verschiedene Bearbeitungsgrade aufweisen: Sie können aus Beobachtungen mit Hilfe von Instrumenten stammen (z. B. Wetterbeobachtungen mit Satelliten, Online-Befragungen, neuartige und massenhaft verfügbare Sensoren, Hochdurchsatzsequenzierung in den Biowissenschaften), Ergebnis von Berechnungen aus Modellen oder Simulationen sein (z. B. 3-D-Modelle archäologischer Ausgrabungen, Verkehrsmodelle und -simulationen), aus Experimenten (z. B. mit Labortieren oder in Beschleunigeranlagen), aber auch aus unterschiedlichsten Aufzeichnungen (z. B. Akten, Texte, Bilder, Musik oder Filme) hervorgehen, wobei die Grenzen zwischen diesen Kategorien teils fließend sind. Für Forschungszwecke können Daten aus einer oder mehreren Quellen Verwendung finden, die entweder direkt nutzbar sind oder zunächst qualitätsgesichert werden müssen. |[6] Neben Rohdaten und qualitätsgesicherten Primärdaten sind abgeleitete Daten und Metadaten, die ihrerseits Daten beschreiben, zu unterscheiden. Je nach Forschungsfeld kann das Verständnis von Daten weitere Differenzierungen notwendig machen. |[7] Unterschieden wird ferner zwischen Daten, daraus gewinnbaren Informationen und mit deren Hilfe schließlich generierbarem Wissen, wobei die Nutzung als Evidenz in der Begründung von Wissen bisweilen zum Definiens von Daten gezählt wird. |[8]

Nicht nur die Verfügbarkeit von Daten verändert sich außerordentlich dynamisch, sondern auch die Instrumente und Methoden, mit denen diese analysiert werden können. Über die stetige Steigerung der Rechenleistungen hinaus steht eine zunehmend algorithmenbasierte Forschung im Fokus der Aufmerksamkeit, die mit Hilfe neuartiger Software Daten in unterschiedlicher Qualität und aus unterschiedlichen Quellen aggregiert und verknüpft. Durch diese Teilautomatisierung der Forschungsprozesse können Ergebnisse erzeugt werden, deren Zustandekommen für Menschen kaum überprüfbar ist. Dies gilt besonders für bestimmte Verfahren des maschinellen Lernens, die beispielsweise durch die Verarbeitung großer Datenmengen als Trainingsdaten möglich werden oder Strukturen in hochdimensionalen Datensätzen finden und deren Ergebnisse mit gängigen Methoden nicht nachvollziehbar respektive erklärbar sind.

Insgesamt sind Daten und ihre Analyse in einer stark wachsenden Zahl von wissenschaftlichen Fachgemeinschaften zum Trend-Thema geworden. Dabei ist

| 6 Grundsätzlich aus informationswissenschaftlicher Perspektive dazu etwa Borgman, Ch. L.: *Big Data, Little Data, No Data. Scholarship in the Networked World*, Cambridge, MA/London 2015, sowie: Kitchin, R.: *The Data Revolution. Big Data, Open Data, Data Infrastructures and Their Consequences*, Los Angeles u. a. 2014.

| 7 Arbeitsdefinition Daten siehe Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen 2016, http://www.rfii.de/?p=1998, S. A–13.

| 8 So etwa Leonelli, S.: *Data-centric Biology*, *University of Chicago Press*, Chicago 2016, S. 77.

nicht zu übersehen, dass sich die Verfügbarkeit von Daten und die Entwicklung und Akzeptanz neuer Analysemethoden in den einzelnen wissenschaftlichen Fachgemeinschaften gegenwärtig unterschiedlich schnell verändern, was Rückwirkungen auf den Bedarf an und den Entwicklungsstand von Standards und Strukturen hat. Als Vorreiter gelten etwa die Astronomie, die Hochenergiephysik und die Lebenswissenschaften, Teile der Sozial- und Wirtschaftswissenschaften sowie die Archäologie. |[9] Nach wie vor werden jedoch auch Forschungsfragen verfolgt, die weniger auf große Datenmengen angewiesen sind, was innerhalb von Disziplinen zu Auseinandersetzungen führen kann, etwa um die Bedeutung und Angemessenheit quantitativer oder hermeneutischer Methoden. In manchen Disziplinen unterlag die Nutzung quantitativer Ansätze bereits gewissen Konjunkturen, so dass eine Hinwendung zu datenintensiver Forschung teils als Wiederkehr älterer Trends unter neuen Vorzeichen erscheint. Während deshalb manche Wissenschaftlerinnen und Wissenschaftler die neuen Möglichkeiten enthusiastisch aufgreifen, sind andere noch nicht oder gerade erst auf dem Weg zu datenintensiver Forschung, während wieder andere Bereiche datenintensive Vorgehensweisen für Erkenntnissuche und Ergebnisqualität nicht benötigen.

Datenintensive Forschung mit ihren Folgen für das Wissenschaftssystem zum jetzigen Zeitpunkt zu untersuchen, bedeutet eine laufende Entwicklung in den Blick zu nehmen. Rahmenbedingungen wissenschaftlichen Arbeitens erleben an vielen Stellen durch Digitalisierung fortdauernd starke Veränderungen. Deutlich ist aber, dass Big Data und andere Ausprägungen datenintensiver Forschung heute nicht mehr zwingend auf die Rahmenbedingungen von Big Science (früher Großforschung) in großen Organisationsstrukturen angewiesen sind. Soweit ausreichende Netzbandbreiten zur Verfügung stehen, besteht inzwischen fast überall Zugang zu Hardware, Software sowie Datensammlungen, deren Verfügbarkeit und Verwendungsmöglichkeiten noch vor wenigen Jahren die Vorstellungskraft vieler Forschenden überfordert hätte und mit denen viele Wissenschaftlerinnen und Wissenschaftler in ihren jeweiligen Fachgebieten nicht sozialisiert worden sind.

In den letzten Jahren hat sich die Aufmerksamkeit naturgemäß zunächst primär auf die Schaffung der Voraussetzungen datenintensiver Forschung einschließlich des systematischeren Auf- und Ausbaus von Informationsinfrastrukturen gerichtet. Mit dem Fortschreiten dieser Entwicklung werden Heraus-

| 9 Zur Veränderung von Forschungsmöglichkeiten aus Perspektive einzelner Fächer beispielsweise: Leonelli, S.: *Data-Centric Biology. A Philosophical Study*, Chicago/London 2016; Stevens, H.: *Life Out of Sequence. A Data-Driven History of Bioinformatics,* Chicago/London 2013; Jannidis, F.; Kohle, H.; Rehbein, M. (Hrsg.): Digital Humanities. Eine Einführung, Stuttgart 2017; Marres, N.: *Digital Sociology. The Reinvention of Social Research*, Malden, MA 2017; Ash, J.; Kitchin, R.; Leszczynski, A. (Hrsg.): *Digital Geographies*, London u. a. 2019; Sbalzarini, I. F.: Big-Data Analytics transformiert die Lebenswissenschaften, in: Informatik Spektrum, 42 (2020) 6, S. 394–400, DOI: 10.1007/s00287-019-01227-5. Fächerübergreifend: Leonelli, S., Tempini, N. (Hrsg.): *Data Journeys in the Sciences*, *Springer Open* 2020, DOI: 10.1007/978-3-030-37177-7.

forderungen für Methodenausbildung, Unterstützungsleistungen bei Datenkuratierung, -speicherung und -sicherheit, Kooperationsbedarfe über Fachgebiete und Fächergrenzen hinweg, Fragen von Ethik, Recht und wissenschaftlicher Integrität sowie weitere Voraussetzungen und Konsequenzen datenintensiver Forschung erkennbar, ohne dass sich dafür heute schon einfache oder eindeutig beste Lösungen abzeichnen würden.

A.II BEISPIELE DATENINTENSIVER FORSCHUNG

Datenintensive Forschung wird in diesem Positionspapier breit verstanden und umfasst sowohl Forschungsvorhaben oder -felder mit umfangreichen, komplexen oder neu kombinierten Datenbeständen als auch solche mit neuen oder weiterentwickelten Methoden etwa aus dem Bereich des maschinellen Lernens. Entsprechend breit sind die folgenden sieben Fallbeispiele aus verschiedenen Wissenschaftsgebieten ausgewählt, die datenintensive Forschung in ihrer ganzen Bandbreite anschaulich machen sollen. Mit den Fallbeispielen werden Schlaglichter auf veränderte Praktiken und Möglichkeiten, aber auch auf Herausforderungen und Risiken datenintensiver Forschung und deren Folgen für Organisation und Strukturen im Wissenschaftssystem geworfen.

II.1 Automatische Sprachverarbeitung durch Nutzung großer Datensätze

Fallbeispiel 1 schildert deutliche Fortschritte in der maschinellen Sprachübersetzung durch die Entwicklung neuer Methoden in Kombination mit der Nutzung großer Datenbestände und gesteigerter Rechenleistungen. Es steht zugleich stellvertretend für Forschungsbereiche, in denen private Unternehmen durch den Besitz großer Datenbestände und die Verfügbarkeit größerer Rechenkapazitäten den öffentlich finanzierten Forschungseinrichtungen überlegen sind.

Automatische Verarbeitung natürlicher Sprache ist eines der anspruchsvollsten Teilgebiete der Künstlichen Intelligenz (KI). Einsatzfelder sind etwa automatische Übersetzung in Meetings internationaler Organisationen, Vorlesungsübersetzungssysteme an Hochschulen, Dialogübersetzer für Hilfsorganisationen in Krisengebieten oder auch im Tourismus. Ziel ist es, gesprochene Signale in Text umzuwandeln und dabei die Bedeutung des gesprochenen Textes korrekt wiederzugegeben. Bei der maschinellen Übersetzung wird der Text in Text einer anderen Sprache übersetzt und gegebenenfalls wieder in gesprochener Sprache ausgegeben. Um einen Dialog zwischen Menschen in zwei (oder mehr) Sprachen zu ermöglichen, muss die umgekehrte Aufgabe ebenfalls erfüllt werden, also die Rückübersetzung. Erschwerend kommt die große Zahl von mehreren Tausend lebenden Sprachen und die Mehrdeutigkeit menschlicher Sprache hinzu.

Lange verfolgte Ansätze von Informatik und Linguistik, die automatische Sprachverarbeitung mit Hilfe von Regelsystemen und Wörterbüchern zu lösen,

haben zu Fortschritten wie etwa Übersetzungssystemen für spezielle Bereiche geführt. Seit den 1990er Jahren wurden vermehrt neuronale Netze bzw. andere maschinelle Lernmethoden eingesetzt und anhand großer Bestände an Trainingsdaten weiterentwickelt. Die großen Fortschritte der letzten Jahre sind vor allem auf die Nutzung riesiger Datenbestände in Kombination mit methodischen Fortschritten im Bereich des *Deep Learning* und vergrößerten Rechenressourcen zurückzuführen. Aktuelle Erfolge basieren beispielsweise auf der Nutzung von Youtube-Videos, wobei das Training des verwendeten neuronalen Netzes deutlich rechenintensiver ist als mit der typischen Ausstattung eines größeren Universitätsrechenzentrums leistbar.

Einige private Firmen sind öffentlich finanzierten Forschungseinrichtungen inzwischen weit überlegen, aufgrund ihrer Rechenressourcen sowie ihrer nicht öffentlich zugänglichen, umfangreichen und qualitativ hochwertigen Datensätze. Gemeinsame Evaluierungskampagnen in der Forschung im Bereich der maschinellen Übersetzung setzen deshalb auch auf Trainingsdaten, die im Umfang absichtlich beschränkt werden, um die Beteiligung von Forschenden aus öffentlich finanzierten Forschungseinrichtungen mit teils begrenzteren Rechenressourcen zu ermöglichen. | [10]

Weitere Forschungsfragen im Bereich automatischer Sprachverarbeitungssysteme bzw. generell von KI-Systemen, die Fähigkeiten und Arbeitsweisen des Menschen nachbilden sollen, werden vermutlich neben datenintensiver Forschung auch neue methodische Ansätze erfordern. Dazu gehört die Erklärbarkeit von Ergebnissen (warum wurde „Schloss“ mit *castle* und nicht mit *lock* übersetzt?) oder die Entwicklung von Systemen, die lebenslang lernen oder erkennen, wann sie eine Aufgabe nicht (richtig) lösen können. | [11]

II.2 Kooperativ entwickelte Datenbasis im *Global Trade Analysis Project* (GTAP)

Fallbeispiel 2 zeigt die Möglichkeiten einer langfristig kooperativ aufgebauten und nachhaltig fortentwickelten Datensammlung, die ihren Anfang im Bereich der Agrarökonomie hatte und sich zu einer von verschiedenen Fachgebieten nachgefragten Anlaufstelle für Daten sowie schließlich auch für Methoden und Software entwickeln konnte.

Die GTAP-Datenbasis erfasst die weltweite ökonomische Aktivität von 141 Ländern und 65 Sektoren und dient als Grundlage für multi-regionale Simulationsmodelle. Sie entsteht durch eine qualitätskontrollierte Zusammenführung von Daten der öffentlichen Statistik mit weiteren, wissenschaftlich erarbeiteten Da-

| [10] In „gemeinsamen Evaluierungskampagnen“ schließen sich (weltweit) Forschende aus dem Gebiet zusammen und einigen sich auf gemeinsame Benchmark-Daten, Computer, Betriebssystemversionen etc., um ihre Methoden gegeneinander zu evaluieren und dafür Vergleichbarkeit herzustellen.

| [11] Waibel, A.: Sprachbarrieren durchbrechen: Traum oder Wirklichkeit? Nova Acta Leopoldina NF 122, Nr. 410, 2015, S. 101–123; Hinton, G. et al.: *Deep Neural Networks for Acoustic Modelling in Speech Recognition*, in: *IEEE Signal Processing Magazine*, 2 (2012) November, S. 82–97.

ten. Die Datenbasis bestand zum Zeitpunkt ihrer Gründung 1992 aus rein ökonomischen Daten (z. B. Input-Output-Tabellen, bilaterale Handelsdaten, Zolldaten). Diese Kerndatenbasis wurde bis heute stark erweitert (z. B. Daten zur Migration, Rücküberweisungen, Lagerhaltung) und detailliert (z. B. Informationen über Zölle und Steuern). Die ökonomische Kerndatenbasis wurde zudem ergänzt durch kompatibel vernetzte sogenannte *Satellite Databases* mit Daten aus Quellen angrenzender oder auch weiter entfernter Disziplinen (z. B. Informationen zum Klimawandel, demographische Daten oder Daten zum Nahrungsmittelverbrauch), was zu einem deutlich breiteren Einsatz der GTAP-Datenbasis führt.

Die Datenbasis wird finanziell getragen und gemanagt von einem Konsortium aus aktuell 32 nationalen und internationalen staatlichen und privaten Hochschulen, Forschungs- und Wissenschaftsinstitutionen, die einen jährlichen Konsortialbeitrag leisten. Sie wird an der *Purdue University* kuratiert. Einnahmen aus dem Verkauf der Datenbasis fließen in deren Weiterentwicklung. Jede Wissenschaftlerin bzw. jeder Wissenschaftler kann Zugriff auf die GTAP-Datenbasis oder einen definierten Teil davon bekommen, wenn sie/er selbst einen qualitätsgesicherten Beitrag dazu leistet. Voraussetzung für das Funktionieren dieses Ansatzes ist, dass der Aufwand des Einzelnen für den Beitrag zur Datenbasis geringer ist als der dadurch für sie/ihn entstehende Nutzen. Mit der GTAP-Datenbasis oder dem ebenfalls zur Verfügung gestellten GTAP-Modell arbeiten mittlerweile weltweit mehr als 13 000 Wissenschaftlerinnen und Wissenschaftler, deren Kooperation über die Zeit sehr viel enger geworden ist. Es gibt jährliche Konferenzen, auf denen eine qualitätskontrollierte Auswahl der Arbeiten mit der Datenbasis oder der weiter entwickelten Modellstruktur vorgestellt werden.

Mit diesem Ansatz werden Standardwissen und -daten zur Verfügung gestellt, so dass jede Forschungsfrage auf einem höheren Einstiegsniveau begonnen sowie Daten- und methodische Modellentwicklung deutlich weiter vorangetrieben werden können. Um eine Monopolstellung oder Einflussnahme auf Forschungsrichtungen zu verhindern, wurden verschiedene Maßnahmen ergriffen. Hierzu gehören u. a. eine Erhöhung der Transparenz und der Qualitätskontrolle sowie eine verstärkte Entwicklung von Software, die einen Einsatz der gesamten oder von Teilen der Datenbasis oder deren „Verarbeitung“ und Ergänzung zu anderen Zwecken ermöglicht. | [12]

| [12] Die Anzahl der im Text des Fallbeispiels beteiligten Länder und Sektoren entspricht dem Stand im Jahr 2019. Román, A.; Narayanan, B.; McDougall, R.: *An Overview of the GTAP 9 Data Base*, in: *Journal of Global Economic Analysis,* 1 (2016) 1, S. 181–208 http://dx.doi.org/10.21642/JGEA.010103AF; https://www.gtap.agecon.purdue.edu/events/conferences/default.asp.

Fallbeispiel 3 zeigt, dass große Datenbestände, deren Generierung teils sehr kostenintensiv ist, inzwischen durch relativ kleine Forscherteams zusammengeführt werden können. Diese Aggregation von Daten aus unterschiedlichsten Quellen ermöglicht zusätzliche Nutzungen innerhalb und außerhalb der Wissenschaft, die jedoch auch ethische Herausforderungen darstellen können.

Im Projekt So2Sat entstehen hochaufgelöste vierdimensionale und global vernetzte Städtemodelle. Dafür werden unterschiedliche Datenarten mit einem Volumen im zweistelligen Petabyte-Bereich zusammengeführt, vor allem Satellitenbilder verschiedener Art (inkl. Radar und optische Sensoren von DLR, ESA | [13] und privaten Unternehmen), Bilddaten (z. B. Instagram, zur Erfassung von Gebäudefronten u. a.) und Textdaten (z. B. Twitter, zur semantischen Annotation der zeitabhängigen Nutzung von Gebäuden und Verkehrsinfrastruktur) sowie offen zugängliche Daten zum 2-D-Gebäudefußabdruck und zur Gebäudenutzung aus Geographischen Informationssystemen (GIS). Aufgrund der hohen Diversität von optischen Wellenlängen, Blickwinkeln, Sensorarten und Zeitpunkten sowie der Komplexität von Radardaten sind diese Daten hochdimensional. Zu ihrer Aufbereitung werden modellbasierte Signalverarbeitungsmethoden verwendet, zur Datenanalyse Methoden des maschinellen Lernens, insbesondere des *Deep Learning*, und zur Bewältigung der Datenflut Hochleistungsrechner eingesetzt.

So entstehen neuartige Städtemodelle für derzeit 42 Städte, die die Geographie der Stadt sowie die Auslegung und Form ihrer Bauten und Straßen mit einer Auflösung von unter einem Meter umfassen. Dazu kommen zeitabhängige semantische Informationen zu den Gebäuden und Straßen, die deren Nutzung über den Tages- und Jahreszeitraum erfassen. Solche Informationen sind eine wertvolle Grundlage für die Städteplanung, insbesondere im Hinblick auf das perspektivisch bedeutende Smart-City-Konzept. Die Technologie ist insbesondere relevant für die Planung der rasant wachsenden Megastädte der sich entwickelnden Welt. Erstmals führen die neuen Datenanalysemethoden die heterogenen Datenquellen aus Satellitenbildern und sozialen Netzen systematisch zusammen. Die im Projekt generierten Daten werden nach Abschluss frei zugänglich gemacht, ebenso die entwickelten Methoden, deren zugrunde liegenden Prinzipien auf andere Anwendungen der Datenintegration erweiterbar sein werden. Die Projektergebnisse werden vielfältige Forschung in unterschiedlichen Disziplinen über den globalen Wandel bzgl. Städteentwicklung inspirieren. Durch die Möglichkeit der Verknüpfung heterogener, teilweise personen- oder gruppenbezogener Daten ist das Projekt aber auch ein Beispiel für die Herausforderungen, dadurch entstehende Möglichkeiten der Bildung von personenbezogenen Profilen mit ethischen und rechtlichen Aspekten des Datenschutzes zu

| [13] DLR: Deutsches Zentrum für Luft- und Raumfahrt, ESA: *European Space Agency*.

vereinbaren sowie eine mögliche Benachteiligung von Bevölkerungsgruppen mittels Klassifizierung von Nutzungsprofilen zu vermeiden. Beides ist in dem Projekt nicht vorgesehen, erscheint bei der verwendeten Technologie aber unter entsprechenden Bedingungen nicht ausgeschlossen. Zu den beteiligten Disziplinen zählen Erdbeobachtung, Mathematik, Informatik, Elektrotechnik, Sozial-, Städte- und Umweltwissenschaften. Das Projekt wird mit einem Starting Grant des Europäischen Forschungsrats (ERC) finanziert. | [14]

II.4 Erstellung eines jährlichen globalen Kohlenstoff-Budgets

Fallbeispiel 4 zeigt das Assessment von Daten aus verschiedenen Datenbanken, deren Ergebnis eine hohe Informationsverdichtung mit großen politischen Implikationen ist. Das Fallbeispiel zeigt zugleich, welche neuen Herausforderungen durch datenintensive Forschung für die Publikation von Daten und Software sowie deren langfristige Verfügbarkeit entstehen.

Die quantitative Veränderung des Kohlenstoffdioxidanteils (CO_2) in der Atmosphäre und der Beitrag menschlicher Aktivitäten dazu sind Thema immer weiter verfeinerter wissenschaftlicher Untersuchungen. In diesem Zusammenhang ist es unerlässlich, einen Überblick (Budget) über Quellen und Senken von CO_2 zu erhalten. Dies sind auf menschlicher Seite etwa die Energieerzeugung, Zementproduktion und Landnutzung sowie auf natürlicher Seite Ozean, Bio- und Atmosphäre.

Mitarbeiterinnen und Mitarbeiter des *Global Carbon Project* (GCP) haben seit 2006 jährlich ein solches Budget des Eintrags von CO_2 in die Umwelt und dessen Verbleib erstellt. Das GCP als solches erfasst dabei keine Daten und erstellt keine Modelle, sondern wendet die Methode des kritischen Assessments als Bewertung und Auswahl solcher Quellen an, wie sie vom *Intergovernmental Panel on Climate Change* (IPCC) eingesetzt wird. Es ist offensichtlich, dass dazu Daten aus verschiedensten wissenschaftlichen Disziplinen wie auch Behördendaten (z. B. Wirtschaftsstatistiken), Zusammenfassungen von In-situ-Messungen, Satellitendaten und andere herangezogen und Modellvergleiche eingesetzt werden müssen. Sowohl die herangezogenen Quellen als auch deren Bewertung können sich über die Jahre verändern.

