Peter Wittenburg und Kathrin Beck

1.1 Gesellschaftliche, technologische und internationale, nationalstaatliche bzw. bundeslandspezifische Treiber

Abstract: Daten werden von Wissenschaft, Industrie, Verwaltungen sowie zunehmend auch von Privatpersonen als Citizen Scientists und Anwender moderner Technologien wie z. B. Smart Watches erzeugt. Um diesen verschiedenen Interessen gerecht zu werden sowie um die Daten besser nutzbar zu machen, fördern die Öffentliche Hand, Forschungsförderorganisationen sowie Industrieverbände Maßnahmen zur Wiederverwendung von Daten und zur Entwicklung von Software und Daten-Infrastrukturen. Dieses Kapitel gibt einen Überblick über die Entwicklung der datenbasierten Forschung von ihren Ursprüngen bis in die heutige Zeit.

Einleitung

Digitale Daten werden in vielen Wissenschaften seit etlichen Jahren erzeugt. Beispielsweise konnte man diverse physikalische Prozesse nur durch den Einsatz von Sensoren untersuchen, die Daten erzeugten, die dann von den seit 1964 verfügbaren und immer leistungsfähiger werdenden Rechnern[1] verarbeitet wurden.

Als eines der frühen Beispiele kann auf Friedrich Hertweck verwiesen werden, der am Max-Planck-Institut für Plasmaphysik arbeitete. Er war einer der Wegbereiter für neue Verfahren im Umgang mit digitalen Daten, als er 1970 mit AMOS (Advanced multi user operating system) ein Software-System vorstellte, das darauf abzielte, die an Plasmareaktoren anfallenden Datenmengen sinnvoll analysieren zu können.[2]

Es ist einerseits die Menge der durch Sensoren, Simulationen und Crowdsourcing erzeugten digitalen Primärdaten in nahezu allen wissenschaftlichen Disziplinen, die eine neue Qualität formen, und es ist andererseits das große Maß an Verwobenheit zwischen diesen Rohdaten und vor allem auch den abgeleiteten Daten und Annotationen, die wir mit dem Begriff der Komplexität umschreiben, der wir uns mit neuen Methoden stellen müssen. Es gibt zudem keinen Grund anzunehmen, dass sich diese Entwicklung verlangsamen würde. Mit dem Begriff „Internet of

1 S. https://de.wikipedia.org/wiki/Computer. Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

2 Vgl. Max-Planck-Institut für Plasmaphysik 1998.

https://doi.org/10.1515/9783110657807-002

Things“ wird ein Trend hin zu einer Welt voller kleiner Sensoren in all unseren Lebensbereichen umschrieben. Die Firma Intel prognostiziert, dass wir in 2020 mehr als 200 Mrd. dieser kleinen Erzeuger von kontinuierlichen Strömen hochauflösender Daten haben werden.[3]

Große Fragen drängen sich geradezu auf und sind bisher keineswegs beantwortet. Hier seien nur einige wenige genannt:

- Wem gehören all diese Daten, wer hat Zugriff auf sie und wer darf sie kommerziell nutzen?
- Wie verhindern wir einen Missbrauch, insbesondere von sensiblen und/oder personenbezogenen Daten? Wie können wir ihre Löschung sicherstellen?
- Wie sollen wir mit diesen Daten umgehen, d. h. wie sollen wir sie verwalten? Wie können wir sie für den Menschen sinnvoll zusammenführen und analysieren?
- Werden wir alle zu leichtgläubigen „Dataisten“ wie Yuval N. Harari eine Art neue Religion umschreibt?[4]

Es ist keine Frage, dass mit der Verfügbarkeit von immer mehr und detaillierteren Daten große Umbrüche in der Wissenschaft, Gesellschaft und Industrie einhergehen werden. George Strawn vergleicht die Veränderungen, vor denen wir jetzt in der Wissenschaft stehen, mit denen im 17. Jahrhundert, als die britische Royal Society in einem geradezu revolutionären Akt von allen Wissenschaftlerinnen und Wissenschaftlern forderte, dass sie die damals neuen Druckmöglichkeiten nutzen und ihre Erkenntnisse in Form von öffentlichen Publikationen der interessierten Gemeinschaft verfügbar machen sollten.[5] Wir kennen das Ergebnis dieses wegweisenden Beschlusses für die Wissenschaft – es hat uns ein immer noch weitgehend erhaltenes Gebilde von Theorien, experimentellen Nachweisen und Falsifizierungen gebracht. Dieses Gebilde mit all seinen Texten und Verweisen kann auch als wesentlicher Teil unseres wissenschaftlichen „Gedächtnisses“ bezeichnet werden, das den Stand des menschlichen Wissens zu einem großen Teil umfasst und ohne das wir heute nicht auskommen könnten. Heute wollen Wissenschaftlerinnen und Wissenschaftler jedoch nicht nur auf die Publikationen