Die Ergebnisse dieser Arbeit sind z. B. in den fünften Assessment Report des IPCC von 2013 eingeflossen. Seit 2013 werden die jährlichen Budgets parallel in

| [14] Website des Projektes So2Sat – *10^{16} Bytes from Social Media to Earth Observation Satellites*: http://www.so2sat.eu; *TEDx Talk*: https://www.youtube.com/watch?v=VgluoTngkQE; wichtigste Publikationen: Kang, J. et al.: *Building instance classification using street view images*, in: *ISPRS Journal of Photogrammetry and Remote Sensing*, 145, *Part* A (2018), S. 44–59, https://doi.org/10.1016/j.isprsjprs.2018.02.006; Zhu, X. et al.: *Deep Learning in Remote Sensing. A Comprehensive Review and List of Resources*, in: *IEEE Geoscience and Remote Sensing Magazine*, 5 (2017) 4, S. 8–36, DOI: 10.1109/MGRS.2017.2762307; Zhu, X. et al.: *Geodetic SAR Tomography*, in: *IEEE Transactions on Geoscience and Remote Sensing*, 54 (2016)1, S. 18–35, DOI: 10.1109/TGRS.2015.2448686.

Form eines kurzen Artikels in einer Ausgabe der Zeitschrift *Nature* und eines Datenartikels bei *Earth Systems Science Data* (ESSD) veröffentlicht.

ESSD verlangt, dass der Datensatz selbst in einem verlässlichen Daten-Repositorium aufbewahrt wird und mit einem persistenten *Identifier* (vorzugsweise einem *Digital Object Identifier* – DOI) versehen ist. Angesichts der Bedeutung benutzter Software – der Algorithmen wie auch der konkreten Codes – wäre hier der inhaltlich naheliegende nächste Schritt, deren Bereitstellung in verlässlichen Repositorien (mit stabilen *Identifiers*) ebenfalls zu verlangen. Allerdings erweist es sich schon heute mit zwei beteiligten Journalen unterschiedlicher Verlage und einem Daten-Repositorium als schwierig, das korrekte gegenseitige Zitieren der publizierten Objekte zum Zeitpunkt der Publikation (alle Objekte gleichzeitig oder nacheinander?) sicherzustellen. Dieses technisch-organisatorische Problem würde durch das Hinzufügen eines oder mehrerer Software-Repositorien und womöglich assoziierter Artikel in Software-Journalen nochmals deutlich verschärft, ebenso wie der Aufwand deutlich zunähme, alle diese Komponenten, die eigentlich zur Veröffentlichung einer einzigen Erkenntnis dienen, auch dauerhaft zugänglich zu erhalten. | [15]

II.5 Systemmedizinische Ansätze zur Untersuchung von Lungenkrebs

Fallbeispiel 5 steht für die praktischen Herausforderungen von Forschung an der Schnittstelle von Medizin und Bioinformatik. Notwendig sind hier die Entwicklung einer disziplinenübergreifend einheitlichen Sprache und institutionenübergreifender Workflows sowie professionelle Beratung zu umfangreichen und sich dynamisch entwickelnden rechtlichen Vorschriften.

Für den medizinischen Fortschritt bieten interdisziplinäre Forschungsvorhaben, in denen große Datenmengen gesammelt und genutzt werden, enorme Chancen. Lungenkrebs, die Krebsart mit der höchsten Mortalitätsrate weltweit, wird in den meisten Fällen erst spät diagnostiziert. Bislang fehlen dann Möglichkeiten, sicher vorherzusagen, welche Patientinnen und Patienten von der sehr kostenintensiven und mit Nebenwirkungen behafteten Immuntherapie profitieren und welche Therapieform für die anderen Patienten geeignet wäre. Hierfür relevante Daten umfassen typischerweise molekulare Beobachtungen wie Sequenzdaten, mit verschiedenen Methoden erzeugte Bilder des Organs sowie klinische Parameter und Verläufe. Krebsentstehung, Krankheitsprogression und Therapieansprechen sind sehr dynamische Prozesse, die von einer Vielzahl von

| [15] Webseite des GCP: https://www.globalcarbonproject.org/about/; Nutzung des CO_2-Budgets durch IPCC: Ciais, P. et al.: *Chapter 6: Carbon and Other Biogeochemical Cycles*, in: IPCC: *Climate Change 2013. The Physical Science Basis*, hrsg. v. Stocker, Th. F. et al., Cambridge/New York, NY 2013; GCP-Veröffentlichungen zum Budget 2018: Figueres, Ch. et al.: *Emissions are still rising: ramp up the cuts*, in: *Nature*, 564 (2018) 27, S. 27–30, https://doi.org/10.1038/d41586-018-07585-6; Le Quéré, C. et al.: *Global Carbon Budget 2018*, *Earth System Science Data*, 10 (2018), S. 2141–2194, https://doi.org/10.5194/essd-10-2141-2018; der Datensatz selbst: *Global Carbon Project*. *Supplemental data of Global Carbon Budget* 2018 (Version 1.0) [*Data set*]. https://doi.org/10.18160/gcp-2018.

Faktoren bestimmt werden. Auf Grund der Komplexität der Zusammenhänge bietet die datenbasierte mathematische Modellierung Möglichkeiten, molekulare Mechanismen zu entschlüsseln und quantitative Vorhersagen bezüglich des Zellwachstums und schlussendlich des Therapieerfolges zu machen.

Im Verbundforschungsvorhaben „LungSys – Systembiologie von Lungenkrebs" waren sowohl Grundlagenforschende aus der Zellbiologie, Molekularbiologie, Biochemie, Physik, Informatik als auch Forschende aus Onkologie, Radiologie, Chirurgie, Pathologie, sowie weitere Expertinnen und Experten aus Kliniken, Biobanken und Datenmanagement beteiligt. Für die später u. a. im Rahmen des Deutschen Zentrums für Lungenforschung (DZL) fortgesetzten Forschungen galt es, eine gemeinsame Sprache zu etablieren und ein gemeinsames Verständnis für Datenerhebung, Datenfluss und Datenspeicherung zu entwickeln. So mussten für die Zusammenführung von Daten aus verschiedenen Kliniken dokumentierte Parameter harmonisiert und Schnittstellenkompatibilität sichergestellt werden. Plasma- und Gewebeproben der Patientinnen und Patienten werden in einer zentralen Biobank gelagert und können, sofern die Einwilligung der Betroffenen vorliegt und ein Ethikantrag genehmigt wurde, für Forschungszwecke untersucht werden. Um Untersuchungen an einer anderen Institution durchführen zu können, wird wiederum die Zustimmung der lokalen Ethikkommission benötigt. Nach Inkrafttreten der EU-Datenschutzgrundverordnung (DSGVO) muss für den Umgang mit personenbezogenen Daten der Patientinnen und Patienten inzwischen zudem ein Datenschutzkonzept vorgelegt werden, das eine Folgenabschätzung sowie eine Bewertung und Dokumentation der technischen und organisatorischen Maßnahmen enthalten muss. Zu weiteren Projektvorbereitungen gehören vertragliche Vereinbarungen für die Forschungskonsortien, den Transfer von Proben und Daten, so dass in der Summe vielfältige Voraussetzungen zu erfüllen sind für interdisziplinäre Forschungsprojekte mit Patientendaten. | [16]

II.6 Maschinelles Lernen in den Ingenieurwissenschaften

Fallbeispiel 6 zeigt, wie sich die Digitalisierung von Prozess- und Wertschöpfungsketten in der Zusammenarbeit von Ingenieurwissenschaften und Industrie etabliert hat und sie mit Hilfe von Methoden des maschinellen Lernens deutlich effizientere Produktentwicklungen sowie schnellere Ergebniserreichung ermöglichen kann.

In den Ingenieurwissenschaften konzentrieren sich viele Forschungsaktivitäten auf die Weiterentwicklung digitaler Lösungsansätze sowohl in der vertikalen

| [16] Becker, V. et al.: *Covering a broad dynamic range. Information processing at the erythropoietin receptor*, in: *Science*, 328 (2010) 5984, S. 1404–1408; Yin, Y. et al.: *Tumor Cell Load and Heterogeneity Estimation From Diffusion-Weighted MRI Calibrated With Histological Data. An Example From Lung Cancer*, in: IEEE *Trans Med Imaging*, 37 (2018) 1, S. 35–46; Angelidis, I. et al.: *An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics*, in: *Nat Commun*, 2019, Feb., S. 1–17.

Vernetzung der industriellen Steuerungsarchitekturen über alle Prozessebenen einer modernen industriellen Produktion hinweg als auch in der horizontalen Vernetzung der Partner entlang der Wertschöpfungskette. Eine wesentliche Rolle nehmen dabei auch die digitalen *Engineering*-Werkzeuge im gesamten Lebenszyklus von Produktionsanlagen und Produkten ein. Dies betrifft Projektierung, Entwicklung, Inbetriebnahme und Betrieb von Anlagen und Systemen bis hin zur Wiederverwendung und dem Recycling im Rahmen der Kreislaufwirtschaft.

Im Kontext der Digitalisierung industrieller Wertschöpfungsprozesse (Industrie 4.0) nimmt die Diskussion um den Einsatz datenbasierter Lösungen und *big data analytics* einen zunehmend breiten Raum ein und ist Gegenstand umfangreicher Forschungsanstrengungen in nahezu allen Bereichen der Ingenieurwissenschaften. Eine leistungsfähige Dateninfrastruktur mit hoher Datensicherheit und Datenverfügbarkeit sowie auch die Interoperabilität und Echtzeitfähigkeit von Datensystemen ist funktionskritischer Bestandteil einer nachhaltigen souveränen Dateninfrastruktur für unser zukünftiges Wirtschaftswesen. | [17]

Maschinelles Lernen (ML) auf Maschinendaten ist zum Beispiel im Bereich der erneuerbaren Energien für die Überwachung des Zustands von Windkraftanlagen ein wichtiges Thema geworden. Hier werden inzwischen auf der Basis riesiger Mengen von Sensor-Daten Methoden des Maschinellen Lernens eingesetzt und ständig weiterentwickelt, um Betrieb und Wartung – beides kritische Kostenfragen – zu optimieren. Im breiteren Bereich der Produktion wird ML für die Diagnose und Optimierung mit enormen Fortschritten und positiven Ergebnissen eingesetzt, z. B. mit dem Ergebnis höherer Flexibilität.

Daneben ist als weiteres Beispiel die Modellierung turbulenter Strömungen eine Aufgabe, die in einem extrem breiten Anwendungsspektrum von Flugzeugen bis hin zu Gasturbinen, vom Blasensäulenreaktor bis zur Strömung in einer Herzklappe reicht. Daher hat die Fähigkeit, solche Strömungen im Detail zu simulieren, enorme Auswirkungen auf das Verständnis und die wissenschaftliche Fortentwicklung, aber auch auf den wirtschaftlichen Erfolg vielfältiger Anwendungsfelder. Die Turbulenz wird modelliert in komplexen Modellen, die hochauflösende Datenfelder erzeugen. ML-Techniken werden inzwischen vielfach verwendet, um solche Modelle unter Verwendung von High-Fidelity-Daten aus prototypischen Simulationen bereitzustellen. Die jüngsten Erfolge erlauben sogar eine Quantifizierung der damit verbundenen Unsicherheiten. Die Kombina-

| 17 Lenz, J.; Wuest, Th.; Westkämper, E.: *Holistic approach to machine tool data analytics*, in: *Journal of Manufacturing Systems*, 48*, Part C* (2018), S. 180–191; Madni, A. M.; Sievers, M.: *Model-Based Systems Engineering. Motivation, Current Status, and Needed Advances*, in: Madni, A. M. et al. (Hrsg.): *Disciplinary Convergence in Systems Engineering Research*, Cham 2018*; ManuFUTURE High-Level Group*: *ManuFUTURE Vision 2030. Competitive, Sustainable and Resilient European Manufacturing, Brussels* 2018 (http://www.manufuture.org/wp-content/uploads/Manufuture-Vision-2030_DIGITAL.pdf); Zhong, R. Y. et al.: *Big Data for supply chain management in the service and manufacturing sectors. Challenges, opportunities, and future perspectives*, in: *Computers & Industrial Engineering*, 101 (2016), S. 572–591.

tion von physikalischen Gesetzen, großen Datenmengen und ML birgt insgesamt ein hohes Potenzial für ein tieferes Verständnis und bessere Modelle, was zu besseren Werkzeugen zur Optimierung und Kontrolle der Fluidströmungen führt. |[18]

II.7 *Distant Reading* und Vergleich von Heftromanen

Fallbeispiel 7 zeigt, wie neu entstandene digitale Textsammlungen in Bibliotheken datenintensive Forschung an der Schnittstelle von Computerphilologie und Computerlinguistik ermöglichen. Methoden aus der Informatik werden dort für Analysen und Vergleiche großer Textmengen wenig beforschter Gattungen nutzbar gemacht, um ältere Forschungsergebnisse zu überprüfen und neuen Forschungsfragen nachzugehen.

Heftromane wie Perry Rhodan oder John Sinclair sind kein klassischer Gegenstand der Literaturwissenschaft, die sich vielfach der Beforschung von einzelnen Texten der sogenannten Hochliteratur widmet. Heftromane wurden traditionell definiert durch das kompakte Publikationsformat, eigene Formen der Distribution über den Zeitschriftenmarkt statt über den Buchhandel und auch die angenommene Zusammensetzung der Heftromanlesenden. Durch die zunehmende digitale Veröffentlichung von Heftromanen (E-Books) und die Hinzunahme digitaler Objekte in öffentliche Bibliotheken sind entsprechend umfangreiche Sammlungen entstanden. Dieses neue Angebot wird im Projekt zur makroanalytischen Untersuchung von Heftromanen an der Schnittstelle von Computerphilologie und Computerlinguistik genutzt, um mit Hilfe verschiedener Methoden zu verstehen, wie sich Gattungen der Heftromane (z. B. Science-Fiction, Horror, Romantik, Western) untereinander und im Vergleich zur Hochliteratur unterscheiden. Dazu erfolgt eine Kooperation mit der Gedächtnisinstitution Bibliothek, hier der Deutschen Nationalbibliothek. Die beteiligten Forschenden nutzen die rasch zunehmende Verfügbarkeit großer Textsammlungen für die Literaturwissenschaft sowie neuer Methoden, die vorrangig aus der Informatik

|[18] Stetco, A. et al.: *Machine learning methods for wind turbine condition monitoring. A review,* in: *Renewable Energy*, 133 (2019), S. 620–635, https://doi.org/10.1016/j.renene.2018.10.047; Weichert, D. et al.: *A review of machine learning for the optimization of production processes,* in: *International Journal of Advanced Manufacturing Technology*, 104 (2019), S. 1889–1902; Lenz, J.; Wuest, Th.; Westkämper, E.: *Holistic approach to machine tool data analytics*, in: *Journal of Manufacturing Systems*, 48, *Part C* (2018), S. 180–191. Zhong, R. Y. et al.: *Big Data for supply chain management in the service and manufacturing sectors. Challenges, opportunities, and future perspectives*, in: *Computers & Industrial Engineering,* 101 (2016), S. 572–591. Ma, M.; Lu, J.; Tryggvason, G.: *Using statistical learning to close two-fluid multiphase flow equations for a simple bubbly system*, in: *Physics of Fluids*, 27, 092101 (2015) 9, DOI: 10.1063/1.4930004; Duraisamy, K.; Iaccarino, G.; Xiao, H.: *Turbulence Modeling in the Age of Data,* in: *Annual Review of Fluid Mechanics*, 51 (2019), S. 357–377, DOI: 10.1146/annurev-fluid-010518-040547; Xiao, H.; Cinnella, P.: *Quantification of model uncertainty in RANS simulations: A review,* in: *Progress in Aerospace Sciences*, 108 (2019) Juli, S. 1–31, DOI: 10.1016/j.paerosci.2018.10.001; Brunton, S. L.; Noac, B. R.; Koumoutsakos, P.: *Machine Learning for Fluid Mechanics,* in: *Annual Review of Fluid Mechanics*, 52 (2020), *expected* Jan 2020, DOI: 10.1146/annurev-fluid-010719-060214.

entstammen und für die Literaturwissenschaft nutzbar gemacht, angepasst und evaluiert werden. | [19]

Das zunächst explorative Projekt zur makroanalytischen Untersuchung von Heftromanen wertet 9 000 deutschsprachige Heftromane aus dem Zeitraum 2009 bis 2017 aus, um zunächst statistische Aussagen über die gesammelten Texte zu treffen (*Distant Reading* statt *Close Reading*). Um die Kohärenz der Gattungen zu untersuchen, werden die häufigsten Substantive und deren Leistungsfähigkeit für die Klassifikation der Texte auch mit Hilfe von überwachtem Lernen untersucht und visuell aufbereitet. Dabei zeigt sich, dass Heftromane inhaltlich und stilistisch recht deutlich abgegrenzte Gattungen bilden. Texte werden mit verschiedenen Methoden der Inhaltsanalyse untersucht, um Themen, Gegenstände, Figuren und wiederkehrende Formulierungen zu identifizieren. Schließlich wird in einem dritten Teil des Projektes die Gattungskomplexität und der Kontrast zur Hochliteratur untersucht, indem die Variabilität der Sprache und die Größe des Wortschatzes wie auch die Satzbaukomplexität in den Blick genommen werden. Vorläufige Ergebnisse des Projektes zeigen, dass die Gattungen der Heftromane deutlich umrissen und inhaltlich gut erschließbar sind. Zwei Thesen werden bereits als widerlegt betrachtet: Heftromane sind weniger homogen als in älterer Forschung behauptet und zeigen eine deutliche Binnenvarianz. Außerdem ist die Sprache der Heftromane nicht eindeutig schlichter und insbesondere die Science-Fiction-Romane weichen hier deutlich ab. Weitere Untersuchungen sind geplant zur Figurenkonstellation und zur Analyse von Erzähler- und Figurenrede sowohl innerhalb der Heftromane als auch im Vergleich zur sogenannten Hochliteratur.

A.III DIMENSIONEN DES WANDELS

Das Wissenschaftssystem hat im digitalen Zeitalter Anteil an einem umfassenden Transformationsprozess. Es bestimmt dessen Fortgang durch Forschung mit, wird aber auch in vielerlei Hinsicht von diesen Entwicklungen selbst beeinflusst. So verändern sich wissenschaftliche Informationsbeschaffung und -verarbeitung ebenso dynamisch wie – massiv verstärkt durch die Kontakt- und Mobilitätsbeschränkungen während der COVID-19-Pandemie – die Kommunikation von Wissenschaftlerinnen und Wissenschaftlern untereinander bis hin zur Verbreitung ihrer Forschungsergebnisse. Die Anfänge dieser Entwicklung reichen Jahrzehnte zurück, doch weist in den vergangenen Jahren vieles auf eine deutliche Beschleunigung des Wandels hin.

| [19] Jannidis, F.; Konle, L.; Leinen, P.: Makroanalytische Untersuchung von Heftromanen, in: Digital Humanities: multimedial & multimodal. Konferenzabstracts zur DHd-Jahrestagung 2019, S. 167–172, https://doi.org/10.5281/zenodo.2596094.

Im Folgenden werden Beobachtungen zu datenintensiver Forschung im digitalen Zeitalter nach sechs Themenbereichen zusammengetragen, in denen Auswirkungen auf die Organisation und Struktur des Wissenschaftssystems festzustellen sind. Dabei gilt es stets zu berücksichtigen, dass Wissenschaft die Bedingungen, unter denen datenintensive Forschung betrieben wird, nicht vollständig autonom bestimmen kann. Zahlreiche Akteure innerhalb wie außerhalb des Wissenschaftssystems nehmen Einfluss auf den gegenwärtigen Transformationsprozess und Kulturwandel – national und international. Die Entwicklung und Verbreitung datenintensiver Forschung wird von den einzelnen Fachgemeinschaften und Institutionen der Wissenschaft geprägt, aber auch von außen durch Ressourcenbereitstellung und Rechtsrahmen sowie durch Erwartungen, die von verschiedenen gesellschaftlichen Akteuren an Leistungen und Nachvollziehbarkeit von Wissenschaft gerichtet werden.

III.1 Datenverfügbarkeit und Datenhaltung

Die Verfügbarkeit von Daten für die Forschung an Hochschulen und Forschungsinstituten hat sich in den vergangenen Jahren dynamisch entwickelt. Das Spektrum von Datenarten, Datenvolumina und Datenraten hat sich enorm vergrößert, während zugleich Speicherung, Auffindbarkeit und Austausch der Daten erleichtert wurden. |[20] Bereits vorhandene analoge Daten- und Objektebestände aus wissenschaftlichen Sammlungen und Archiven werden für die wissenschaftliche Nutzung nachträglich digitalisiert, in neu aufgebauten digitalen Sammlungen aufbereitet und über Plattformen häufig auch öffentlich zugänglich gemacht (vgl. Fallbeispiel II.7). Andere Datenquellen – etwa aus der öffentlichen Verwaltung oder von Bürgerwissenschaftsprojekten (Citizen Science) – werden durch die Wissenschaft systematischer erschlossen. Weitere neuartige Datenbestände aus Unternehmen – beispielsweise Daten aus der Nutzung von mobilen Endgeräten, Social-Media-Plattformen oder auch Produktionsprozessen – werden für die Wissenschaft erstmals verfügbar, können teils jedoch nur eingeschränkt genutzt werden oder unterliegen Beschränkungen bei der Veröffentlichung. Viele zusätzliche Daten bringt die Verbreitung weiterentwickelter oder neuer Messtechnik und Sensorik, die häufig vollautomatisiert eingesetzt wird und Daten in Echtzeit zur Verfügung stellen kann (vgl. Fallbeispiel II.6). Nicht all diese Daten sind jedoch für die Wissenschaft leicht verfügbar – aus verschiedenen rechtlichen und technischen Gründen oder weil sie sich in der Hand privater Unternehmen befinden. |[21]

|[20] Vgl. etwa die Übersicht von Datenbanken und Forschungsdaten-Repositorien verschiedener wissenschaftlicher Fachgebiete unter www.re3data.org.

|[21] Vgl. Rat für Sozial- und Wirtschaftsdaten: Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften. Datenzugang und Forschungsdatenmanagement, RatSWD Output 4 (6), Berlin (2019), https://doi.org/10.17620/02671.39.

Fachgemeinschaften haben sich datenintensiver Forschung mit digital verfügbaren Daten zu verschiedenen Zeitpunkten zugewandt und blicken auf unterschiedlich lange Traditionen zurück. Schon Ende der 1950er Jahren haben etwa die Erd- und Umweltwissenschaften begonnen, sich über Standards der Erhebung und Speicherung geophysikalischer Daten in sogenannten *World Data Centers* zu verständigen (heute *ISC World Data System*). Auch Teile der Sozialwissenschaften haben in Deutschland bereits Anfang der 1960er Jahre ein „Zentralarchiv" für die Aufbereitung und Archivierung bestimmter sozialwissenschaftlicher Umfragedaten geschaffen (heute GESIS – Leibniz-Institut für Sozialwissenschaften). Andere Fachgebiete haben deutlich später begonnen, gemeinsame Datensammlungen in fachbezogenen oder fachübergreifenden, nationalen oder internationalen Strukturen und mit unterschiedlichen Nutzungsbedingungen und Zugangsmodellen aufzubauen. Im Gesundheitsbereich sind inzwischen die NAKO Gesundheitsstudie zur Erforschung bestimmter Volkskrankheiten oder auch die übergreifende Medizininformatik-Initiative zum standortübergreifenden Austausch klinischer Daten auf den Weg gebracht worden. International sind etwa das *Human Epigenome Project*, das Datensammlung und -konsolidierung im Bereich der zellulären Regulation vorantreibt, oder der *Human Cell Atlas* zu nennen, der alle menschlichen Zelltypen kartieren und charakterisieren will. Insgesamt stellen in Fachgemeinschaften gemeinsam aufgebaute, gepflegte und kontinuierlich ausgebaute digitale Datensammlungen bislang keineswegs den Regelfall dar (vgl. Fallbeispiel II.2). Wo sie entstehen und gepflegt werden, sind diese Bemühungen oft verbunden mit fächerbezogenen Verständigungen über Standards datenintensiver Forschung – beispielsweise zu Metadaten, Speicherformaten, -menge und -dauer sowie über Zugangsregimes zu Daten. Solche Standardisierungsinitiativen laufen häufig international ab. | [22]

Die dynamische Entwicklung der Datenverfügbarkeit führt dazu, dass die Handhabung der Datenströme an vielen Stellen im Wissenschaftssystem angepasst oder – abhängig von der Tradition datenintensiver Forschung im jeweiligen Fachgebiet und den jeweils vorrangigen Forschungsformen – weitgehend neu geregelt und stabil finanziert werden muss. Um für die Forschung genutzt werden zu können, müssen die verfügbaren Daten einschließlich reichhaltiger und standardisierter Metadaten gespeichert, aufbereitet, qualitätsgesichert, auffindbar, nutzbar und zusammenführbar gemacht werden. Zahlreiche unterschiedliche Arbeitsschritte sind hierfür erforderlich, die teilweise unter dem Begriff Datenkuratierung zusammengefasst werden. Sie werden von dedizierten Stellen innerhalb des Wissenschaftssystems oft unter Zuhilfenahme weiterer externer Dienstleister übernommen, erfordern häufig aber noch aufwändige Mitwirkung der jeweiligen Wissenschaftlerinnen und Wissenschaftler. Weitere anspruchsvolle Herausforderungen sind die Datenauswahl bei schnell anfallenden Massendaten durch Lösch- und Prüfkonzepte, die Speicherung einschließlich der Siche-

| 22 Etwa die *Global Alliance for Genomes and Health*, https://www.ga4gh.org.

rung der Datenintegrität sowie das Versionsmanagement. Hieran schließen sich wiederum Fragen der Bereitstellung an, die nicht nur das Suchen und Finden von Datenbeständen, sondern auch deren Transport und den gestuften Zugang zu ihnen mit einem entsprechenden Rechtemanagement beinhalten. Weitere Fragen wirft der Einsatz spezieller Software für die Erhebung und Nutzung von Daten auf. Hierzu gehört die langfristige Zugänglichkeit bzw. Lesbarkeit von Daten, die nur mit Hilfe spezieller und häufig wissenschaftsextern von Unternehmen entwickelter Software erhoben, dargestellt und analysiert werden können (vgl. Fallbeispiel II.4).

Eine Schlüsselrolle für die Verfügbarkeit und Nutzung von qualitätsgesicherten Datenbeständen für die Forschung spielen Informationsinfrastrukturen von wissenschaftlichen Einrichtungen wie Bibliotheken bzw. Zentren für Kommunikation und Information (Rechenzentren), von und für Fachgemeinschaften betriebene Repositorien zur Aufbewahrung und Recherche sowie spezielle Datensammlungen. Eine nachhaltige Sicherung der Daten über lange Zeiträume erfordert passende Infrastrukturen und institutionelle Strukturen und kann nicht bspw. durch die Auslagerung an Repositorien von Verlagen garantiert werden. Entsprechende Infrastrukturen sind vielfach im Entstehen, wie die Änderungen bei der Organisation zentraler Einrichtungen an Hochschulen und Forschungseinrichtungen, Ergänzungen um neue lokale Einheiten zur Datenanalyse sowie der Aufbau der Nationalen Forschungsdateninfrastruktur (NFDI), der *European Open Science Cloud* (EOSC) sowie mit GAIA-X ein über die Wissenschaft hinausweisendes Projekt zur Implementierung eines offenen digitalen Ökosystems zeigen. |[23] Zu erwarten ist, dass große, über Disziplinengrenzen hinausreichende Datensammlungen und Infrastrukturen generell eine Anziehungskraft für datenintensive Forschung entfalten werden. Allerdings können entsprechend hohe Investitionen auch zur Weiterverfolgung von einmal eingeschlagenen Wegen führen, selbst wenn Alternativen sich später als überlegen herausstellen sollten („Pfadabhängigkeiten“).

III.2 Fragestellungen und Analysemethoden

Die Faszination durch und Aufmerksamkeit für datenintensive Forschung hängt neben der veränderten Verfügbarkeit von Daten wesentlich mit der Weiterentwicklung von Analysemethoden sowie Hardware und Software-Werkzeugen zu-

|[23] Im Rahmen des Projektes GAIA-X entwickeln Vertreterinnen und Vertreter aus Politik, Wirtschaft und Wissenschaft gemeinsam eine vernetzte und offene Dateninfrastruktur auf Basis europäischer Werte. Ursprünglich als deutsch-französische Initiative gestartet, beteiligen sich mittlerweile Vertreter aus sieben europäischen Ländern an der Projektentwicklung. Ziel ist es, „eine sichere und vernetzte Dateninfrastruktur, die den höchsten Ansprüchen an digitale Souveränität genügt und Innovationen fördert“, zu schaffen (vgl. https://www.bmwi.de/Redaktion/DE/Dossier/gaia-x.html). Zum aktuellen Stand des sich dynamisch entwickelnden Projekts vgl. https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html).Als Übersicht zu Datenpools zuletzt Heumann, St.; Jentzsch, N.: Wettbewerb um Daten. Über Datenpools zu Innovationen, hrsg. v. Stiftung Neue Verantwortung, Berlin 2019, https://www.stiftung-nv.de/sites/default/files/wettbewerb_um_daten.pdf.

sammen. So treten neben die Datenanalyse mit klassischen statistischen Methoden zunehmend Verfahren wie etwa maschinelles Lernen, bildgebende Methoden (*visual analytics*) oder kombinatorische Algorithmen. Mit ihrer Hilfe wird es möglich, sehr große und komplexe Datenbestände zu analysieren und in ihnen Muster zu erkennen. Diese Analysemethoden können für die Bearbeitung alter und neuer, strukturierter und nicht-strukturierter, homogener und heterogener sowie völlig neu kombinierter und zusammengeführter Datenbestände eingesetzt werden. Sie können Dinge sichtbar machen, die vorher für die Forschung nicht sichtbar waren, indem auch hochdimensionale Datenbestände für Analysen verknüpft werden, bei denen dies bisher nicht möglich war (vgl. Fallbeispiele II.3 und II.5). Datengeleitete Fragestellungen ergänzen damit inzwischen theoriegeleitete und wissensbasierte Forschung, was auch als „Empirifizierung" in der Wissenschaft beschrieben wird und eine neue Herangehensweise an Datenbestände beinhalten kann, bei der Hypothesen ex ante nicht mehr explizit gemacht werden. Grundsätzliche, strittige Fragen sind in diesem Zusammenhang, ob der hohe Vorhersageerfolg statistischer Modelle es erlaubt, auf die Unterscheidung von Korrelation und Kausalität zu verzichten, oder, falls nicht, ob es möglich ist, kausale Hypothesen (teil-)automatisiert zu erzeugen. |[24]

Ein Beispiel für neue Fragestellungen ergebnisoffener Forschung ist die vergleichende Genomik, die durch die Verfügbarkeit einer Vielzahl vollständiger Genomdaten möglich geworden ist. Auf der anderen Seite führt die Definition eines sehr konkreten Ziels (vgl. Fallbeispiel II.1 zu automatischer Übersetzung) oft zu der fokussierten Verfolgung der benötigten neuartigen Methodenforschung und zu großen Durchbrüchen. Dieser anwendungsorientierte Ansatz kann insbesondere dann hilfreich sein, wenn hochkomplexe Forschungsansätze unter Beteiligung ganz unterschiedlicher Disziplinen und unter Verwendung riesiger Datenmengen zum Erfolg geführt werden sollen. Datenintensive Forschung wird allerdings zu „datengetriebener" Forschung, wenn die Verfügbarkeit von neuen Datenbeständen und Analysemethoden zum alleinigen Treiber der Forschungsprogrammatik wird oder Forschungsgebiete dominiert. Problematisch sind außerdem kurzlebige Entwicklungen, die dazu führen, dass Daten nicht mehr ausreichend analysiert werden oder ihre Nachnutzung unter dem Erneuerungszyklus der Messtechnik oder mangelhaften Metadaten leidet.

Die Erweiterungen des Methodenspektrums werden in den Fachgemeinschaften je nach Bedarf unterschiedlich aufgegriffen (vgl. Kapitel A.II). Der unterschiedliche Bedarf an neuen Methoden und Fragestellungen kann aber auch innerhalb einzelner Disziplinen zu Frontbildungen führen, wie Auseinandersetzungen zwischen Vertretern qualitativer und quantitativer Perspektiven in den Sozial-

| 24 Weiter häufig diskutiert hierzu der Beitrag von Anderson, Ch.: *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, in: WIRED, 23.06.2008, https://www.wired.com/2008/06/pb-theory/. Vgl. Peters, J.; Janzing, D.; Schölkopf, B.: *Elements of Causal Inference. Foundations and Learning Algorithms*. *MIT Press*, Cambridge/MA, 2017.

oder Geschichtswissenschaften oder von Vertretern der „klassischen“ Geisteswissenschaften mit jenen der Digital Humanities zeigen.

Insgesamt ist die Verwendung neuer Analysemethoden und entsprechender Softwarepakete bislang voraussetzungsvoll und erfordert die Einbindung von entsprechend ausgebildeten Personen sowie teils auch die Verfügbarkeit geeigneter Trainings- und Testdaten für lernende Algorithmen. Eine kritische Auseinandersetzung mit neuen Methoden ist für ihre Verbesserung unverzichtbar. Die Archivierung von Software ist ebenfalls eine große Herausforderung, die wegen Versionswechseln, unzureichender Dokumentation und der Abhängigkeit von bestimmten Betriebssystemen vielfach noch ungelöst ist. Auf diese und weitere Herausforderungen eines Softwareeinsatzes bei datenintensiver Forschung, der wissenschaftlichen Anforderungen und Qualitätsansprüchen genügen können muss, richtete sich zuletzt zunehmend Aufmerksamkeit, so dass fachübergreifend eine wachsende Nachfrage nach Austausch über Forschungssoftwareentwicklung und der Förderung von mehr wissenschaftsspezifischer Open-Source-Software zu beobachten ist. | [25]

III.3 Forschungsalltag und Wissenschaftskultur

Datenintensive Forschung verändert den Forschungsalltag von Wissenschaftlerinnen und Wissenschaftlern auch aus individueller Perspektive. Um mit neuen Datenarten und -mengen sowie Analysemethoden umgehen zu können, werden zusätzliche Kompetenzen benötigt. Forschungsdatenmanagement muss geplant und betrieben sowie vielfältige Aspekte der Qualitätssicherung berücksichtigt werden, die sich aus dem kritischen Umgang mit neuen oder fachgebietsfremden Daten und Methoden ergeben, wie etwa Fehlinterpretationen von Daten, Umgang mit manipulierten oder illegal akquirierten Daten, Fehlanwendungen von Methoden, Aufdeckung von Rechenfehlern oder Reaktionen auf nicht erklärbare Ergebnisse. Die Reproduktion bestimmter Ergebnisse mit anderen Daten oder Methoden muss möglich sein.

Künftige Wissenschaftlerinnen und Wissenschaftler benötigen zusätzliche Kompetenzen aus überarbeiteten oder neu geschaffenen Curricula und bereits aktive haben Fort- und Weiterbildungsbedarf. | [26] Obwohl beiden Gruppen dafür inzwi-

| [25] Hierzu etwa Vermeir, K. et al.: *Global Access to Research Software: The Forgotten Pillar of Open Science Implementation. Global Young Academy*, Halle/Saale 2018, https://globalyoungacademy.net/wp-content/uploads/2018/03/18013_GYA_Report_GARS-Web.pdf, sowie 2019 Gründung eines Vereins und einer zugehörigen Konferenz für Forschungssoftwareentwickler in Deutschland nach angelsächsischem Vorbild, siehe https://www.de-rse.org/de/conf2019/index.html. Zuletzt auch: Davenport, J. H.; Grant, J.; Jones, C. M.: *Data Without Software Are Just Numbers*, in: *Data Science Journal*, 19 (2020) 1, 3, S. 1–6. DOI: https://doi.org/10.5334/dsj-2020-003.

| [26] Rat für Informationsinfrastrukturen: Digitale Kompetenzen – dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft, Göttingen 2019, http://www.rfii.de/?p=3883.

schen erste Angebote aus Hochschulen und Forschungseinrichtungen sowie neue institutionenübergreifende Initiativen zur Verfügung stehen, reichen diese noch nicht aus. |[27] Nicht alle neuen Aufgaben zur Vorbereitung und Durchführung datenintensiver Forschung können außerdem von Wissenschaftlerinnen und Wissenschaftlern selbst mit den ihnen bisher zur Verfügung stehenden Ressourcen übernommen werden. Vielfältige Beratungs- und Unterstützungsstrukturen werden daher heute schon in Anspruch genommen, sind aber längst noch nicht flächendeckend verfügbar. So wandelt sich auch das Verhältnis zwischen Forschenden und Infrastrukturanbietern, die immer enger kooperieren müssen. Viele technische und datenbezogene Services sind weniger fachspezifisch als in Fachdisziplinen angenommen (z. B. digitale Laborbücher oder Online-Informationsplattformen), weshalb Wissenschaftlerinnen und Wissenschaftler immer öfter auf Infrastrukturexpertinnen und -experten zurückgreifen können, beide Seiten sich häufig aber auf unbekanntes Terrain wagen und die Neuaufteilung eigener Ressourcen prüfen müssen.

Eine neue Arbeitsteiligkeit datenintensiver Forschung zeigt sich schließlich nicht nur in der Ausdifferenzierung neuer und eigenständiger Aufgabenprofile im Forschungsprozess, sondern auch in der Zusammenarbeit innerhalb von Fachgebieten und zwischen diesen. Da Methoden und Daten stärker fachübergreifend genutzt werden können, erhöht sich der Bedarf an interdisziplinärem Austausch hierzu. Dies zeigt die Gründung der 2013 als internationales Netzwerk formierten *Research Data Alliance* (RDA) oder der seit 2016 verfolgten FAIR-Initiative (FAIR für *findability*, *accessibility*, *interoperability*, *reusability*), die bessere Voraussetzungen für den Austausch, die Nutzung und Nachnutzbarkeit von maschinenlesbaren Daten schaffen wollen und Expertise aus Fachgemeinschaften mit jener aus Infrastrukturen verbinden. |[28] Die Notwendigkeit für solche Foren und Abstimmungen nimmt mit einer Nutzung von Daten zu, die Grenzen von Fachkulturen, Institutionen, Technologien, Ländern und Rechtsräumen überschreitet.

Jenseits akuter Qualifizierungs- und Ressourcenfragen wirft der durch datenintensive Forschung ausgelöste Transformationsprozess auch Fragen nach einer Kultur des Teilens im Wissenschaftssystem auf. Diese Fragen beziehen sich sowohl auf die Anwendung und Weiterentwicklung von facheigenen und -fremden Methoden als auch auf die Erhebung, Pflege und Nutzung von Daten ein-

| 27 Etwa Workshops zu Daten und Software von *The Carpentries*, https://carpentries.org/, oder auf eine Organisation bezogen etwa die *Helmholtz Information & Data Science Academy*, vgl. https://www.helmholtz-hida.de/. Umfangreiche Aktivitäten verfolgen Stifterverband, Centrum für Hochschulentwicklung (CHE) und Hochschulrektorenkonferenz (HRK) mit Unterstützung des Bundesministeriums für Bildung und Forschung (BMBF) im Hochschulforum Digitalisierung, vgl. https://hochschulforumdigitalisierung.de/.

| 28 https://www.rd-alliance.org/about-rda; Wilkinson, M. D. et al.: *The FAIR Guiding Principles for scientific data management and stewardship*, in *Scientific Data*, 3:160018 (2016), DOI: 10.1038/sdata.2016.18, https://www.nature.com/articles/sdata201618.pdf; https://www.go-fair.org.

schließlich ihrer Zugänglichkeit. Sie werden vielfach unter den Überschriften Open Access oder Open Data geführt und setzen auf ein aktives und frühzeitiges Teilen von Datenbeständen über engste Kooperationspartner hinaus. |[29] Dabei darf nicht übersehen werden, dass die Bereitschaft von Wissenschaftlerinnen und Wissenschaftlern zur Anpassung ihrer Arbeitsweise davon abhängt, was ihre Fächer und Heimateinrichtungen zulassen oder aktiv einfordern, aber auch wie und wofür Anerkennung und Reputation im Wissenschaftssystem zugewiesen wird. |[30]

Traditionell wird der Beitrag von Einzelpersonen zur kollektiven Wissensproduktion an ihrer Autorschaft von Publikationen bemessen, deren Anerkennung sich maßgeblich nach der Prioritätsregel bestimmt. Daten gelten in diesem Regime als Produktionsmittel, über die zu verfügen einen Wettbewerbsvorteil bedeutet. Infolgedessen sind Verhaltensweisen in der Forschungspraxis bisher vielfach darauf ausgerichtet, schnell publizierbare Ergebnisse vorzulegen, was sich auf die Bereitschaft zu nachhaltiger Datenpflege und gestuft zugänglich gemachten Datenbeständen negativ auswirkt. Abgesehen von der Abfrage von Forschungsdatenmanagementplänen spielen Daten in Begutachtungsprozessen für Publikationen und Anträge der Forschungsförderung bislang noch keine große Rolle. Im Publikationsbereich wird ebenfalls nach Wegen gesucht, die Anerkennung wichtiger Teilbeiträge zum Erkenntnisfortschritt jenseits der Fixierung auf Publikationen zu verbessern. Neuere Vorschläge im Publikationswesen, aber auch aus Forschungsförderorganisationen zielen deshalb auf eine differenziertere Berücksichtigung von Einzelbeiträgen (*contributorship*) zu einer Publikation an Stelle des bislang dominierenden Verständnisses der Autorenrolle (*authorship*), um dadurch Beiträge zum Fortschritt der Forschung jenseits hergebrachter Autorschaft besser sichtbar zu machen und Kooperationen sowie Software-Entwicklung und Datenanalysen zu befördern. |[31]

III.4 Außenbeziehungen und Wettbewerbssituation

Datenintensive Forschung in öffentlich finanzierten Hochschulen und Forschungseinrichtungen ist auf vielfache Weise mit dem privaten Sektor verbun-

| 29 Als jüngster Appell aus der Wissenschaft hierzu die Erklärung mehrerer internationaler Universitätsverbünde im Januar 2020: Sorbonne-Erklärung für offene Forschungsdaten, https://www.german-u15.de/presse/ressourcen-2020/20200130_Sorbonne-Declaration-on-Research-Data-Rights.pdf. Als Überblick entsprechender Richtlinien und Förderaktivitäten von Forschungsförderern im europäischen Vergleich: Fosci, M.; Richens, E.; Johnson, R.: *Insights into European research funder Open policies and practices*, September 2019, https://doi.org/10.5281/zenodo.3401278.

| 30 Dazu grundsätzlich Fecher, B.: *A reputation economy. How individual reward considerations trump systemic arguments for open access to data*, in: *Palgrave Communications* 3 (2017), Art. 17051, DOI: 10.1057/palcomms.2017.51. Als Überblick über aktuelle Veränderungen in der Bewertung von Forschungsleistungen im europäischen Vergleich: *European University Association: Research Assessment in the Transition to Open Science. 2019 EUA Open Science and Access Survey Results*, Brüssel Oktober 2019, https://eua.eu/downloads/publications/research%20assessment%20in%20the%20transition%20to%20open%20science.pdf.

| 31 Dazu etwa Holcombe, A. O.: *Contributorship, Not Authorship: Use CRediT to Indicate Who Did What*, in: *Publications* 7 (2019) 3, S. 48 ff., DOI: 10.3390/publications7030048.

den, teils auf ihn angewiesen, steht aber auch im Wettbewerb mit ihm, etwa um Personal. Sie verwendet kommerziell entwickelte Hard- und Software. Sie nutzt zudem privatwirtschaftliche Dienstleistungen wie Speicherplatz, Rechenkapazitäten oder Webtechnologien und Vermittlungsplattformen etwa für sozialwissenschaftliche Umfragen. Vielfach erlaubt dies der Wissenschaft, auf aufwändige Eigenentwicklungen zu verzichten, Hardwareeinkäufe durch Mieten zu ergänzen oder zu ersetzen und Teilschritte des Forschungsprozesses, etwa in der Datenerhebung, auszulagern und sich zuliefern zu lassen. Derartige Services und Produkte werden jedoch nicht immer nur für die Wissenschaft entwickelt oder von dieser genutzt. Die entwickelnden Unternehmen stehen in der Regel in Konkurrenz zu anderen Anbietern und schützen ihr Entwicklungswissen im Wettbewerb. Die damit verbundenen Herausforderungen sind nicht durchweg neu, können sich aber mit der Verbreitung datenintensiver Forschung verstärken: Forschung kann durch fehlendes Verständnis der oder mangelnde Transparenz von Tools oder nicht veränderbaren Voreinstellungen von Messgeräten oder langfristige Abhängigkeit bei fehlenden Alternativangeboten unbeabsichtigt beeinflusst oder verzerrt werden. Auch Sicherheitsfragen bei der externen Speicherung und Analyse sensibler Daten müssen geklärt werden. Inländische Forschung, die auf im Ausland liegende Datenkorpora zugreift, unterliegt dem Risiko, den Datenzugang unter veränderten politischen Bedingungen zu verlieren. Umgekehrt unterliegen deutsche Datengeber dem Risiko, die Verwendung ihrer Daten im Ausland bezüglich Normen und Standards nicht kontrollieren zu können. Schließlich sind die Qualitätsansprüche in Kooperationen von Wissenschaft mit privaten Unternehmen Gegenstand von Aushandlungsprozessen.

Vergleichsweise neu ist für die Wissenschaft, dass enorm große Datenbestände in Unternehmen vorhanden sind, die entscheidende Fortschritte in der Methodenentwicklung oder bei der Verfolgung bestimmter Forschungsfragen versprechen (vgl. Fallbeispiel II.1). Diese Daten, etwa aus Social-Media-Plattformen oder von Suchmaschinenbetreibern, sind jedoch für die Wissenschaft teils gar nicht, teils unvollständig, befristet oder nur unter bestimmten Auflagen zugänglich, die mit den üblichen Anforderungen an wissenschaftliches Arbeiten in Spannung stehen können. Um auch mit geschützten Daten aus privaten Unternehmen forschen zu können, werden aus der Wissenschaft neue Wege gesucht und erprobt, etwa durch Partnerschaften zwischen Wissenschaftsförderern und Unternehmen, Kooperationsplattformen oder Datentreuhandstellen. |[32] In der

| 32 Als Beispiel eines neuartigen Kooperationsversuchs zwischen Facebook und dem *Social Science Research Council* vgl. https://socialscience.one/ bzw. https://www.ssrc.org/programs/view/social-data-initiative/; King, G.; Persily, N.: *A New Model for Industry-Academic Partnerships*, in: *Political Science & Politics*, August 2019, S. 1–7, DOI: 10.1017/S1049096519001021; als Beispiel für eine neue Kooperationsplattform zwischen Industrie und Wissenschaft etwa das *ADA Lovelace Center for Analytics, Data and Applications* am Fraunhofer IIS, vgl. https://www.scs.fraunhofer.de/de/referenzen/ada-center.html. Zu einer Datentreuhand als Broker zwischen Unternehmen und Forschenden zuletzt die Datenethikkommission der Bundesregierung in ihrem Gutachten: https://datenethikkommission.de/ sowie der Rat für Sozial- und Wirtschaftsda-

Wissenschaft sind sowohl Befürworter neuer, einem breiteren Datenzugang dienender Kooperationsformate vertreten als auch kritische Stimmen, die den Zugang zu Forschungsdaten aus öffentlich finanzierten Projekten unter bestimmten Bedingungen kontrollieren können wollen. Einvernehmlich ist die Sorge vor Monopolbildungen, die etwa dazu führen könnten, dass öffentlich produzierte Daten wegen unzumutbarer Kosten nur unzureichend genutzt werden, ähnlich wie dies in einzelnen Bereichen des Publikationswesens zu beobachten ist.

III.5 Rechtliche Rahmenbedingungen

Datenintensive Forschung unterliegt vielfältigen rechtlichen Regelungen. Durch die Dynamik der Entwicklung von Methoden und Datennutzungsarten ist sie dabei mit ungeklärten Fragen und Grauzonen konfrontiert, da rechtliche Regelungen wie auch Rechtsprechung den technischen und gesellschaftlichen Entwicklungen erst mit einem gewissen Zeitabstand nachfolgen können. In den letzten Jahren hat sich der Möglichkeitsspielraum datenintensiver Wissenschaft so rasant erweitert, dass Lücken zwischen Regelungsbedarf und tatsächlicher Regelung noch deutlicher sichtbar geworden sind. Zentrale Rechtsbereiche für datenintensive Forschung sind Datenschutz und Urheberrecht. Die Europäische Datenschutzgrundverordnung regelt bereits viele Fragen, doch Unklarheiten bestehen noch fort, etwa was den Ausgleich zwischen Forschung und Datenschutzrecht oder die Definition von Forschung betrifft, für deren Zwecke im Sinne von Art. 89 Ausnahmen von den Bestimmungen der Verordnung gemacht werden dürfen. |[33] Besonders herausfordernd sind der Umgang mit personenbezogenen Daten wie etwa Patientendaten im Bereich der Gesundheitsforschung sowie Fragen zur Verfügbarmachung von Daten zum Zweck bestimmter Studien oder aber deren Löschung, die unter engen Bedingungen ebenfalls geboten sein kann (vgl. Fallbeispiel II.5). Im Urheberrecht existiert ebenfalls ein Rechtsregime, das angewendet werden kann und muss, sobald urheberrechtlich geschützte Werke ihrerseits zum Gegenstand datenintensiver Forschung werden. |[34] Weniger de-

ten in seinen Empfehlungen: Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften. Datenzugang und Forschungsdatenmanagement. Mit Gutachten „Web Scraping in der unabhängigen wissenschaftlichen Forschung“, RatSWD Output 4 (6), Berlin 2019, https://doi.org/10.17620/02671.39. Zuletzt außerdem: Rat für Informationsinfrastrukturen: Stellungnahme Datentreuhandstellen gestalten – Zu Erfahrungen der Wissenschaft, Göttingen 2020, http://www.rfii.de/?p=4259.

| 33 Dazu Rat für Informationsinfrastrukturen: Datenschutz und Forschungsdaten. Rat für Informationsinfrastrukturen plädiert für zeitgemäßeren Datenschutz in der Wissenschaft, Pressemitteilung vom 06.03.2017, http://www.rfii.de/?p=2253; Kreutzer, T.; Lahmann, H.: Rechtsfragen bei Open Science. Ein Leitfaden, Hamburg 2019, https://dx.doi.org/10.15460/HUP.195; Grundlegend: Spiecker gen. Döhmann, I.: Wissensverarbeitung im Öffentlichen Recht, in: Rechtswissenschaft 1 (2010) 3, S. 244–282, https://www.rechtswissenschaft.nomos.de/archiv/2010/heft3/.

| 34 Umfrage des DataJus Projekts (Leitung: Professorin Lauber-Rönsberg) zu rechtlichen Unterstützungsangeboten im Zusammenhang mit dem Forschungsdatenmanagement (FDM) im Dezember 2018 und weitere Informationen: https://tu-dresden.de/gsw/jura/igetem/jfbimd13/forschung/forschungsprojekt-datajus; als Beratungsangebot z. B. das Portal http://forschungslizenzen.de; Lauber-Rönsberg, A.; Krahn, Ph.; Baumann, P.: Gutachten zu den rechtlichen Rahmenbedingungen des Forschungsdatenmanagements im Rahmen des DataJus-Projektes. Kurzfassung, Stand: 12.07.2018, https://tu-dresden.de/gsw/jura/igetem/jfbimd13/ressourcen/dateien/publikationen/DataJus_Kurzfassung_Gutachten_12-07-18.pdf?lang=de&set_language=de.

tailliert ist bisher die Regulierung von Entscheidungen, die auf Grundlage von Ergebnissen datenintensiver Forschung getroffen werden. Hier sind gesellschaftlich-politische Aushandlungsprozesse noch in Recht zu fassen.

Während die Wissenschaftlerinnen und Wissenschaftler Datenschutz und Urheberrecht Dritter berücksichtigen müssen, haben sie zugleich ein eigenes Interesse, selbst Rechte nicht nur an Werken im Sinne des Urheberrechts, sondern auch an den von ihnen erzeugten oder bearbeiteten Daten zu erwerben und anerkannt zu sehen. Das schließt die Abtretung von Rechten an Dritte ein, wo grundsätzlich zwischen externen und kommerziellen Dienstleistern zu unterscheiden ist und Risiken der Abhängigkeit und Monopolbildung bestehen. Da Daten jedoch nicht unter den für das Urheberrecht zentralen Werkbegriff fallen, gibt es kein rechtlich fundiertes Eigentums- oder Zugangsrecht an Daten. Rechte der Datenproduzierenden an Zweitveröffentlichung oder Mitnahme von Daten können daher nur in professionsethischen Regelungen für die Datenzuordnung fundiert sein. An verschiedenen Stellen wird auf Autonomiespielräume der Wissenschaft verwiesen, sich angesichts bestehender Regelungsdefizite eigene personen- oder institutionenzentrierte Regelungen im Rahmen ihrer Selbstverwaltungsstrukturen zu setzen, etwa im Rahmen der Kodifizierung guter wissenschaftlicher Praxis. Solche Bemühungen um dezentrale Regelungen können jedoch dort erschwert werden, wo etwa im Publikationswesen private Anbieter hohe Durchsetzungsmacht für ihre Vorgaben entfalten können.

Neben offenen Fragen an Forschungsdesigns und deren rechtlich korrekte Durchführung besteht aber auch noch Unsicherheit, inwiefern das Bewusstsein für rechtliche Aspekte datenintensiver Forschung bereits hinreichend ausgeprägt ist. Eine besondere Herausforderung ist hierbei, dass wissenschaftliche Fragestellungen und der Austausch darüber häufig in international organisierten Fachgemeinschaften verfolgt werden, deren Mitglieder vor Ort jeweils unterschiedlichen rechtlichen Regelungen unterliegen. Bei Forschungsvorhaben, die in Kooperation mit dem privaten Sektor erfolgen, stellt sich zudem die Frage, welche Grenzen (z. B. im Datenschutz) für die Vorhaben von Wissenschaftlerinnen und Wissenschaftlern aus dem deutschen und europäischen Rechtsraum bestehen und was diese für die Verfolgung bestimmter Fragestellungen zur Folge haben. Hier ist insbesondere zu bedenken, dass durch die Rechtsfigur der „gemeinsamen Verantwortlichkeit“ etwaiges datenschutzrechtliches Fehlverhalten eines Partners auch die anderen erfasst und diese unter Umständen dafür haften.

III.6 Gesellschaftliche Erwartungen und Unsicherheiten

Datenintensive Forschung ist mit vielfältigen positiven und negativen Erwartungen aus Öffentlichkeit und Politik konfrontiert, die Antworten erfordern sowie Anlässe zur Reflexion von Fragestellungen und Methoden bieten. Mediale Darstellungen der Potenziale datenintensiver Forschung konzentrieren sich bislang

häufig auf plakative Möglichkeiten, etwa zum Verlust der individuellen Privatsphäre durch die Anhäufung von Informationsmacht, der Einzelne nichts mehr entgegensetzen können, auf der einen oder optimistische Annahmen zur Verbesserung von Heilungschancen bis hin zur raschen Ausrottung verbreiteter Krankheitsbilder auf der anderen Seite. Zu den Herausforderungen für die Berichterstattung gehört die besondere Diskrepanz zwischen scheinbar leicht fassbaren, ansprechenden Visualisierungen und den dahinter stehenden, hoch komplexen Methoden sowie die Neigung zur anthropomorphisierenden Beschreibung von automatisierten Prozessen, was die Gefahr birgt, das scheinbar Einfache nicht weiter zu hinterfragen. Der Umgang der Medienberichterstattung und veröffentlichten Meinung mit neuen Möglichkeiten datenintensiver Forschung ist daher anspruchsvoll. Das gilt ebenso für die Fähigkeit der Mediennutzenden, Gegenstände und Bewertungen der Berichterstattung über Veränderungen in der Forschung, aber auch Versprechungen von Unternehmen in diesem Bereich aufzunehmen, Implikationen zu erfassen, kritisch zu bewerten und Übertreibungen zu erkennen. Angesichts dieser Unsicherheiten ist Wissenschaft gefordert, Funktionsweise, Chancen und Risiken neuer datenbasierter Analysemethoden proaktiv zu erklären, ihre Qualitätsansprüche zu verteidigen, aber auch Fragen aus Teilen der Gesellschaft zum Anlass eigener Reflexion zu machen, wenn es z. B. um die Konzipierung von Fragestellungen oder auch die Erweiterung von Studiengängen um rechtliche und ethische Fragen geht. Es bleibt jedoch häufig im Ungefähren, an wen solche Erwartungen adressiert und mit welchen Ressourcen sie zu bedienen sind.

Auf Seiten der Politik zeigt sich deutlich der Gestaltungswunsch, datenintensive Forschung einerseits zu fördern und ihre Innovationspotenziale zu heben. Zugleich ist es selbstverständlich, dabei menschen- und grundrechtliche Positionen zu schützen und Beeinträchtigungen solcher Rechte auch durch Anpassung und Ergänzung der rechtlichen Rahmenbedingungen zu verhindern. Zugänglichkeit und Nachnutzbarkeit von Daten wurden in den letzten Jahren durch politische Akteure auf nationaler und internationaler Ebene thematisiert und mit eigenen Erwartungen – vor allem an die Ermöglichung und Beschleunigung von Innovationen – verbunden. Im Fokus stand hier der Zugang zu Daten aus unterschiedlichen Quellen für Nachnutzungen, mit Folgen für die Entwicklung gemeinsamer Standards, den Infrastrukturausbau, verschiedenste Unterstützungsleistungen wie auch die Entwicklung von Richtlinien und Gesetzen. |[35] Die Europäische Kommission verfolgt eine ganze Reihe von Aktivitäten, die sich auf den Wissensaustausch zum Themenfeld, die Gestaltung der Forschungsförde-

| 35 Mit der Einlösung der FAIR-Prinzipien werden beispielsweise auch große wirtschaftliche Vorteile verbunden, vgl. die Studie von PwC *EU Services* für die Generaldirektion Forschung der EU-Kommission: *Cost-Benefit analysis for FAIR research data. Cost of not having FAIR research data*, Brüssel 2018, DOI: 10.2777/02999, https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en.

rung sowie die Ausrichtung der künftigen europäischen Wissenschaftspolitik mit Hilfe externer Beratung (z. B. die *Open Science Policy Plattform* – OSPP) beziehen. |[36] Im Fokus sind zudem konkrete E-Infrastrukturen wie der Aufbau der *European Open Science Cloud* (EOSC), um bestimmte technologische Fähigkeiten und Daten in Europa eigenständig vorzuhalten. Auch in Deutschland haben sich die Regierungen von Bund und Ländern in den letzten Jahren verstärkt der Gestaltung von Rahmenbedingungen datenintensiver Forschung zugewandt. Aktivitäten betreffen die Erarbeitung von Strategien wie auch die Initiierung von Förderprogrammen für den Aufbau von Kompetenzzentren und Infrastrukturen. |[37] Dabei sind Schnittmengen zwischen den Themen Open Access und Open Data sowie datenintensiver Forschung durchaus vorhanden, es handelt sich jedoch nicht um deckungsgleiche Themen. Erst allmählich wird dabei auch auf die Folgen und Konsequenzen eines breiten Zugangs eingegangen, etwa für Verbraucherinteressen, den Datenschutz oder die Autonomie der Bürger und Bürgerinnen.

Erheblicher Beratungsbedarf besteht in der Politik zur Ausgestaltung gesellschaftlicher Transformationsprozesse im digitalen Zeitalter, die mit dem Wissenschaftssystem verbunden sind, aber darüber hinausgehen. Dies zeigen etwa die inzwischen erarbeiteten Stellungnahmen des Ethikrats zu Big Data im Gesundheitsbereich und von Leopoldina, acatech und Akademienunion zu Privatheit in Zeiten der Digitalisierung. |[38] Normative Erwägungen sind im Transformationsprozess des digitalen Zeitalters erneut in den Fokus geraten, wobei durchaus an ältere Traditionen angeknüpft wird, etwa was angewandte Ethik als Partnerin der länger institutionalisierten Technikfolgenabschätzung anbelangt oder auch die in wissenschaftlichen Einrichtungen seit den 1980er Jahren nicht nur im Bereich der Gesundheitsforschung eingerichteten Ethikkommissionen. |[39] Viele Forschungsfelder in Soziologie, Recht und Politikwissenschaft untersuchen Auswirkungen der Nutzung neuer Datenbestände und Analysemethoden auf die Gesellschaft, etwa Folgen für Chancengleichheit, Teilhabe oder Verteilungsentscheidungen. In den Fokus geraten dabei zudem zentrale Fragen

| 36 https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-policy-platform.

| 37 Zuletzt jeweils mit Bezügen auch auf die Wissenschaft – Europäische Kommission: Eine europäische Datenstrategie, Brüssel, 19.02.2020, https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_de.pdf, sowie Bundesregierung: Eckpunkte einer Datenstrategie der Bundesregierung, November 2019, https://www.bundesregierung.de/resource/blob/975226/1693626/60b196d5861f71cdefb9e254f5382a62/2019-11-18-pdf-datenstrategie-data.pdf.

| 38 Deutscher Ethikrat: Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Stellungnahme, Berlin 2018, https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-big-data-und-gesundheit.pdf; Privatheit in Zeiten der Digitalisierung, Stellungnahme 2018, hrsg. v. Leopoldina, acatech und Akademienunion, Halle/Saale 2018, https://www.leopoldina.org/uploads/tx_leopublication/2018_Stellungnahme_BigData.pdf.

| 39 Gutachten der Datenethikkommission der Bundesregierung, Berlin 2019, https://datenethikkommission.de/; vgl. auch den im August 2019 eingerichteten öffentlichen Beirat des Cyber Valley zu ethischen und gesellschaftlichen Implikationen geplanter Forschungsprojekte, http://www.cyber-valley.de/public-advisory-board.

des Selbstbildes von Wissenschaft, nach Anwendungsmöglichkeiten und auch Grenzen von Forschung.

B. Empfehlungen

Datenintensive Forschung bewirkt einen Transformationsprozess im Wissenschaftssystem, der große Chancen bietet, wenn er klug gestaltet und zu diesem Zweck systematisch beobachtet, analysiert und bewertet wird. Mit den folgenden Leitlinien und Empfehlungen will der Wissenschaftsrat Aufmerksamkeit auf jene Herausforderungen richten, die über den Auf- und Ausbau neuer Organisationen hinausgehen, wie er derzeit beispielsweise mit dem Aufbau der Nationalen Forschungsdateninfrastruktur oder dem Verbund für Nationales Höchstleistungsrechnen sowie den Kompetenzzentren für Künstliche Intelligenz erfolgt. Für die Durchführung und Weiterentwicklung datenintensiver Forschung sind erweiterte und auch neue Infrastrukturen elementar, ebenso wie die Neuzuweisung oder Neuaufteilung von Ressourcen in den wissenschaftlichen Einrichtungen. Diese Notwendigkeiten dürfen aber nicht den Blick dafür verstellen, dass ein umfassenderer Kulturwandel in den Wissenschaften notwendig ist, um datenintensive Forschung erfolgreich organisieren und betreiben und um den Nutzen des Teilens von Daten und Software voll entfalten zu können. Dieser Kulturwandel setzt eine intensive Auseinandersetzung der einzelnen Wissenschaftlerinnen und Wissenschaftler mit der Frage voraus, welche Voraussetzungen und Konsequenzen datenintensive Forschung hinsichtlich der Nutzung von Daten und der damit verbundenen Methoden hat. Auf entsprechende Verständigungsprozesse in den Fachgemeinschaften und deren Foren müssen Hochschulen und Forschungseinrichtungen sowie die Forschungsförderer mit Bund und Ländern aufsetzen und die Transformation nachdrücklich unterstützen.

B.I LEITLINIEN ZUM KULTURWANDEL IN DEN WISSENSCHAFTEN

I.1 Leitlinie 1: Teilen und Kooperieren

Technische, rechtliche, organisatorische und soziale Voraussetzungen und Regeln der Forschung müssen so gestaltet werden, dass sie das Teilen von Daten und Software in der Wissenschaft und ihre kooperative Bearbeitung fördern. Dabei können Zugangsbeschränkungen zu in der Wissenschaft erzeugten oder bearbeiteten Daten nicht nur aus rechtlichen Gründen notwendig sein, bedürfen aber in allen anderen Fällen jeweils einer Begründung.

Zielvorstellung: Die Nutzung, Wiedernutzung und Verknüpfung gut gepflegter und kontinuierlich fortentwickelter Datenbestände verspricht vielfältige Gewinne an Erkenntnis und Effizienz für die Forschung. Damit Daten von Wissenschaftlerinnen und Wissenschaftlern in verschiedensten Gebieten und mit unterschiedlichsten Fragestellungen geteilt, weiterverwendet und neu verknüpft werden können, müssen technische, rechtliche, organisatorische und soziale Aspekte des Datenteilens überzeugend und nachhaltig geregelt sein. Zu den Voraussetzungen gehören Standards und Infrastrukturen, Zugänglichkeit notwendiger Software, rechtliche und untergesetzliche Regelungen sowie die Anerkennung der erbrachten Leistungen über disziplinäre, institutionelle und geographische/politische Grenzen hinweg. Entsprechende Anstrengungen sollten auch dort vorangetrieben werden, wo ein kooperatives Vorgehen im Umgang mit Daten und Software bislang nicht oder nur in Ausnahmefällen üblich ist. Ein verantwortungsvolles und faires Teilen von Daten in der Wissenschaft beinhaltet transparente Regelungen des Zugangs zu und der Verwendung von diesen Daten. Personenbezogene und personenbeziehbare wie auch urheberrechtlich oder durch Regelungen von Vertragsparteien geschützte Daten können spezielle Bestimmungen über gestufte Regelungen des Zugangs zu diesen und ihrer Weiterverwendung notwendig machen. Eine exklusive Aneignung öffentlich produzierter Daten durch kommerzielle Anbieter ist ebenso abzulehnen wie eine nicht offengelegte und Regeln guter wissenschaftlicher Praxis zuwiderlaufende Nutzung von fremden Daten.

Handlungsbedarf: Die Klärung technischer, rechtlicher, organisatorischer und sozialer Aspekte des Datenteilens erfordert Beiträge unterschiedlicher Akteure innerhalb und außerhalb des Wissenschaftssystems. Grundsätzlich sind die einzelnen Wissenschaftlerinnen und Wissenschaftler gefordert, an Verständigungsprozessen in ihren Fachgemeinschaften mitzuwirken und bei neuartigen Fragestellungen und Möglichkeiten selbstverständlich auch den vorhandenen Rechtsrahmen zu berücksichtigen. Wo Regelungslücken bestehen, sind innerhalb der Fachgemeinschaften Üblichkeiten zu vereinbaren, auf die auch von außen Bezug genommen werden kann. Sofern entsprechende Verständigungen in den (meist internationalen) Fachgemeinschaften und übergreifenden Foren erzielt wurden oder noch erarbeitet werden, sollten Hochschulen und Forschungseinrichtungen wie auch Forschungsförderer nach Möglichkeit darauf aufsetzen, wenn sie Anforderungen zum Zugang zu Daten (durch Produzenten oder externe Nutzer und Anbieter) und zur Anerkennung von Teilbeiträgen und Rollen bei deren Erhebung, Aufbereitung und Analyse festlegen. Die sogenannten FAIR-Prinzipien für den Austausch, die Nutzung und Nachnutzbarkeit von maschinenlesbaren Daten bieten dafür einen geeigneten Ausgangspunkt, der im Rah-

men der guten wissenschaftlichen Praxis genutzt werden sollte und auf den weiter aufgebaut werden kann. |[40]

I.2 Leitlinie 2: Themen- und Methodenvielfalt

In den Forschungsfeldern müssen sowohl methodisch verschiedene datenintensive Ansätze als auch andere Forschungsformen nebeneinander bestehen, weiterentwickelt werden und neu entstehen können. Dafür sind Freiräume zu erhalten und Angebote in den Organisations- und Förderbedingungen der Forschung auszubalancieren.

Zielvorstellung: Die dynamische Entwicklung datenintensiver Forschung ist auf allen Ebenen des Wissenschaftssystems so aufzunehmen, dass neue Möglichkeiten erkannt und genutzt, daneben aber auch Gefahren reflektiert sowie Schieflagen und Fehlentwicklungen vermieden werden. Neue Methoden und Datenbestände, weiterentwickelte Kompetenzprofile, Dienstleistungen und Organisationsstrukturen können dann die Wissenschaft produktiv unterstützen und Erkenntnisgewinne ermöglichen, ohne bereits bestehende Forschungsfelder oder weniger datenintensive Forschungsformen (z. B. hermeneutisch-interpretierende) zu benachteiligen, einseitig bestimmten Trends zu folgen und schwer korrigierbare Pfadabhängigkeiten zu schaffen.

Handlungsbedarf: Wissenschaftlerinnen und Wissenschaftler müssen die Erweiterung ihrer Fragestellungen, Datengrundlagen und Analysemöglichkeiten in ihren Fachgemeinschaften und deren Foren reflektieren. Dies kann beispielsweise in Zeitschriften, Internetforen oder auf Tagungen, aber auch in großen Forschungsverbünden oder Konsortien geschehen. Abhängig vom Entwicklungsstand datenintensiver Forschung in der jeweiligen Fachgemeinschaft ist auch das Verhältnis unterschiedlich datenintensiver Fragestellungen und entsprechender Forschungsrichtungen zueinander zu bestimmen. Dabei ist zugleich in den Blick zu nehmen, inwiefern neu verfügbare Datenbestände zu einer Überbetonung bestimmter Forschungsthemen führen und welche unbeabsichtigten Folgen dies haben kann. Sinnvoll wäre es, die Ergebnisse solcher Beratungen innerhalb von Fachgemeinschaften in einem Weißbuch oder einer vergleichbaren strategischen Positionsbestimmung festzuhalten, um erzielte Verständigungen zu fixieren sowie Organisations- und Sprechfähigkeit nach außen zu erlangen.

Diese fachinterne Verständigung muss um einen fachübergreifenden Austausch ergänzt werden, um die Weitergabe fachspezifischer und die Übernahme fachfremder Methoden und Standards sowie die Interoperabilität der erarbeiteten Lösungen zu ermöglichen. Dies betrifft sowohl Datenbestände als auch Stan-

| 40 Zum Ergänzungsbedarf der FAIR-Kriterien auch Rat für Informationsinfrastrukturen: Herausforderung Datenqualität. Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage, Göttingen 2019, http://www.rfii.de/?p=4043, insbes. S. 23 ff.

dards, Workflows und Methoden einzelner Fachgemeinschaften. Verschiedene nationale und internationale Foren bestehen inzwischen hierfür (z. B. *Research Data Alliance* – RDA, *Committee on Data of the International Science Council* – CODATA, de-RSE zur Softwareentwicklung sowie die im Aufbau befindlichen Strukturen der Nationalen Forschungsdateninfrastruktur – NFDI und der *European Open Science Cloud* – EOSC). An diesen Foren sollten sich Wissenschaftlerinnen und Wissenschaftler aktiv beteiligen, um den Austausch zwischen Forschenden und Anbietern von Services für die Forschung zu verbessern.

Auch in den wissenschaftlichen Einrichtungen, von Wissenschaftsförderern und Zuwendungsgebern in Bund und Ländern müssen Veränderungen der Forschungsmöglichkeiten und -schwerpunkte durch datenintensive Forschung berücksichtigt werden. Die wissenschaftlichen Einrichtungen müssen den Austausch datenintensiv Forschender unterstützen und dafür ihre Strukturen und Ressourcenzuordnungen in den Blick nehmen. Die verschiedenen Förderer von Wissenschaft sollten ihre Portfolios überprüfen und anpassen. Neue Förderformate können erprobt und Anpassungen bestehender an die Erfordernisse datenintensiver Forschung geprüft werden. Dabei muss das Risiko einer zeitweiligen Überbetonung der Förderung von einzelnen Bereichen datenintensiver Forschung mit speziellen Methoden sowohl in Programmgestaltung und Ausschreibung als auch in den zugehörigen Begutachtungs- und Auswahlprozessen beobachtet und begrenzt werden.

I.3 Leitlinie 3: Kompetenzaufbau und Spezialisierungen

Kompetenzen im Umgang mit Daten und Methoden ihrer Verarbeitung müssen für Tätigkeiten innerhalb und außerhalb der Wissenschaft systematisch vermittelt werden. Dabei sind unterschiedliche Spezialisierungsgrade und Karrierepfade zu berücksichtigen, die von allgemeiner Data Literacy über fachspezifische bis zu fachübergreifenden Kompetenzprofilen reichen.

Zielvorstellung: Eine systematische Weiterentwicklung von Kompetenzen im Umgang mit Daten und Methoden ist erfolgskritisch für die weitere Gestaltung des Transformationsprozesses in der Wissenschaft. Entsprechend breit gilt es die Erweiterung der Kompetenzen anzugehen, von der Aktualisierung und Erweiterung der Studiengänge und Lehrinhalte für Tätigkeiten innerhalb und außerhalb der Wissenschaft über die Weiterbildung der bereits an Hochschulen und Forschungseinrichtungen tätigen Personen bis hin zur Ergänzung von Tätigkeitsprofilen und Berufsbildern im Wissenschaftssystem.

Handlungsbedarf: Die umfassende Vermittlung datenbezogener Kompetenzen erfordert Beiträge zahlreicher Akteure im Wissenschaftssystem. Impulse für die Anpassung von Curricula und Weiterqualifizierungsmöglichkeiten in Hochschulen und Forschungseinrichtungen müssen aus den Fachgemeinschaften kommen. Lehrkapazitäten und Lehrinhalte in Informatik, Statistik, Recht, Ökonomie und Ethik müssen erweitert und aufeinander bezogen werden. Zugleich

gilt es das vorhandene Personal im Forschungs- und Infrastrukturbereich besser in Austauschforen miteinander zu verbinden, Fort- und Weiterbildungsangebote und neue Tätigkeitsprofile zu schaffen sowie stabil zu verankern. Auch im Rahmen des Infrastrukturaufbaus sollten diese Angebote adressiert werden. Detaillierte und weitreichende Vorschläge hierfür hat zuletzt der Rat für Informationsinfrastrukturen unterbreitet. |[41]

I.4 Leitlinie 4: Anerkennung von Daten- und Softwarearbeit

Hochwertige Beiträge zu kuratierten Datensammlungen und Datenpublikationen, zum Forschungsdatenmanagement wie auch zur Methoden- und Softwareentwicklung müssen als genuine Leistungen von Wissenschaftlerinnen und Wissenschaftlern anerkannt werden und Unterstützung erfahren.

Zielvorstellung: Ein Kulturwandel ist dann erfolgreich vollzogen, wenn Daten nicht mehr bloß als Zwischenprodukt für eine Publikation betrachtet werden, sondern ihre Aufbereitung für verschiedenste wissenschaftliche Nutzungen stabil organisiert werden kann. Dafür müssen die anfallenden Arbeiten an Datenbeständen sichtbarer werden, mehr Anerkennung finden und zu mehr wissenschaftlicher Reputation führen als dies bislang der Fall ist. Gleiches gilt für die Entwicklung von Standards, Methoden und Software, die für die Analyse und Nachnutzung von Daten unverzichtbar sind. Die Anerkennung dieser Beiträge zur datenintensiven Forschung steht in direktem Zusammenhang mit dem Kulturwandel, der auch beim Teilen, kooperativen Bearbeiten und Zugänglichmachen von Daten erforderlich ist (dazu Leitlinie 1).

Handlungsbedarf: Damit Beiträge zur Entwicklung von qualitätsgesicherten Datenbeständen und neuen Analysemöglichkeiten als Voraussetzung von Fortschritten in der Forschung besser sichtbar werden, müssen die unterschiedlichen Schritte im Prozess der Produktion von Forschungsdaten und die zugehörigen Rollen definiert werden. Festgelegt werden muss dabei auch, welche Rechte sich aus diesen verschiedenen Rollen ergeben, etwa beim Wechsel von Wissenschaftlerinnen und Wissenschaftlern zwischen verschiedenen Einrichtungen oder gegenüber externen Anbietern für die Speicherung von Daten und die Durchführung von Analysen. Die Wissenschaftlerinnen und Wissenschaftler sollten Engagement für hochwertige Datenarbeit in ihrer ganzen Bandbreite von Tätigkeiten (Standards und Interoperabilität, Kuratierung u. a. m.) befördern, indem sie dessen Bedeutung dem wissenschaftlichen Nachwuchs in ihrer jeweiligen Fachgemeinschaft vermitteln und sich selbst in Foren zu entsprechenden Fragen engagieren. Die Fachgemeinschaften können die Reputationswirksam-

| 41 Rat für Informationsinfrastrukturen: Digitale Kompetenzen – dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft, Göttingen 2019, http://www.rfii.de/?p=3883.

keit solcher Aktivitäten befördern, in dem sie auf die Kenntlichmachung unterschiedlicher Teilbeiträge etwa in Publikationen achten (vgl. zum *contributorship*-Standard S. 30) oder die Auszeichnung entsprechenden Engagements mit Preisen fördern. |[42] Höhere Ansprüche an die Dokumentation der Schritte im Prozess der Produktion und Verarbeitung von Forschungsdaten in wissenschaftlichen Publikationen, die mit Blick auf die Reproduzierbarkeit ohnehin gefordert sind (vgl. Leitlinie 5), können ebenfalls zur Steigerung der Anerkennung der Teilbeiträge zu diesem Prozess führen. Wissenschaftliche Einrichtungen müssen Datenzuordnungsfragen in Arbeitsverträgen und institutioneninternen Richtlinien sowie Kooperationsvereinbarungen mit externen Partnern regeln. Sie sollten die Bedeutung hochwertiger Datenarbeit in ihrer Rekrutierungspraxis und bei der Zuteilung von Ressourcen berücksichtigen. Letzteres gilt auch für die Forschungsförderer, die mit der Nachfrage nach Forschungsdatenmanagementplänen bereits zur verbesserten Problemwahrnehmung beigetragen haben und weitere Elemente ihrer Antrags- und Begutachtungsprozesse überprüfen sollten.

I.5 Leitlinie 5: Nachnutzen und Reproduzieren

Forschungsdaten und die bei ihrer Erzeugung und Verarbeitung verwendeten digitalen Instrumente müssen im Interesse von Wissenschaft und Gesellschaft zuverlässig verfügbar bleiben, um die Reproduzierbarkeit von Ergebnissen zu sichern und weitere Nachnutzungen auch nach mittleren und langen Zeiträumen zu ermöglichen. Eine zentrale Voraussetzung dafür ist die systematische und nachvollziehbare Dokumentation der einzelnen Verarbeitungsschritte.

Zielvorstellung: Forschungsergebnisse und ihre Grundlagen können in einem breit verstandenen, sowohl Daten als auch Software umfassenden Sinn dann für verschiedene Zwecke verlässlich (nach-)genutzt werden, wenn stabile Prozesse und Strukturen zur Datenpflege, -auswahl und -aufbewahrung gefunden und implementiert werden. Der Nutzen solcher nachhaltigen Lösungen kann sich innerhalb der Wissenschaft auf neue Forschungsfragen beziehen, die mit bereits bestehenden oder neu verknüpften Datenbeständen verfolgt werden, aber auch auf die Überprüfung wissenschaftlicher Erkenntnisse abzielen. Die Nachnutzung von Datenbeständen verspricht zugleich Nutzen für das Innovationsgeschehen über die Wissenschaft hinaus, wenn mit Unsicherheiten in der Nutzbarkeit und Qualitätseinschätzung von Datenbeständen transparent umgegangen wird. Dafür müssen die Annotation der Primärdaten durch Metadaten mit maschinell verarbeitbaren kontrollierten Vokabularien geklärt sein und die Verarbeitungsschritte von Rohdaten zu Primärdaten systematisch dokumentiert werden. Sind die Methoden zur Vorverarbeitung der Rohdaten langfristig

|[42] Hierzu auch Rat für Informationsinfrastrukturen: Herausforderung Datenqualität. Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage, Göttingen 2019, http://www.rfii.de/?p=4043.

anerkannt, kann deren Löschung geprüft werden. Wenn Primärdaten leicht neugeneriert werden können oder es starke Argumente dafür gibt, dass sie irrelevant sind, kann auf ihre langfristige Speicherung unter Umständen ebenfalls verzichtet werden.

Handlungsbedarf: Wissenschaftlerinnen und Wissenschaftler müssen die qualitätsgesicherte Datenaufbereitung, -auswahl und -archivierung als wichtige Aufgabe anerkennen, weiterentwickeln und Unterstützung dafür erfahren. Dabei können sie vielfach bei Förderern und den eigenen Heimateinrichtungen auf entsprechende Handreichungen und Richtlinien zurückgreifen. Die Notwendigkeit hierfür wird inzwischen von verschiedenen Akteuren im Wissenschaftssystem betont und wurde mit der sogenannten FAIR-Initiative wesentlich befördert. Handreichungen und Richtlinien sollten stetig weiterentwickelt und fachspezifisch angepasst werden. Hierzu gehört in vielen Fällen eine Verständigung über Basis- und Referenzdatensätze, um Standards für die Fachgemeinschaft und auch in Kooperationen nach außen zu bestimmen und nachvollziehbar zu dokumentieren. Angepasst an das jeweilige Fachgebiet sind die Prozesse der Erarbeitung der Daten und notwendige Metadaten sowie deren Strukturen ebenfalls nach allgemeingültigen Standards oder Protokollen zu dokumentieren. Ferner gilt es, sich über fachspezifische und wissenschaftsadäquate Kriterien für die Auswahl oder aber Löschung von Daten zu verständigen, die bis zu Festlegungen über das langfristig zu bewahrende „Daten-Erbe" einer Fachgemeinschaft führen können, dessen öffentliche Zugänglichkeit gesichert sein soll. Stärker als bislang ist der Blick zudem auf Soft- und Hardwareentwicklungen zu richten, die teils in der Wissenschaft, teils von externen Dienstleistern im privaten Sektor erbracht werden. Hier fehlen vielfach noch Standards der Archivierung, die – beginnend mit der Aushandlung der erforderlichen Berechtigungen im Falle proprietärer Software – das Variantenmanagement von Software oder auch die Unabhängigkeit der Ausführbarkeit von bestimmten Betriebssystemen betreffen. Wichtig ist ein systematisiertes Vorgehen, um Daten und Software digital zu verknüpfen und Insellösungen sowohl für die Archivierung von Daten als auch von Soft- und Hardware zu vermeiden. Oberhalb der einzelnen Fachgemeinschaften ist die Nachnutzbarkeit personenbeziehbarer Daten zu regeln. Zu klären ist dabei unter anderem, unter welchen Bedingungen die Forschungsklausel nach Art. 89 DSGVO greift, wie Einverständniserklärungen rechtssicher so formuliert werden können, dass sie Nachnutzungen ausdrücklich erlauben, |[43] und welche Anwendungsmöglichkeiten es für neuere Vorschläge wie Datenspenden gibt.

| 43 Für die krankheits- und institutionenübergreifende Mehrfachnutzung von Patientendaten wurden im Rahmen der Medizininformatik-Initiative Einwilligungsdokumente entwickelt, die die Erteilung eines *broad consent* ermöglichen (vgl. https://www.medizininformatik-initiative.de/) und inzwischen auf Zustimmung der Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder gestoßen sind (vgl. Datenschutz-

Das Wissenschaftssystem muss sich den Herausforderungen der Schnelllebigkeit von Datenangeboten, Hard- und Software-Entwicklungen sowie dem immer schnelleren Aufkommen neuer Methoden stellen. Die Offenheit für teils radikale Neuerungen muss zugleich in ein ausgewogenes Verhältnis zum Interesse an Beständigkeit, Reproduzierbarkeit und nachhaltiger Nutzung bewährter Lösungen gebracht werden.

Zielvorstellung: Datenintensive Forschung in Hochschulen und Forschungseinrichtungen bewegt sich in einem dynamischen, von privaten Kooperationspartnern wie auch Wettbewerbern beeinflussten Umfeld, auf das die wissenschaftlichen Einrichtungen in angemessener Geschwindigkeit reagieren können müssen. Gelingt dies, werden Aus- und Weiterbildungsmöglichkeiten regelmäßig den technischen und methodischen Entwicklungen angepasst und gleichzeitig der Arbeitsmarkt Wissenschaft attraktiv gehalten. Prozesse und Strukturen datenintensiver Forschung müssen regelmäßig und häufig schneller als bisher überprüft und angepasst und zugleich möglichst nachhaltige Lösungen erreicht werden. Zu beachten ist, dass neue, skalierbare Methoden die Zukunft eines Forschungsfeldes nachhaltig prägen können und daher substantielle Ressourcen für das Hinterfragen, Vergleichen und Weiterentwickeln von Methoden bereitgestellt werden müssen.

Handlungsbedarf: Die Geschwindigkeit, mit der sich datenintensive Forschung entwickelt, erfordert dementsprechend beschleunigtes Handeln auf verschiedenen Ebenen im Wissenschaftssystem. Neue Methoden müssen disziplinenübergreifend schneller aufgenommen, überprüft und weiterentwickelt werden, sowohl von den Wissenschaftlerinnen und Wissenschaftlern selbst (z. B. über Konferenzen, aber auch neuere Formate wie Hackathons und Datathons |[44]) als auch von den Hochschulen, die diese Entwicklungen zeitnah in ihre Curricula aufnehmen müssen (vgl. Leitlinie 3). Die wissenschaftlichen Einrichtungen müssen im Kontext datenintensiver Forschung vielfach ihre Reaktionsgeschwindigkeit grundsätzlich erhöhen, um neben Aus- und Weiterbildungsinhalten und -angeboten interne Ressourcenzuweisungen für Forschung und Infrastrukturen sowie Strukturen und Prozesse ebenfalls anzupassen, ohne dabei die Vielfalt von Fragestellungen und Methoden zu gefährden (vgl. Leitlinie 2). |[45] Die Forschungsförderer wiederum müssen nicht nur die Maßstäbe, sondern auch das

konferenz: Beschluss der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder zu den Einwilligungsdokumenten der Medizininformatik-Initiative des Bundesministeriums für Bildung und Forschung, 15. April 2020, https://datenschutzkonferenz-online.de/media/dskb/20200427_Beschluss_MII.pdf).

|[44] Hackathons und Datathons sind Veranstaltungsformate, bei denen spezifische daten- oder methodenbezogene Probleme im Wettbewerb von Einzelnen oder in Teams gelöst werden sollen.

|[45] Die COVID-19-Pandemie hat in manchen Feldern als Katalysator gewirkt und Entwicklungen, die bereits angelegt waren, massiv beschleunigt (vgl. Teil C, S. 59 ff.). Dies ändert aber nichts daran, dass sich die Veränderungsbereitschaft und -geschwindigkeit auch außerhalb von Krisenzeiten wird erhöhen müssen.

Tempo ihrer Förderentscheidungen und Reaktionen auf das Aufkommen andersartiger und zusätzlicher Förderbedarfe (z. B. zur Methodenreflexion und zur Anbahnungsförderung) überprüfen.

Gleichzeit sind an verschiedenen Stellen im Wissenschaftssystem aber auch Maßnahmen zur Stabilisierung und zur nachhaltigen Entwicklung von Prozessen und Strukturen zu treffen. Die Wissenschaftlerinnen und Wissenschaftler setzen sich mit ihren Fachgemeinschaften dem Vorwurf mangelnder Nachhaltigkeit aus, wenn sie Daten nicht auf Nachnutzbarkeit prüfen und neu entwickelte Methoden aus anderen Disziplinen nicht systematisch prüfen, anpassen und weiterentwickeln. Die wissenschaftlichen Einrichtungen müssen organisatorische Vorkehrungen für nachhaltige Lösungen bei der Datensammlung und Datenkuratierung schaffen und diese angemessen ausstatten. Förderer und Zuwendungsgeber müssen Methodenentwicklung und -überprüfung zusätzlich unterstützen, auch über die Förderung nachhaltiger Softwareentwicklung und -nutzung. Datensammlung und -validierung sind langfristig stabil anzulegen und dabei ist zu beachten, dass die Nachnutzung von Daten und Software prinzipiell auf unbestimmte, unter Umständen unbegrenzte Zeit möglich sein muss (vgl. Leitlinie 7).

I.7 Leitlinie 7: Wissenschaftliche Standards in Kooperationen

Öffentlich finanzierte Wissenschaft soll neben eigenen Daten auch Forschung mit Daten der öffentlichen Hand und des privaten Sektors vorantreiben und deren Nutzungen in der Gesellschaft unterstützen. In Kooperationen mit nichtwissenschaftlichen Partnern müssen jedoch wissenschaftliche Standards, die Einhaltung rechtlicher Regelungen und übergreifende Regeln guter wissenschaftlicher Praxis gewährleistet sein.

Zielvorstellung: Datenintensive Forschung in Hochschulen und Forschungseinrichtungen verwendet Daten der öffentlichen Verwaltung und des privaten Sektors für eigene Fragestellungen sowie in wechselseitigen Kooperationen mit diesen gesellschaftlichen Teilsystemen. Auch vollständig oder in Teilen kommerziell angebotene Software, Mess- und Analyseinstrumente werden in der öffentlich finanzierten datenintensiven Forschung verwendet. In beiden Fällen sichert eine kritische Auseinandersetzung mit den Rahmenbedingungen der jeweiligen Nutzungen und Kooperationen, dass die Wissenschaft ihre fachspezifischen Standards und die übergreifenden Regeln guter wissenschaftlicher Praxis reflektiert, gegenüber ihren Kooperationspartnern thematisiert und auf deren Einhaltung achtet.

Handlungsbedarf: Damit wissenschaftsinterne Standards datenintensiver Forschung eingehalten werden können, müssen Wissenschaftlerinnen und Wissenschaftler besondere Sensibilität im Umgang mit solchen Daten zeigen, die im öffentlichen oder privaten Bereich anfallen. Herkunft und Qualität von Daten müssen beurteilt werden können und die Angemessenheit der in Kooperationen

verfolgten Fragestellungen nach wissenschaftlichen, professionsethischen und rechtlichen Standards geprüft werden. Diese Aufgabe fällt allen Wissenschaftlerinnen und Wissenschaftlern zu, die entsprechende Forschungsfragen verfolgen. Sie sollten dafür auf Richtlinien ihrer Heimateinrichtung sowie Standards ihrer Fachgemeinschaften zurückgreifen und beratende Unterstützung lokal oder von spezialisierten Stellen in Anspruch nehmen können. Lösungen müssen so jeweils für Rechtskonformität, die Wahrung guter wissenschaftlicher Praxis und wissenschaftlicher Integrität, wie auch für Spannungen zwischen genuinen Unternehmensinteressen an exklusiv nutzbaren Daten und dem Gemeingutcharakter von Daten in der Wissenschaft gesucht werden (dazu Leitlinie 1 und Leitlinie 5). |[46] Es ist hilfreich, wenn die Forschung noch intensiver auf Daten der öffentlichen Hand zurückgreifen kann. Der Wissenschaftsrat bittet Bund und Länder zu prüfen, ob diese Daten der Forschung – trotz der damit verbundenen rechtlichen Herausforderungen – noch besser zugänglich gemacht werden können. Denn die vermehrte Bereitstellung von Daten der öffentlichen Hand stärkt nicht allein die Forschung, sondern die Ergebnisse wirken in unterschiedlicher Weise in die Gesellschaft zurück.

Auch die Nutzung (teil-)kommerziell entwickelter Software und Analyseinstrumente muss von den Wissenschaftlerinnen und Wissenschaftlern auf die Einhaltung der in den Fachgemeinschaften formulierten Standards geprüft werden. Langfristige, irreversible Abhängigkeiten von Diensten aus privater Hand (sogenannte Lock-in-Effekte) müssen vermieden werden. Dies kann dazu führen, dass Fachgemeinschaften die bisherige Praxis bei der Erhebung, Analyse, Archivierung, Verfügbarkeit und Nachnutzung von Daten überprüfen und neu ordnen müssen und dafür die Unterstützung durch öffentlich finanzierte Infrastrukturangebote benötigen.

I.8 Leitlinie 8: Gesellschaftlicher Austausch

Wissenschaft muss gerade in diesem Feld den Dialog mit der Gesellschaft suchen, um die Veränderungen ihrer Grundlagen, Fragestellungen und Ergebnisse durch datenintensive Forschung transparent zu machen. Sie muss in diesem Austausch Impulse aus der Gesellschaft zur eigenen Weiterentwicklung aufnehmen.

Zielvorstellung: Datenintensive Forschung mit neuartigen Methoden und neuartigen oder neu verknüpften Datenbeständen birgt Chancen, bedarf aber besonderer Aufmerksamkeit für den Dialog mit allen gesellschaftlichen Bereichen, der

| 46 Wichtige Überlegungen zu Datentreuhandsystemen zuletzt durch die Datenethikkommission der Bundesregierung: Gutachten der Datenethikkommission, Berlin 2019, https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf, sowie den Rat für Sozial- und Wirtschaftsdaten in seinen Empfehlungen: Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften. Datenzugang und Forschungsdatenmanagement. Mit Gutachten „Web Scraping in der unabhängigen wissenschaftlichen Forschung“, RatSWD Output 4 (6), Berlin 2019, https://doi.org/10.17620/02671.39.

über übliche Bedarfe der Wissenschaftskommunikation hinausgeht. Es gilt den Erklärungsbedarf all jener zu berücksichtigen, die außerhalb der Wissenschaft auf wissenschaftliche Ergebnisse zurückgreifen – in Politik, Wirtschaft, Medien, Öffentlichkeit aus je unterschiedlichen Interessen – oder diese auch nur beobachten und verstehen möchten. Der Austausch von datenintensiver Wissenschaft und Gesellschaft muss so angelegt sein, dass nicht nur Veränderungen innerhalb der Wissenschaft in die Gesellschaft hinein kommuniziert, sondern auch Beiträge aus der Gesellschaft, einschließlich Sorgen und Ideen, von der Wissenschaft aufgenommen werden.

Handlungsbedarf: Wissenschaftlerinnen und Wissenschaftler sollten dafür ihre Bemühungen verstärken, Vorgehen und Erkenntnisse datenintensiver Wissenschaft, deren Nutzen, aber auch Chancen und Risiken von Veränderungen der wissenschaftlichen Praxis an ein breites Publikum zu vermitteln, um das Bild solcher Forschung in der Öffentlichkeit differenziert auszugestalten. Diese anspruchsvollen Kommunikationsaufgaben müssen nicht zwingend von jeder Wissenschaftlerin und jedem Wissenschaftler wahrgenommen werden, sondern können auch arbeitsteilig organisiert werden. Besonders naheliegende Mittel der interaktiven Vermittlung sollten genutzt und Ressourcen dafür bereitgestellt werden (z. B. Online-Kurse mit „Do-it-yourself"-Datenanalysen/Analysedemos, interaktive Ergebnisberichte, Datenpublikationen und dynamische Visualisierungen, Hackathons). Vielfach werden verantwortliche Wissenschaftlerinnen und Wissenschaftler dabei auf Unterstützung aus ihren Einrichtungen zurückgreifen und Expertise in der Informationsvermittlung und Wissenschaftskommunikation in Anspruch nehmen können.

Für einen umfassenden Dialog zwischen Wissenschaft und Gesellschaft, der neben Daten (z. B. aus Apps von Bürgerwissenschaftsprojekten) auch Impulse für die Weiterentwicklung von Forschungsfragen und -methoden in die Wissenschaft zurückträgt und die Reflexivität der Wissenschaft erhöhen kann, ist der Bedarf neuer Interaktionsformate, Foren oder Arenen zu prüfen, wie sie bereits im Kontext bürgerwissenschaftlichen Engagements erprobt worden sind. Hier könnten Stiftungen wertvolle Anstöße geben und Unterstützung leisten.

B.II EMPFEHLUNGEN AN ZENTRALE AKTEURE IM WISSENSCHAFTSSYSTEM

Zur Konkretisierung der Leitlinien werden im Folgenden Empfehlungen an zwei zentrale Akteursgruppen im Wissenschaftssystem formuliert: An die Hochschulen und Forschungseinrichtungen sowie an Forschungsförderer mit Bund und Ländern. Diese Akteure können den Kulturwandel im Zusammenhang mit datenintensiver Forschung wesentlich befördern und unterstützen. Dazu sind sie auf Positionierungen der Wissenschaftlerinnen und Wissenschaftler angewiesen. Auf die grundsätzliche Bedeutung fachspezifischer Verständigungsprozesse, die abhängig von der jeweiligen Tradition der Datennutzung wie auch der

kritischen Auseinandersetzung mit der Replikation von Forschungsergebnissen unterschiedlich weit fortgeschritten sind, hat die Deutsche Forschungsgemeinschaft bereits hingewiesen. |[47] Dieser fachspezifische Diskussionsprozess über Aufbereitung und Nachnutzung von Daten muss fortgesetzt werden und in eine breitere Auseinandersetzung mit Chancen und Risiken datenintensiver Forschung münden, wozu etwa die Reflexion veränderter Arbeitsweisen, methodischer Profile und Erkenntnisansprüche gehört. Die Fachgemeinschaften (im Sinne von wissenschaftlichen Gemeinschaften und *sub-communities*) sind aufgefordert, entsprechende Strategien zu entwickeln und ihre Bedarfe an unterschiedlichen Infrastrukturen und Diensten wie auch Datensammlungen klar zu formulieren, so dass diese von wissenschaftlichen Einrichtungen, Förderern sowie Bund und Ländern aufgenommen werden können. Eine Herausforderung in diesem Prozess ist die unterschiedlich stabile Organisationsform, die solche Fachgemeinschaften gefunden haben. So bestehen teils national, teils international formale Organisationen wie Fachgesellschaften mit eigener Rechtsform neben loseren, etwa über Foren wie Zeitschriften und Tagungen oder auch Forschungsverbünde organisierten Fachgemeinschaften. Daneben werden voraussichtlich die Konsortien der Nationalen Forschungsdateninfrastruktur einen wichtigen Beitrag hierzu leisten. |[48]

II.1 Hochschulen und Forschungseinrichtungen

Hochschulen und Forschungseinrichtungen sind durch datenintensive Forschung mit breit gefächerten Herausforderungen konfrontiert, die Spielräume für individuelle Entwicklungen eröffnen. Sie sollten die Chancen des Transformationsprozesses wahrnehmen und sich mit den nachfolgenden Themen, für die nur bisher noch wenige gute Praxisbeispiele vorliegen, auseinandersetzen. |[49]

| 47 Deutsche Forschungsgemeinschaft: Leitlinien zum Umgang mit Forschungsdaten, 30.09.2015, https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf, sowie eine Übersicht fachspezifischer Empfehlungen zum Umgang mit Forschungsdaten: https://www.dfg.de/foerderung/antrag_gutachter_gremien/antragstellende/nachnutzung_forschungsdaten/index.html#anker62194854; ein *Framework for Discipline-specific Research Data Management* bietet *Science Europe* seit 2018 an unter: https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf.

| 48 Vertreter einer Mehrzahl der Fachkonsortien haben ihre Absicht erklärt, gemeinsam an Querschnittsthemen zu arbeiten, um den Aufbau einer Nationalen Forschungsdateninfrastruktur zu begleiten und den nötigen Kulturwandel zu unterstützen, vgl. Bierwirth, M., et al.: Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung, 15. Juni 2020, http://doi.org/10.5281/zenodo.3895209.

| 49 Die sogenannte Allianz der Wissenschaftsorganisationen stimmt sich seit 2008 bereits in ihrer Schwerpunktinitiative „Digitale Information“ ab (https://www.allianzinitiative.de/) und bietet den Wissenschaftseinrichtungen in Deutschland seitdem vielfältige Informationen und Positionen zu verschiedenen Herausforderungen an.

Hochschulen und Forschungseinrichtungen bietet sich durch datenintensive Forschung – abhängig von Forschungsanteil, Forschungsschwerpunkten und institutionellen Vorerfahrungen – die Chance der Profilbildung: Vorreiter können national und international standardbildend wirken. Solche Positionierungen können einzeln, aber auch in der Kooperation und in Netzwerken/Verbünden von wissenschaftlichen Einrichtungen erfolgen (z. B. Digitale Hochschule NRW, NFDI-Konsortien, NHR-Verbund der NHR-Zentren |[50]). Dazu ist auf eine rechtssichere Ausgestaltung der Kooperationen zu achten. Bisher haben insbesondere Einrichtungen der außerhochschulischen Forschung diese Chancen durch institutsübergreifende Abstimmung der Dachorganisationen bzw. internationaler infrastrukturbezogener Fachgemeinschaften genutzt (z. B. Projekte der Helmholtz-Gemeinschaft wie *Helmholtz-Inkubator Information & Data Science* oder gemeinschaftsweite Services rund um Künstliche Intelligenz).

II.1.b Einrichtungsinterner Rahmen datenintensiver Forschung

Hochschulen und Forschungseinrichtungen müssen interne Vereinbarungen darüber treffen, welche Anforderungen an datenintensive Forschung in ihrem Regelungsbereich gelten und wie das Erreichen dieser Anforderungen organisatorisch und mit den nötigen Ressourcen unterstützt werden kann. Dies darf nicht mit einer Abwertung der Forschungsfragen und -methodik nicht-datenintensiver Forschung einhergehen. Dabei ist eine proaktive Positionierung der Einrichtungen wichtig, ergänzend zu jenen Aufgaben, die ihnen als Durchsetzungsinstanzen des 2019 aktualisierten Kodex der Deutschen Forschungsgemeinschaft (DFG) mit Leitlinien zur Sicherung guter wissenschaftlicher Praxis ohnehin obliegen. |[51] Wissenschaftlerinnen und Wissenschaftler müssen die Erwartungen ihrer Heimateinrichtungen kennen, was lokal ausgeprägte Standards des Forschungsdatenmanagements einschließlich Anforderungen an die Datensicherheit, Dokumentationspflichten, den erwarteten Umgang mit Software, Kooperationsregeln, aber auch Konditionen der Datenmitnahme oder den Zugriff auf die Daten bei Einrichtungswechseln betrifft.

II.1.c Lokale Beratungsangebote zum Rahmen datenintensiver Forschung

Datenintensiv forschende Wissenschaftlerinnen und Wissenschaftler benötigen neben Beratung zur Durchführung von Datenanalysen auch Unterstützung bei der Einhaltung von allgemeinen rechtlichen und speziellen professionsethischen Regelungen sowie Richtlinien ihrer Heimateinrichtung einschließlich Satzungs- und Vertragsgestaltung. Hierzu gehört insbesondere der Umgang mit personenbezogenen Daten, was erforderliche Anonymisierungen oder Pseudo-

| 50 NHR: Nationales Hochleistungsrechnen (https://www.nhr-gs.de/).

| 51 Vgl. Kapitel B.II.2.

nymisierungen, die Ausgestaltung von Verfahrensanforderungen und nationalen und internationalen Kooperationen sowie besondere ethische Anforderungen betrifft. Je nach Inhalt und Aussagekraft der Daten, z. B. im Gesundheitsbereich, müssen besondere Vorkehrungen getroffen werden. Zudem muss Datensicherheit im Sinne eines Schutzes vor unberechtigtem Zugang, Veränderung und Manipulation gewährleistet sein. Zur angemessenen und rechtssicheren Handhabung bedarf es fundierter Beratung und einfach zugänglicher Informationsressourcen. Informiert werden muss zu Anforderungen an die Qualität und Sicherheit der Daten, zum Umgang mit offengelegten und nicht offengelegten Daten aus Forschungsprojekten einrichtungsintern, aber auch in Kooperationen mit Partnern aus Wissenschaft und Wirtschaft, zur Aufbewahrung oder Löschung und zur Datenmitnahme mit Blick auf Institutionenwechsel von Wissenschaftlerinnen und Wissenschaftlern. Hochschulen und Forschungseinrichtungen werden bei der Bereitstellung der Informationen und Beratung vielfach an bestehende Strukturen anknüpfen können, müssen aber Erweiterungen ihres Angebots prüfen, etwa was forschungsethische Regelungen und neuere Entwicklungen der Rechtsprechung betrifft.

II.1.d Spezifischer Ressourcenbedarf für datenintensive Forschung

Datenintensive Forschung erfordert eine Vielzahl von Tätigkeiten, die in einer Zeitkonkurrenz mit anderen Aufgaben stehen. Ein erheblicher Teil dieses Aufwandes entsteht nicht projektbezogen, sondern muss langfristig und nachhaltig gedeckt werden. Dies darf nicht zu Lasten anderer Aufgaben der Hochschulen und Forschungseinrichtungen gehen. Erforderlich sind daher mehr dauerhaft finanzierte, spezialisierte Stellen in den wissenschaftlichen Einrichtungen für die Bereiche Datenanalyse und Forschungsdatenmanagement, wobei sich disziplinäre Bedarfe unterscheiden können. Datenintensive Forschung verursacht zudem Kosten für aufwändige Analysemethoden und qualitätsgesicherte, nachhaltig nutzbare Daten in institutionellen und fachspezifischen Repositorien sowie umfangreichen Datensammlungen, die bei der Zuteilung von Ressourcen anerkannt und beachtet werden müssen. Finanzierungsbedarfe sind in bereits laufenden Projekten datenintensiver Forschung oftmals nicht ausreichend berücksichtigt und werden auf 5 bis 15 % der Gesamtkosten geschätzt. |[52] Auch bei der

| 52 Die Schätzungen zu den Kosten des Datenmanagements variieren abhängig davon, welche Teilaufgaben im Einzelnen eingerechnet werden. Hohe Schätzung etwa in der Studie des *Joint Information Systems Committee* (JISC) von Beagrie, N.; Lavoie, B.; Woollard, M.: *Keeping Research Data Safe 2*, 30.04.2010, http://repository.essex.ac.uk/2147/1/keepingresearchdatasafe2.pdf. Die Kommission Zukunft der Informationsinfrastruktur ging 2011 aus von „mittelfristig etwa 5 % bis 10 % der Forschungskosten zusätzlich für nachhaltige ‚Datenbereitstellung'", vgl. Gesamtkonzept für die Informationsinfrastruktur in Deutschland. Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder, April 2011, S. 44. Niedrigere Schätzungen zuletzt durch eine Expertengruppe der EU-Kommission, die 2016 schätzte „*on average about 5 % of research expenditure should be spent on properly managing and stewarding data*", vgl. *Realising the European Open Science Cloud: First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud*, Brüssel 2016, S. 19,

Anschubfinanzierung neuer Projekte muss ein datenspezifischer Bedarf ermittelt und berücksichtigt werden. Wichtig ist es, entsprechende Kosten zu überprüfen und einrichtungsintern Transparenz darüber herzustellen. Die Verteilung der innerhalb wissenschaftlicher Einrichtungen verfügbaren Ressourcen zwischen Forschungseinheiten sowie Infrastruktur- und Serviceeinrichtungen ist gegebenenfalls anzupassen, wobei dann auf die Themen- und Methodenvielfalt mit Blick auf verschiedene datenintensive Ansätze als auch unterschiedliche Formen datenintensiver und nicht-datenintensiver Forschung zu achten ist (dazu Leitlinie 2).

II.1.e Qualifizierung und neue Tätigkeitsprofile für datenintensive Forschung

Hochschulen müssen darauf reagieren, dass mit der zunehmenden Datenverfügbarkeit und der Verbreitung neuer Analysemethoden entsprechende Kompetenzen in vielen Fachgemeinschaften ebenso wie außerhalb der Wissenschaft stark nachgefragt werden. Entsprechend werden bereits vielfach Bemühungen unternommen, Studiengänge um Grundkompetenzen zum Datenverstehen (Data Literacy) zu erweitern und auch neue Studiengänge mit der Datenanalyse im Mittelpunkt anzulegen. Erweiterte Curricula sollten eine Sensibilisierung für gesellschaftliche, rechtliche und ethische Rahmen- und Grenzbedingungen datenintensiver Forschung unterstützen. In dem sich schnell verändernden Feld muss mit der kurzen Halbwertszeit der Lehrinhalte umgegangen werden. Wünschenswert ist vielfach ein Team- oder Co-Teaching durch Methodenwissenschaftlerinnen und -wissenschaftler („Data Scientists“ im engeren Sinne) mit Fachwissenschaftlerinnen und -wissenschaftlern (z. B. Computational Engineers). Dem stehen allerdings häufig kapazitätsbezogene Vorbehalte entgegen, die bei interdisziplinären Studiengängen generell auftreten. | [53]

Der Weiterbildungsbedarf von forschungsunterstützendem Personal und von Wissenschaftlerinnen und Wissenschaftlern, die nicht im datenintensiven Forschen sozialisiert wurden, ist dringend zu decken. Hier müssen die Bemühungen verstärkt werden, die Methoden- und Interpretationskompetenz der eigenen Wissenschaftlerinnen und Wissenschaftler auszubauen sowie Standards zur

https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf. Diese Größenordnung gibt auch die *Research Data Alliance* an, in: *The Data Harvest: How sharing research data can yield knowledge, jobs and growth. A Special Report by RDA Europe*, 2014, S. 33, https://rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html. Noch niedriger schätzt eine PwC-Studie für die Generaldirektion Forschung der EU-Kommission die Kosten mit 2,5 % ein und sieht demgegenüber enorme Einsparungen und Effizienzgewinne an anderer Stelle, vgl. *Cost-Benefit analysis for FAIR research data. Cost of not having FAIR research data*, Brüssel 2018: DOI: 10.2777/02999, https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en. Aktuell zu Kostentypen und Budgetkalkulation vgl. https://www.forschungsdaten.info/themen/planen-und-strukturieren/fdm-budgetplanung/.

| [53] Zu spezifischen Herausforderungen in interdisziplinären Studiengängen siehe Wissenschaftsrat: Wissenschaft im Spannungsfeld von Disziplinarität und Interdisziplinarität | Positionspapier (Drs. 8694-20), Köln Oktober 2020, https://www.wissenschaftsrat.de/download/2020/8694-20.pdf.

Statistik und in der wissenschaftsinternen Softwareentwicklung und -nutzung weiterzuentwickeln und auch in speziellen Austauschforen, etwa im Zuge des Infrastrukturausbaus von NFDI-Konsortien, EOSC und NHR-Verbund, zu verbreiten. Zusätzlich sind neue Tätigkeitsprofile und Laufbahnen für Personal im Forschungs- und Infrastrukturbereich zu schaffen und stabil zu verankern. Entsprechende Stellen müssen nicht zwangsläufig professoral sein, sollten aber über professionelle und wissenschaftliche Autonomie verfügen. | [54]

II.1.f Anreize für qualitätsvolle Datenarbeit

Hochschulen und Forschungseinrichtungen sollten datenintensiv forschende Wissenschaftlerinnen und Wissenschaftler aktiv unterstützen, indem sie qualitativ hochwertige Datenarbeit über die Bereitstellung der erforderlichen Ressourcen für eine systematische Verankerung von Datenverarbeitungsprozessen hinaus wertschätzen. Dies kann von der Ergänzung institutioneller Leitbilder und Strategien, über die Aufnahme entsprechender Anforderungen in Stellenausschreibungen und Berufungsverfahren bis hin zur Auszeichnung und Prämierung in der internen leistungsorientierten Mittelvergabe reichen. | [55]

II.1.g Verbindung von Forschungs- und Infrastrukturkompetenzen

Datenintensive Wissenschaft benötigt unterschiedliche Typen von Infrastruktur und Diensten: disziplinübergreifende Services mit Basisdiensten wie Rechenkapazitäten, Speicher, Datenpublikation, technische und methodische Beratung, zusätzlich aber auch forschungsfeldbezogene Services für spezifische Analysemethoden, Editions- und Visualisierungswerkzeuge sowie Datenkuratierung. Die Organisation beider Infrastrukturtypen kann in wissenschaftlichen Einrichtungen unterschiedlich erfolgen, sowohl in gemeinsamen Organisationsstrukturen von Zentren für Kommunikation und Information (Rechenzentren) und Bibliotheken als auch in virtuellen Kooperationen verschiedener Organisationseinheiten. Wichtig ist, dass bei der Übernahme fach- und forschungsnaher Services enge Kooperationen gesichert sind und die entsprechenden Expertinnen und Experten aus dem Infrastrukturbereich mit Wissenschaftlerinnen und Wissenschaftlern aus der Forschungspraxis in direktem Austausch stehen. Nur so können Infrastruktur- und Serviceangebote für die Forschung passgenau und einfach nutzbar (*convenient*) entwickelt und angeboten werden. Daher kommen

| 54 Zum Thema Kompetenzaufbau und mit Vorschlägen zu dessen Organisation umfassend Rat für Informationsinfrastrukturen: Digitale Kompetenzen – dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft, Göttingen 2019, http://www.rfii.de/?p=3883.

| 55 Vorschläge für weitere Anreize zur Stärkung der Datenqualität in wissenschaftlichen Einrichtungen hat zuletzt der Rat für Informationsinfrastrukturen unterbreitet in Rat für Informationsinfrastrukturen: Herausforderung Datenqualität. Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage, Göttingen 2019, http://www.rfii.de/?p=4043, S. 89 ff. Als Beispiel für die Honorierung der Datenbereitstellung vgl. Berliner Institut für Gesundheitsforschung mit QUEST-Preis für die Wiederverwendung von Daten, vgl. https://www.bihealth.org/de/forschung/quest-center/ausschreibungen-und-preise/quest-ausschreibungen-und-preise/offene-ausschreibungen-quest.

Orten und Personen zur Verbindung von Forschungs- mit Infrastrukturperspektiven vor dem Hintergrund des durch datenintensive Forschung verursachten Wandels in den Wissenschaften große Bedeutung zu. Um Interoperabilität zu sichern und Kosteneinsparungen durch Skaleneffekte zu ermöglichen, ist ein institutionenübergreifender Austausch über Lösungsansätze anzustreben, etwa in den künftigen NFDI-Konsortien oder in Forschungsverbünden.

II.1.h Zusätzliche Forschungskapazitäten und Data Science Center

Viele Einrichtungen werden sich im Rahmen ihrer Profilbildung entscheiden, Bereiche datenintensiver Forschung personell zu verstärken. Förderprogramme des Bundes wie auch vieler Länder unterstützen sie darin. |[56] Neue Professuren werden nicht nur in der Informatik, Mathematik und Statistik, sondern insbesondere auch in Fachdisziplinen bis hin zu Ökonomie, Rechtswissenschaften und Philosophie beheimatet sein. In Abhängigkeit von der strategischen Zielsetzung einer Hochschule können vereinzelt neue Professuren eingerichtet werden oder eine Gruppe von Professuren – koordiniert über Fachbereichs- bzw. Fakultätsgrenzen hinweg.

Verschiedene Optionen sind denkbar, um neue Forschungsstrukturen für datenintensive Forschung anzulegen. Aus heutiger Sicht erscheint die Einrichtung von Data Science Centern an einzelnen Universitäten oder auch für mehrere Universitäten gemeinsam erfolgversprechend, die den datenintensiv Forschenden aus verschiedenen Fachdisziplinen einen identifizierbaren Ort mit interdisziplinärem Charakter geben und Lehrangebote machen können (z. B. gemeinsame Studiengänge, Promotionsprogramme, Kolloquien, Tagungen und weitere Veranstaltungen). Solche Data Science Center können Methodenkompetenz aus Informatik, Mathematik und Statistik mit Anwendungen aus verschiedensten Fächern, die den Stärken der jeweiligen Universität entsprechen, zusammenbringen. Als dritte Säule solcher Center ist reflexive Kompetenz aus Disziplinen wie Rechtswissenschaft, Psychologie, Soziologie und Philosophie erforderlich. Entsprechende Strukturen können zugleich einen Anreiz für die Gewinnung von Personal in einem hochkompetitiven Umfeld bieten. |[57]

| 56 Z. B. Niedersachsen, wo 50 „Digitalisierungsprofessuren“ dauerhaft mit zusätzlichen Mitteln an den Hochschulen gefördert werden nach einem wettbewerblichen Auswahlverfahren. Das Land erwartet zugleich nachweisbare eigene Anstrengungen, vgl. zur Ausschreibung https://www.mwk.niedersachsen.de/download/140217 und zur Bekanntmachung der Ergebnisse https://www.mwk.niedersachsen.de/startseite/aktuelles/presseinformationen/acht-hochschulen-erfolgreich-bei-bewerbung-um-geforderte-digitalisierungsprofessuren-182518.html.

| 57 Zu spezifischen Herausforderungen an die Informatik in diesem Kontext äußert sich der Wissenschaftsrat in seinen Empfehlungen: Perspektiven der Informatik in Deutschland, Köln 2020, https://www.wissenschaftsrat.de/download/2020/8675-20.pdf.

Bund und Länder und die von ihnen getragene öffentliche Forschungsförderung haben im vergangenen Jahrzehnt große Vorhaben der Infrastrukturförderung insbesondere beim Hoch- und Höchstleistungsrechnen sowie für eine Nationale Forschungsdateninfrastruktur auf den Weg gebracht. Die kontinuierliche Aktualisierung und Weiterentwicklung dieser Infrastrukturen wie auch der Hard- und Software aller öffentlich finanzierten wissenschaftlichen Einrichtungen bleiben unverzichtbar. In jüngster Zeit wurde daneben die Policy- und Strategieentwicklung weitergetrieben. Auch die Deutsche Forschungsgemeinschaft, die als wichtiger Akteur der Forschungsförderung den Kulturwandel vorantreiben kann, hat eine intensive Auseinandersetzung mit dem digitalen Wandel aufgenommen, ihr Förderportfolio geändert (z. B. SFB-INF-Projekte, Software-Förderung) und in der Aktualisierung des Kodex mit Leitlinien zur Sicherung guter wissenschaftlicher Praxis zentrale Fragen der Qualitätssicherung, der Nutzungsrechte, des Zugangs und der Archivierung von Forschungsdaten adressiert. |[58] Mit den vorliegenden Empfehlungen des Wissenschaftsrats werden ergänzende Bereiche in den Blick genommen. Damit Wissenschaftlerinnen und Wissenschaftler in Deutschland den Kulturwandel in der Wissenschaft mitgestalten und die neuen Möglichkeiten auf höchstem Niveau nutzen können, muss das Portfolio der verfügbaren Förderinstrumente durch Förderinstitutionen und Zuwendungsgeber überprüft und weiterentwickelt werden.

II.2.a Vernetzte Beratungsstrukturen zum Rahmen datenintensiver Forschung

Datenintensive Forschung unterliegt komplexen rechtlichen und ethischen Rahmenbedingungen für deren Berücksichtigung Wissenschaftlerinnen und Wissenschaftler auf Beratung und Unterstützung angewiesen sind. Entsprechende Angebote der einzelnen wissenschaftlichen Einrichtungen werden heute schon durch übergreifende ergänzt. |[59] Auf überregionaler Ebene ist besonders die Beratung zur nationalen und internationalen Rechtsanwendung für datenintensive Forschung einschließlich der Beobachtung der Rechtsprechung und der Umsetzungsfortschritte von Bedeutung, aber auch solche zur Forschungsethik. Die vorhandenen lokalen Angebote in Hochschulen und außerhochschulischen Forschungseinrichtungen gilt es mit den spezialisierten überregionalen Angeboten besser zu vernetzen und abhängig von der Nachfrage diese möglicherweise auch zu erweitern.

| 58 Deutsche Forschungsgemeinschaft: Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex, Bonn 2019, https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf.

| 59 Beispielsweise die Angebote zum Bereich Forschungsdatenmanagement über https://www.forschungsdaten.info/ oder der Forschungsstelle Recht des Vereins zur Förderung eines Deutschen Forschungsnetzes e. V. (DFN) unter https://www.dfn.de/rechtimdfn/.

Die potenziell unbeschränkte Wiedernutzung digitaler Daten verlangt danach, organisatorische und finanzielle Lösungen für das langfristige, nachhaltige Aufbewahren und Kuratieren von Daten stets mitzudenken und den daraus resultierenden Ressourcenbedarf einzuplanen. Strukturen für die Datenkuratierung und -speicherung, die sich nach und nach etablieren, müssen mit dem Ziel eines schlüssigen Gesamtsystems dynamische Anpassungen zulassen und dafür auch begleitend evaluiert werden. |[60] Laufende Kosten für eine effiziente und sichere Datenkuratierung, die von Personal- bis zu Energiekosten reichen und teils von Dritten übernommen werden können, müssen transparent gemacht und ihre Übernahme von den Heimatinstitutionen oder über die Projektfinanzierung geklärt werden. Die Förderer müssen daher prüfen, welche Anpassungen bei der Berücksichtigung von Datenkuratierungskosten notwendig sind, um die Gesamtsystemausgaben im Zusammenhang mit datenintensiver Forschung adäquat abzubilden. Sie sollten Anstrengungen unternehmen, die Mehrkosten für Datenkuratierung mit zusätzlichen Mitteln zu decken, damit der Ausbau datenintensiver Forschung nicht zu Lasten anderer Forschungsformen geht (vgl. Leitlinie 2). Zu den verschiedenen Optionen zählen eine Aufstockung oder Ergänzung der institutionellen Förderung oder der projektbezogenen Overheads (Programmpauschalen) und auch die Verlängerung einer Projektförderung durch zweckgebundene Zusatzmodule, sofern generierte Daten nachhaltig nutzbar gemacht werden sollen.

II.2.c Experimentierräume in der Forschungsförderung

Datenintensive Forschung zeichnet sich durch die hohe Geschwindigkeit aus, in der sich Methoden verändern. Sie braucht deshalb neben enger Abstimmung zwischen projektfinanzierter Forschung und langfristig, teils institutionell finanzierten Dienstleistungen und Infrastrukturen auch mehr Flexibilität und Experimentierräume in der Projektförderung. Während im Infrastrukturbereich inzwischen stabilere Förderstrukturen geschaffen werden, sollte in der Projektförderung die wachsende Nachfrage nach kurzfristig verfügbaren und agilen Förderelementen, in denen auch Kurswechsel, ungeplante Ergebnisse oder ein Scheitern in Kauf genommen werden darf, von Förderorganisationen und Zuwendungsgebern adressiert werden. Derartige Fördermöglichkeiten können beispielsweise den testweisen Zugang zu bestimmten Daten oder den kurzfristigen Transfer von Methoden unterstützen (z. B. Personalkosten für Programmieraufgaben), um schneller Ergebnisse in neuen Anwendungsfeldern zu erreichen.

| 60 Entsprechend hat die Gemeinsame Wissenschaftskonferenz von Bund und Ländern (GWK) bereits eine Evaluation der NFDI-Aufbauphase durch den Wissenschaftsrat vorgesehen, vgl. GWK: Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI), Bonn 28.11.2018, § 13.

Datenintensive Forschung ist häufig interdisziplinär und kann Brücken zwischen bislang getrennten Forschungsfeldern schlagen. Interdisziplinarität umfasst dabei auch die Verbindung mit nicht-datenintensiver Forschung, etwa bei der Integration gesellschaftlicher, rechtlicher und ethischer Grenzen und Rahmenbedingungen in Forschungsfragen. Um diese Möglichkeiten umfassend zu nutzen und die Koordinationserfordernisse interdisziplinärer Verbünde („Architektenleistung“) in diesen Bereichen anzuerkennen, sollten neuartige Formate für Anbahnungsfinanzierungen datenintensiver und über Disziplinengrenzen hinwegreichender Forschungsprojekte entwickelt werden. Solche Projekte könnten zunächst mit einer kurzen Explorationsphase gefördert werden (z. B. 6–12 Monate), bevor eine längere Förderung nach Zielerreichung operationalisierter Meilensteine freigegeben werden könnte. |[61]

II.2.e Datensammlungen als Infrastrukturen

Die Probleme reiner Projektförderung im Infrastrukturbereich insbesondere mit Blick auf die Stabilität und Nachhaltigkeit der Strukturen und Dienstleistungen sind inzwischen erkannt worden und werden an verschiedenen Stellen bereits korrigiert, etwa beim Aufbau einer Nationalen Forschungsdateninfrastruktur, wo eine langfristige Finanzierung und entsprechende Strukturen vorbereitet werden, die allerdings eine stabile Finanzierung der Beschaffung und des Betriebs von Hardware ihrerseits voraussetzen. Auch jenseits der bisherigen Infrastrukturförderung sollte die Notwendigkeit weiterer Reaktionen der Forschungsförderer und Zuwendungsgeber auf Bedarfe datenintensiver Forschung geprüft werden. Hierzu zählte der Aufbau solcher Datensammlungen von Fachgemeinschaften, der infrastrukturähnlich zu verstehen und in entsprechender Weise zu fördern ist, um auch außerhalb der NFDI-Konsortien ein nachhaltiges und stabiles Angebot an eine Fachgemeinschaft und weitere Nutzende unterbreiten zu können. Mit Blick auf Effizienz und Nachhaltigkeit gilt es zudem Synergien zu heben sowie eine engere Abstimmung des Förderhandelns etwa für die NFDI-Konsortien und die NHR-Zentren im Zuge vorgesehener Zwischenevaluationen zu überprüfen.

II.2.f Nachhaltige Softwareentwicklung

Die modularisierte Entwicklung und Archivierung offener Forschungssoftware sollte verstärkt und auf neue Weise gefördert werden, um Abhängigkeiten von privaten Anbietern und Kompromisse hinsichtlich der Standards zu vermeiden.

| 61 Vgl. Wissenschaftsrat: Wissenschaft im Spannungsfeld von Disziplinarität und Interdisziplinarität | Positionspapier (Drs. 8694-20), Köln Oktober 2020, S. 56 f., https://www.wissenschaftsrat.de/download/2020/8694-20.pdf.

Dies sollte den Unterstützungsbedarf neuer Foren zum Austausch über Forschungssoftware-Entwicklung einbeziehen. Was die nachhaltige Nutzbarmachung von Forschungssoftware über ihren ursprünglichen Verwendungskontext hinaus betrifft, macht die Deutsche Forschungsgemeinschaft bereits ein neues Förderangebot. |[62] Dieses Angebot ist, wie entsprechende andere Angebote internationaler Förderer, zu begrüßen und sollte angesichts des großen Bedarfs eine Ausweitung erfahren. Ferner sollte beobachtet werden, ob der Übergang in eine nachhaltige Bereitstellung und Nutzbarhaltung nach der DFG-Förderung gelingt.

II.2.g Begutachtungsprozesse und Förderauflagen

Mit den Veränderungen durch datenintensive Forschung sind auch die Begutachtungsprozesse in der Forschungsförderung, für institutionelle Evaluationen und bei der Leistungsbewertung von einzelnen Personen auf den Prüfstand zu stellen. So reicht es nicht mehr aus, Forschungsdatenmanagementpläne anzufordern und formal zu prüfen, sondern diese müssen auch inhaltlich und in der Umsetzung bewertet werden können. Datenintensive Forschung ist häufig interdisziplinär und stellt daher besondere Anforderungen an die Zusammensetzung von Begutachtungsgremien, wie etwa der Begutachtungsprozess zu den NFDI-Konsortien zeigt. |[63] Begutachtungen müssen den Transfer von Methodenentwicklungen zwischen verschiedenen Forschungsfeldern unterstützen und helfen, Fehlinvestitionen in Insellösungen zu vermeiden. Datenintensive Forschung in neuen Kooperationsformen und Verbünden bringt auch neue Herausforderungen für die Leistungszuordnung mit sich, wie die zunehmenden Forderungen nach einer Differenzierung von Autorschaft und Beiträgerschaft zeigen. Förderer und Zuwendungsgeber müssen sich außerdem bewusst sein, dass ihre Förderauflagen und entsprechende Informationsabfragen bereits handlungsleitend und verhaltensändernd wirken können sowie einen Kulturwandel beim Teilen von Daten in der Wissenschaft und ihrer kooperativen Bearbeitung wirksam unterstützen können.

II.2.h Intensivierung der Wissenschaftskommunikation

Die Wissenschaftskommunikation zu datenintensiver Forschung bedarf zusätzlicher Unterstützung, um die Funktionsweise und Folgen datenintensiver Forschung auf innovative Art und Weise zu vermitteln. So können interaktive For-

| 62 Vgl. Information für die Wissenschaft Nr. 44 | 18. Juni 2019: Qualitätssicherung von Forschungssoftware durch ihre nachhaltige Nutzbarmachung, https://www.dfg.de/foerderung/info_wissenschaft/2019/info_wissenschaft_19_44/index.html.

| 63 Zu Bewertung und Begutachtung interdisziplinärer Forschung vgl. Wissenschaftsrat: Wissenschaft im Spannungsfeld von Disziplinarität und Interdisziplinarität | Positionspapier (Drs. 8694-20), Köln Oktober 2020, https://www.wissenschaftsrat.de/download/2020/8694-20.pdf, S. 67 f.

mate es Laien ermöglichen, die Folgen bestimmter methodischer Entscheidungen besser nachzuvollziehen. Zudem bietet sich in besonderer Weise die Chance, Bürgerinnen und Bürger im Rahmen von Citizen Science-Projekten zu beteiligen, beispielsweise wenn diese sich bei der Nutzung von Wearables |[64] an der Schnittstelle von technischen Entwicklungen und ihrer auch wissenschaftlichen Nutzung bewegen. Daher sollten in größeren Forschungsprojekten entsprechende Zusatzmodule und Mindestbudgets vorgesehen werden. Daneben ist zu prüfen, ob sogenannter *Public Service Journalism* zu datenintensiver Forschung gefördert werden kann, wobei sich hier auch ein Feld für forschungsfördernde Stiftungen auftut. |[65]

II.2.i Bedarfsprüfung zusätzlicher Forschungskapazitäten

Bund und Länder haben in jüngster Zeit erhebliche Investitionen in den Aufbau neuer Infrastrukturen und Forschungskapazitäten für datenintensive Forschung getätigt. Dennoch wird der Bedarf mit den bereits laufenden Programmen mittelfristig voraussichtlich nicht gedeckt werden können. Vor diesem Hintergrund sollten sich Bund und Länder Handlungsspielräume für eine weitere Investitionsstufe sichern und in zwei bis drei Jahren eine Zwischenprüfung des erreichten Standes vornehmen. Geprüft werden sollte, ob mit den Fördervorhaben etwa im Bereich Big Data und Künstliche Intelligenz die Förderung ausreichend ausgestaltet wurde und welcher Bedarf an zusätzlichen oder weiter zu fokussierenden Kompetenzzentren oberhalb einzelner Hochschulen und Forschungseinrichtungen zu speziellen Themen datenintensiver Forschung besteht. Im Zuge dessen wird auch zu evaluieren sein, ob eine noch engere Vernetzung und Koordination der bereits geförderten Vorhaben und Zentren auf nationaler oder europäischer Ebene sinnvoll sind.

| 64 Wearables sind tragbare Computer bzw. mit Computern oder Smartphones vernetzte Sensoren, die in Kleidungsstücke oder Accessoires integriert sind.

| 65 Beispiele aus diesem Bereich etwa international https://www.quantamagazine.org/ sowie in Deutschland die Aktivitäten des *Science Media Center*, z. B.: https://www.sciencemediacenter.de/alle-angebote/fact-sheet/details/news/ki-textgeneratoren-neue-entwicklungen-und-moegliche-gesellschaftliche-folgen/.

C. Nachwort: Forschungsdaten und COVID-19

BEOBACHTUNGEN IM HERBST 2020

Der Umgang mit der *Coronavirus Disease 2019* (COVID-19), die sich seit Ende 2019 von China aus weltweit verbreitet hat, illustriert paradigmatisch Herausforderungen der Bereitstellung und Verwendung von Forschungsdaten sowie den mit neuen Methoden einhergehenden Wandel in den Wissenschaften auf eindrucksvolle Weise. Forschung war und ist für den Umgang mit der Pandemie zentral: für die Identifizierung des Coronavirus SARS-CoV-2; für die Erforschung seiner Struktur, seiner Evolution und der Mechanismen der Ansteckung und Pathogenese; für die Entwicklung von Diagnostika, die Beobachtung und die Analyse der globalen COVID-19-Pandemie; für die Entwicklung und Testung von Therapien und prophylaktischen – pharmazeutischen wie nicht-pharmazeutischen – Maßnahmen; und für das Verständnis und die Gestaltung der gesellschaftlichen Folgen. Die globale Bedeutung der Pandemie spiegelt sich in einem enormen Wachstum fokussierter wissenschaftlicher Aktivität wider, das schon im ersten Halbjahr des Jahres 2020 zu mehr als 63 000 COVID-19-bezogenen wissenschaftlichen Publikationen führte. |[66] Wie sich diese Forschungsaktivitäten und damit der Kenntnisstand über die Pandemie entwickeln, hing und hängt weiter maßgeblich von der Verfügbarkeit, der Qualität, dem Teilen und Zusammenführen sowie der Analyse und Interpretation von Forschungsdaten ab. Auch wenn Vieles von Öffentlichkeit und Politik als neuartig empfunden wurde, hat der wissenschaftliche Umgang mit der Pandemie bislang vor allem Trends sichtbarer gemacht und dank des Einsatzes kurzfristig umgewidmeter wie auch zusätzlicher Ressourcen massiv verstärkt, die sich bereits vorher abzeichneten.

Zu den auffallendsten Phänomenen während der Corona-Pandemie gehörte die Geschwindigkeit, mit der Forschungsergebnisse verbreitet, von anderen Wissen-

| 66 Porter, S. J.; Hook, D.: *How COVID-19 is Changing Research Culture. Digital Research Report, Digital Science, June* 2020, https://doi.org/10.6084/m9.figshare.12383267. Zahlen aktualisiert nach https://covid-19.dimensions.ai, Stand 26.06.2020.

schaftlerinnen und Wissenschaftlern aufgegriffen sowie bei der Bearbeitung weiterer Forschungsfragen verwendet wurden. Während diese Beschleunigung im Publikationsbereich durch die große Bedeutung von Preprints, also von noch nicht referierten Manuskripten, auch für die Öffentlichkeit besonders sichtbar und teils zu einer Herausforderung für die Wissenschaftskommunikation wurde (siehe unten), war das Teilen von Forschungsdaten für diesen Geschwindigkeitsgewinn mindestens ebenso entscheidend (vgl. Leitlinie 1: Teilen und Kooperieren, S. 37). Auch wenn kritisiert wird, dass China der Weltgesundheitsorganisation (WHO) seine Erkenntnisse über die neue Erkrankung zunächst vorenthalten hat, |[67] ist doch zu konstatieren, dass die ersten DNA-Sequenzdaten des in Wuhan (zunächst unter dem Namen 2019-nCoV) neu isolierten Coronavirus der weltweiten Forschungsgemeinschaft dank existierender Datenbanken für virale Genome rasch zur Verfügung standen. |[68] Innerhalb weniger Tage konnten Forscherinnen und Forscher des Deutschen Zentrums für Infektionsforschung auf Basis dieser Daten einen zuverlässigen diagnostischen Test für das Virus veröffentlichen. |[69] Plattformen, die schon bei früheren, weniger gravierenden Epidemien zum Einsatz gekommen waren, wurden sofort genutzt, um DNA-Sequenzen von verschiedenen Infektionsherden zusammenzuführen, phylogenetische Analysen vorzunehmen sowie so die Ausbreitung und Evolution des Virus verfolgen zu können. |[70] Zahlreiche Organisationen weltweit schlossen sich Ende Januar 2020 einer Initiative des Wellcome Trust an und vereinbarten, Forschungsdaten und Forschungsergebnisse, die zum Verständnis und zur Bekämpfung der Pandemie beitragen könnten, rasch und offen zu teilen. |[71] Große Internetkonzerne wie auch unabhängige Institute schufen verschiedene, mit Künstlicher Intelligenz ständig aktualisierte, strukturierte und annotierte Portale, um den Zugang nicht nur zu wissenschaftlichen Publikationen, sondern auch zu den zugrunde liegenden Forschungsdaten mit Bezug auf COVID-19 zu erleichtern. |[72] Vielfach bestand die Erwartung, dass Widerstände gegen die Offenlegung von Forschungsdaten überwunden werden können, wenn konkrete

| 67 *Associated Press* (AP): *China delayed releasing coronavirus info, frustrating WHO*, 3. Juni 2020. https://apnews.com/3c061794970661042b18d5aeaaed9fae.

| 68 Ein erstes Genom wurde am 5. Januar 2020 auf GenBank eingereicht und der weltweiten Forschergemeinschaft am 10. Januar auf https://virological.org angekündigt, vier weitere Genome folgten am 12. Januar auf der Plattform der *Global Initiative on Sharing All Influenza Data* (GISAID). Nahezu zeitgleich wurden Genomdaten des neu isolierten Virus auch beim *China National Microbiological Data Center* sowie der *China National GeneBank* hinterlegt. Lu, R. et al.: *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding*, in: *The Lancet* 395 (2020) 10224, S. 565–574, veröffentlicht 29. Januar 2020, https://doi.org/10.1016/S0140-6736(20)30251-8.

| 69 Corman, V. M. et al.: *Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PC*, in: *Eurosurveillance* 25 (2020) 3, https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.3.2000045.

| 70 Erste Analysen erschienen ab dem 14. Januar 2020 auf *Nextstrain*, vgl. https://nextstrain.org/ncov/global.

| 71 https://wellcome.ac.uk/coronavirus-covid-19/open-data.

| 72 https://www.semanticscholar.org/cord19; https://covid19-research-explorer.appspot.com/; https://www.kaggle.com/covid-19-contributions; https://beta.covid-19.openaire.eu/; https://covidscholar.org.

übergeordnete Ziele die Interessen einzelner Akteure überwiegen, und dass dies zu einer enormen Beschleunigung des Forschungsprozesses beitragen kann. Nichtsdestotrotz bleibt es eine Herausforderung, bessere Anreize für das frühzeitige Teilen qualitätsgesicherter und gut dokumentierter Daten zu schaffen. |[73]

Die Bedeutung des freien Flusses von Forschungsdaten bei der Bekämpfung der Pandemie wurde von der Politik frühzeitig erkannt. Die Schaffung einer Datenplattform für COVID-19-bezogene Forschungsdaten war Gegenstand des am 7. April 2020 veröffentlichten Aktionsplans „ERAvsCORONA" der Generaldirektion Forschung bei der Europäischen Kommission und trat Mitte April mit der Schaffung des COVID-19-Datenportals in Kraft. |[74] Über dieses Portal, das auf den Vorarbeiten für die EOSC aufbaut und zugleich ein wichtiger Anwendungsfall für das Gesamtprojekt ist, kann unter anderem auf Sequenzdaten, Daten zu Genexpressionsmustern, Proteinstrukturen sowie Erkenntnisse über mögliche Wirkstoffe und deren Ziele zugegriffen werden. Die parallel entstehende Taskforce COVID-19 des Europäischen Netzwerks Klinischer Forschungsinfrastrukturen (ECRIN) |[75] machte neben aktueller Forschungsliteratur lediglich Metadaten klinischer Versuchsreihen in einem Repositorium offen zugänglich, da klinische Daten typischerweise personenbezogene oder personenbeziehbare Daten beinhalten, die nur zweckgebunden und mit Einwilligung verarbeitet werden dürfen. Hier zeigt sich, dass Offenheit als Paradigma des Teilens von Forschungsdaten nicht überall anwendbar ist, sondern verschiedene Zugangsregime gleichberechtigt nebeneinander existieren können und müssen. Um die Zusammenführung und Nachnutzung von Patientendaten zu erleichtern, werden derzeit neue Formen der Einwilligungserklärung entwickelt, die – rechtssicher dokumentiert und nachverfolgbar – eine Zustimmung zur Nutzung für erweiterte, gleichwohl bestimmte Zwecke der öffentlichen Forschung ermöglichen sollen. In diesem Kontext ist aktuell der in der Medizininformatik-Initiative konzipierte Mustertext zur Patienteneinwilligung im Sinne eines *broad consent* zu nennen, der zwischenzeitig auch auf Zustimmung der Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder gestoßen ist, auch wenn noch unklar ist, ob dieser den Vorgaben der DSGVO vollständig genügt. |[76]

|[73] *RDA COVID-19 Working Group: RDA COVID-19. Recommendations and Guidelines on Data Sharing. Research Data Alliance*, 2020, https://doi.org/10.15497/rda00052.

|[74] *DG Research and Innovation: First "ERAvsCORONA" Action Plan. Short-term coordinated Research & Innovation actions*, Brüssel, 7. April 2020, https://ec.europa.eu/info/sites/info/files/research_and_innovation/research_by_area/documents/ec_rtd_era-vs-corona_0.pdf; https://www.covid19dataportal.org/.

|[75] https://ecrin.org/covid-19-taskforce.

|[76] https://www.medizininformatik-initiative.de/sites/default/files/2020-04/MII_AG-Consent_Einheitlicher-Mustertext_v1.6d.pdf; Datenschutzkonferenz: Beschluss der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder zu den Einwilligungsdokumenten der Medizininformatik-Initiative

Neben den rechtlichen Rahmenbedingungen waren für die Zusammenführung von Forschungsdaten zu COVID-19 – insbesondere jenseits der genetischen und molekularbiologischen Daten – auch Fragen der Standardisierung zu lösen. Durch die Medizininformatik-Initiative war in der deutschen Universitätsmedizin ein organisatorischer und konzeptioneller Rahmen für die Definition gemeinsamer Standards für klinische Daten bereits vorhanden. Darauf aufbauend definierten führende Akteure des Gesundheits- und des Gesundheitsforschungssystems, die sich mit Förderung des Bundesministeriums für Bildung und Forschung (BMBF) im Nationalen Netzwerk Universitätsmedizin zusammengeschlossen hatten, einen Kerndatensatz *German Corona Consensus* (GECCO) |[77] und etablierten die *Corona Component Standards-Initiative* (cocos) |[78], um eine möglichst weitgehende Interoperabilität von spezifischen, COVID-19-bezogenen klinischen Daten mit verschiedensten anderen, relevanten Daten auch über die Einrichtungen hinweg sicherzustellen. Nichtsdestotrotz blieben die wissenschaftliche Begleitung und Aufarbeitung des Verlaufs der Pandemie in Deutschland hinter der Sicherstellung der Krankenversorgung zurück, was nicht zuletzt auf Verzögerungen beim Austausch und der Zusammenführung klinischer Daten zurückgeführt wird. Eine Arbeitsgruppe des Forums Gesundheitsforschung erarbeitet derzeit allgemeine Handlungsempfehlungen zur Nutzbarmachung digitaler Daten für die Gesundheitsforschung. |[79]

Eine Herausforderung bleibt insbesondere die wissenschaftliche Analyse des Verlaufs der Pandemie anhand von epidemiologischen Daten. Solange sie nicht über Ergebnisse großer, repräsentativer Studien verfügt, ist die Forschung wesentlich auf amtliche Daten angewiesen, deren Erhebung und Zusammenführung anderen Berichtslogiken folgt. Positive Testergebnisse müssen den Gesundheitsämtern gemeldet werden, von wo sie spätestens am nächsten Arbeitstag elektronisch an die zuständige Landesbehörde und weiter an das Robert-Koch-Institut (RKI) als zuständiger Ressortforschungseinrichtung des Bundes übermittelt werden. Die Prozesse sind zwar weitgehend, aber nicht vollständig vereinheitlicht und führen zu leichten Asynchronitäten in der Datenübermittlung. |[80] Zugleich macht die Beschleunigung der Berichtsprozesse fallweise Nachbesserungen einzelner Daten notwendig. |[81] Nach wie vor problematisch ist auch,

des Bundesministeriums für Bildung und Forschung, 15. April 2020, https://datenschutzkonferenz-online.de/media/dskb/20200427_Beschluss_MII.pdf. Zur Einordnung dieses und alternativer Lösungsansätze vgl. auch Gutachten der Datenethikkommission der Bundesregierung, Berlin 2019, S. 126 f., https://datenethikkommission.de/.

| 77 https://www.bihealth.org/de/aktuell/deutschlandweite-standards-fuer-corona-daten/.

| 78 http://cocos.team/.

| 79 https://www.gesundheitsforschung-bmbf.de/de/forum-gesundheitsforschung-5787.php.

| 80 „In der aktuellen Lage übermitteln die meisten Ämter sogar täglich.“, Website des RKI: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Fallzahlen.html.

| 81 „In der aktuellen Krise werden die Daten des infektionsepidemiologischen Meldewesens zu COVID-19 mit einem möglichst geringen Zeitverzug publiziert. Dies trägt der sehr hohen Dynamik der Lage Rechnung.

dass nur positive Testergebnisse berichtspflichtig sind und die Rate positiver Tests, die für eine Beurteilung der epidemiologischen Lage wichtig wäre, nur indirekt erschlossen werden kann. Zwar melden die meisten Testlabors freiwillig die Testhäufigkeiten. Diese wöchentlich veröffentlichten Zahlen liefern nach Einschätzung des RKI lediglich Hinweise auf die Situation in den Laboren, erlauben jedoch „keine detaillierten Auswertungen oder Vergleiche mit den gemeldeten Fallzahlen". |[82]

Besonders sichtbar wurde das Fehlen einheitlicher Standards und Berichtsprozesse für epidemiologische Daten dadurch, dass die Zahlen zum Verlauf der Pandemie aus verschiedenen Ländern nur bedingt vergleichbar sind. |[83] Begrenzte Testkapazitäten, unterschiedliche Testregimes, erste Erkenntnisse über symptomlose Infektionsverläufe und in manchen Ländern auch Staatschefs, die öffentlich darüber räsonierten, ob die berichteten Infektionszahlen nicht durch Begrenzung des Testens gesenkt werden sollten, säten Zweifel an den Daten und führten zu intensiven Diskussionen über das Ausmaß der Dunkelziffern. Selbst der Versuch, COVID-19-bezogene Mortalitätsraten als Korrektiv zu verwenden, brachte keine endgültige Sicherheit, weil auch hier unterschiedliche Praktiken hinsichtlich der Differenzierung zwischen COVID-19-verursachter und COVID-19-korrelierter Mortalität bestanden. Zu denken gibt, dass für die Beschreibung des Verlaufs der Pandemie oftmals nicht auf amtliche, von der WHO auf Basis nationaler Statistiken erhobene Daten, sondern die aus unterschiedlichen Quellen teil-automatisiert zusammengetragenen epidemiologischen Daten der Johns Hopkins-Universität |[84] zurückgegriffen wird. Insgesamt zeigte die Pandemie nicht nur auf, dass Bemühungen um einheitliche Metadatenstandards und Anforderungen an die Dokumentation von Forschungsprozessen in unterschiedlichen wissenschaftlichen Gemeinschaften unterschiedlich weit gediehen waren; sie machte auch den Bedarf deutlich, Metadatenstandards ihrerseits so weit zu

Allerdings werden hierdurch zuweilen auch Daten vor Qualitätskontrollen und Validierungen veröffentlicht. Durch weitere Ermittlungen der Gesundheitsämter und Plausibilitätsprüfungen kann es zu Nachträgen oder Korrekturen kommen, was vereinzelt zu Abweichungen gegenüber den zuvor berichteten Daten führt.", Website des RKI, a. a. O.

| 82 RKI: Epidemiologisches Bulletin 34/2020, 20.08.2020, https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2020/Ausgaben/34_20.html.

| 83 „*Despite our need for evidence-based policies and medical decision-making, there is no international standard or coordinated system for collecting, documenting, and disseminating COVID-19 related data and metadata, making their use and reuse for timely epidemiological analysis challenging due to issues with documentation, interoperability, completeness, methodological heterogeneity, and data quality.*", *RDA COVID-19 Working Group: RDA COVID-19. Recommendations and Guidelines on Data Sharing. Research Data Alliance*, 2020, S. 37, https://doi.org/10.15497/rda00052.

| 84 Dong, E.; Du, H.; Gardner, L.: *An interactive web-based dashboard to track COVID-19 in real time*, in: *The Lancet. Infectious Diseases*, 20 (2020) 5, veröffentlicht 19. Februar 2020, https://doi.org/10.1016/S1473-3099(20)30120-1; https://arcg.is/0fHmTX.

standardisieren, dass die Interoperabilität über verschiedene Gemeinschaften hinweg gewährleistet werden kann. |[85]

Die starke Beschleunigung der Forschungsprozesse und die globale Bedeutung möglicher Resultate stellten die Qualitätssicherungsmechanismen des Wissenschaftssystems auf die Probe. Schnell wurde kritisiert, viele Studien beruhten auf zu geringen Fallzahlen, seien schlecht geplant oder würden nicht sorgfältig dokumentiert. |[86] Nahe lag zudem die Sorge, Zeitdruck würde zu einem vermehrten Vorkommen handwerklicher Fehler und Ungenauigkeiten beitragen, eine Sorge, die durch die große Bedeutung der Kommunikation über nicht begutachtete, auf Plattformen im Internet frei verfügbare Manuskripte (Preprints) noch bestärkt wurde. Die renommierte *MIT Press* startete im Sommer 2020 ein sogenanntes *Overlay Journal* mit dem Titel „RR:C19", das schnelle Reviews von Preprints in bekannten Repositorien veröffentlicht und damit einen Beitrag zur Qualitätssicherung der rasch anwachsenden Manuskriptliteratur leisten soll. |[87] Die Notwendigkeit, trotz aller Beschleunigung für stringente Qualitätssicherungsprozesse zu sorgen, wurde nicht zuletzt anhand einzelner Betrugsfälle offensichtlich. So kam es zu mindestens einem prominenten Fall, in dem in Kooperationen zwischen akademischen Forscherinnen und Forschern und einem Unternehmen Standards wissenschaftlicher Qualität und Integrität offenkundig verletzt wurden (vgl. Leitlinie 7: Wissenschaftliche Standards in Kooperationen, S. 45). Zwei Studien, in denen es um Nebenwirkungen der Behandlung von COVID-19-Patientinnen und -Patienten mit Hydrochloroquin oder Chloroquin ging und die in hoch angesehenen medizinischen Zeitschriften erschienen waren, |[88] mussten binnen weniger Tage zurückgezogen werden. |[89] Es hatte sich herausgestellt, dass die Beschreibung des angeblich verwendeten Datensatzes zahlreiche Ungereimtheiten enthielt. Eine Firma namens Surgisphere, die einem der Ko-Autoren gehörte, verweigerte den anderen Ko-Autoren wie auch den Gutachtenden und unabhängigen Prüfern der Verlage den Zugang zu den Rohdaten der Studie und zu den Verträgen, die angeblich mit den Kliniken geschlossen worden waren. Es zeigte sich, dass die Sammlung und Zusammenführung

| 85 Vgl. *RDA COVID-19 Working Group: RDA COVID-19. Recommendations and Guidelines on Data Sharing. Research Data Alliance*, 2020, S. 18 f., https://doi.org/10.15497/rda00052.

| 86 Glasziou, P. P.: *Waste in COVID-19 research*, in: *British Medical Journal* 2020;369:m1847, https://www.bmj.com/content/369/bmj.m1847.full.pdf; Alexander, P. E. et al.: *COVID-19 coronavirus research has overall low methodological quality thus far. Case in point for chloroquine/hydroxychloroquine*, in: *Journal of Clinical Epidemiology* 123 (2020), S. 120–126, https://doi.org/10.1016/j.jclinepi.2020.04.016.

| 87 https://rapidreviewscovid19.mitpress.mit.edu/.

| 88 Mehra, M. R. et al.: *RETRACTED*: *Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19. A multinational registry analysis*, published on May 22, 2020, at *The Lancet*. https://doi.org/10.1016/S0140-6736(20)31180-6; Mehra, M. R. et al.: *Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. This article has been retracted, published on May 1, 2020, at New England Journal of Medicine*, https://www.nejm.org/doi/full/10.1056/NEJMoa2007621.

| 89 Vgl. *Lancet, NEJM retract controversial COVID-19 studies based on Surgisphere data*, https://retractionwatch.com/2020/06/04/lancet-retracts-controversial-hydroxychloroquine-study/.

von Patientendaten auch Geschäftsfeld einer gewinnorientierten „Gesundheitsdatenindustrie“ geworden ist, die mit problematischen Interessenkonflikten behaftet ist. |[90] Angesichts dieser Umstände kamen Zweifel auf, ob die Zeitschriften über hinreichend stringente Regularien zum Umgang mit Forschungsdaten verfügten oder im Rahmen des Peer Review die Datenquellen nicht hätten besser überprüfen müssen. Auch unabhängig von Betrugsfällen bleibt zu fragen, ob alle Verfahrensbeteiligten – Ko-Autorinnen und -Autoren, Gutachtende sowie Herausgeberinnen und Herausgeber – über die nötige Datenkompetenz verfügen, um eingereichte Publikationen und die darin beschriebenen Prozesse im Rahmen ihrer jeweiligen Rolle angemessen beurteilen und so die methodische Qualität von Studien sicherstellen zu können (vgl. Leitlinie 3: Kompetenzaufbau und Spezialisierungen, S. 40).

Besonders problematisch waren und sind Qualitätsprobleme, weil Indikatoren und Simulationen, die auf Forschungsdaten beruhen, während der COVID-19-Pandemie als Grundlage politischer Entscheidungen von enormer gesellschaftlicher Tragweite dienen und Zweifel an der Qualität der Daten leicht zu instrumentalisieren wären (vgl. Leitlinie 8: Gesellschaftlicher Austausch, S. 46). In dieser Situation haben transparente und vertrauensfördernde Mechanismen der Wissenschaftskommunikation und Politikberatung große Bedeutung. Verstärkte öffentliche Aufmerksamkeit wurde beispielsweise der Tatsache beigemessen, dass sich die Bundesregierung unter anderem auf mehrere Ad-hoc-Stellungnahmen der Nationalen Akademie Leopoldina stützte |[91]. Auch die vier großen außeruniversitären Forschungsorganisationen publizierten eine Stellungnahme, in der auf Basis von Simulationen gewonnene Erkenntnisse über die Ausbreitungsdynamik der COVID-19-Pandemie genutzt wurden, um die Wirksamkeit verschiedener Maßnahmen zu beurteilen und adaptive Strategien zur Eindämmung der Pandemie zu empfehlen. |[92] Viele Wissenschaftlerinnen und Wissenschaftler beteiligten und beteiligen sich mit großem persönlichem Einsatz daran, die Genese aktueller Forschungsdaten, ihre Aussagekraft und die Unsicherheiten, mit denen sie behaftet sind, in traditionellen Medien, Podcasts, Blogs, Videos und anderen Formaten allgemeinverständlich darzustellen. Sie fanden, häufig in Kooperation mit erfahrenen Wissenschaftsjournalistinnen und -journalisten, mit differenzierten Darstellungen der aktuellen Kenntnislage teilweise ein Millionenpublikum. Für seinen innovativen Podcast im Norddeut-

| 90 Manancourt, V.; Furlong, A.: *Bungled Lancet study casts shadow over health data industry*, in: *Politico* vom 24. Juni 2020, https://www.politico.com/news/2020/06/24/lancet-study-hydroxychloroquine-health-data-industry-337663.

| 91 Leopoldina – Nationale Akademie der Wissenschaften: Ad-hoc-Stellungnahmen zur Coronavirus-Pandemie, 5. August 2020, https://www.leopoldina.org/uploads/tx_leopublication/2020_08_05_Leopoldina_Stellungnahme_Coronavirus_Bildung.pdf.

| 92 Dabei griffen die Organisationen auf Ergebnisse der Max-Planck COVID19-Modellierungs-Gruppe sowie der Helmholtz-Initiative „Systemische Epidemiologische Analyse der COVID-19-Epidemie“ zurück. Meyer-Hermann, M. et al.: Adaptive Strategien zur Eindämmung der COVID-19-Epidemie, 28. April 2020, https://www.mpg.de/14760567/28-04-2020_Stellungnahme_Teil_02.pdf.

schen Rundfunk, in dessen Rahmen er häufig aktuelle Studien, die Bedeutung neuer Daten und deren Analyse erörterte, wurde der Virologe Christian Drosten mit einem Sonderpreis des Communicator-Preises und dem Grimme Online Award sowohl in der Kategorie „Information“ als auch dem Publikumspreis ausgezeichnet.

Die ersten Monate der Pandemie waren generell von einem großen öffentlichen Interesse an wissenschaftlichen Erkenntnissen zu ihren Ursachen, ihrer Entwicklung und möglichen Maßnahmen geprägt. Teilweise beteiligten sich Bürgerinnen und Bürger an der pandemierelevanten Forschung. Eine Plattform, die im Sinne der Citizen-Science-Bewegung ermöglicht, Rechenzeit auf privaten Rechnern für verteilte Analysen von Strukturproteinen zur Verfügung zu stellen, berichtete im Zuge der Pandemie von einer so hohen Beteiligung, dass es erstmals nicht gelinge, die gespendete Rechenzeit voll auszulasten. |[93] Im Zusammenhang mit der Effektivität der Schließung von Kindergärten und Schulen als nicht-pharmazeutische Interventionen zur Begrenzung der Pandemie wurde wiederholt selbst in Tageszeitungen und Fernsehnachrichten diskutiert, welche Tragweite die Ergebnisse einzelner Studien zum Infektionsgeschehen bei Kindern und in Familien haben und zu welchem Zeitpunkt im Publikationsprozess diese Studien als hinreichend validiert gelten können, um politische Entscheidungen auf ihnen zu basieren. Im Zuge dessen wurde der interessierten Öffentlichkeit bewusst gemacht, wie komplex die Planung und Auswertung einer großen Studie dieser Art ist. Auch wenn es in Einzelfällen Kritik an voreiligen Presseaktivitäten und überzogenen Versprechungen gab, herrschte in Deutschland zunächst der Eindruck vor, dass es während der COVID-19-Pandemie relativ gut gelang, ein differenziertes Bild von der Arbeit der Wissenschaft zu vermitteln, das zumindest teilweise auch einen guten Eindruck von der Bedeutung großer und komplexer Datenmengen und den zu ihrer Verarbeitung notwendigen Prozessen beinhaltete. Befragungen während der ersten Phase der Pandemie ergaben ein hohes Vertrauen breiter Teile der Bevölkerung in die Wissenschaft. |[94] Dieses Vertrauen ist nicht nur für die Akzeptanz entscheidend, die politische Entscheidungen in einer Krisensituation finden, sondern auch eine Voraussetzung dafür, dass Bürgerinnen und Bürger künftig bereit sind, der Verwendung ihrer Daten für bestimmte Forschungszwecke zuzustimmen.

Insgesamt ist durch die COVID-19-Pandemie die Bedeutung von Forschungsdaten und somit auch die von datenbezogenen Tätigkeiten im Forschungsprozess außerordentlich deutlich sichtbar geworden. Damit hoch qualifizierte Wissenschaftlerinnen und Wissenschaftler – gerade auch solche, die noch am Anfang

| 93 „*Usually, your computer will never be idle, but we've had such an enthusiastic response to our COVID-19 work that you will see some intermittent downtime as we sprint to setup more simulations*.“, https://foldingathome.org/covid-19/.

| 94 Wissenschaftsbarometer Corona Spezial – Wissenschaft im Dialog/Kantar, April 2020, https://www.wissenschaft-im-dialog.de/fileadmin/user_upload/Projekte/Wissenschaftsbarometer/Dokumente_20/2020_WiD-Wissenschaftsbarometer_Corona_Spezial_Ergebnispraesentation.pdf.

ihrer Karriere stehen – , derartige Aufgaben zu übernehmen bereit sind, ist es wichtig, dass diese Anerkennung finden und zu wissenschaftlicher Reputation beitragen (vgl. Leitlinie 4: Anerkennung von Daten- und Softwarearbeit, S. 41). Nicht nur traditionelle Autorschaft, sondern auch andere wissenschaftliche Beiträge im Zusammenhang mit der Erzeugung, Auswertung und Pflege von Forschungsdaten müssen künftig als signifikante wissenschaftliche Leistungen gewürdigt werden. Inwieweit ein solcher Wandel in der Wissenschaftskultur gelingt, wird sich zeigen, wenn die akute Phase der aktuellen Krise abgeklungen sein wird. Denn damit Wissenschaft und Gesellschaft Lehren aus ihr ziehen können, müssen virologische, klinische und epidemiologische Datensätze, die unter Hochdruck gewonnen wurden, qualitätsgesichert aufbereitet und archiviert werden, so dass Metastudien durchgeführt werden können (vgl. Leitlinie 5: Nachnutzen und Reproduzieren, S. 42). In Zukunft können Daten aus der COVID-19-Pandemie auch eine Rolle spielen als Vergleichswerte für die Beurteilung des epidemischen Potenzials neuer Erreger. Wann dieser Fall eintreten wird, lässt sich naturgemäß nicht vorhersehen. Umso wichtiger ist, dass Wissenschaft und Gesellschaft anerkennen, wie bedeutsam nachhaltiges Forschungsdatenmanagement, das Teilen von Daten und der Ausbau von Fähigkeiten zu ihrer Analyse als Teile einer Vorsorgestrategie sind, und dafür ausreichende Ressourcen bereitstellen.

Anhang

Abkürzungsverzeichnis

DFG	Deutsche Forschungsgemeinschaft
DOI	*Digital Object Identifier*
DSGVO	Datenschutzgrundverordnung
EOSC	*European Open Science Cloud*
ESSD	*Earth Systems Science Data*
FAIR	*Findability, Accessibility, Interoperability, Reusability*
GCP	*Global Carbon Project*
GTAP	*Global Trade Analysis Project*
GWK	Gemeinsame Wissenschaftskonferenz von Bund und Ländern
IPCC	*Intergovernmental Panel on Climate Change*
KI	Künstliche Intelligenz
ML	Maschinelles Lernen
NFDI	Nationale Forschungsdateninfrastruktur
NHR	Nationales Hochleistungsrechnen
OSPP	*Open Science Policy Plattform*
RatSWD	Rat für Sozial- und Wirtschaftsdaten
RDA	*Research Data Alliance*
RfII	Rat für Informationsinfrastrukturen
RKI	Robert-Koch-Institut
WHO	Weltgesundheitsorganisation
WR	Wissenschaftsrat

Das Publikationsformat „Positionspapier“ wurde 2010 eingeführt, um mit kurzen, zugespitzt formulierten Papieren in absehbarer Zeit auf aktuelle Themen und Entwicklungen reagieren zu können. Im Positionspapier wird deshalb auch - anders als in den übrigen Publikationsformaten des Wissenschaftsrats - darauf verzichtet, umfangreiche empirische Informationen zeitaufwändig aufzuarbeiten und in den Text zu integrieren. Generell zeichnet sich das Format durch eine große prozedurale, thematische und formale Flexibilität aus.

Drs. 8667-20
Köln im Oktober 2020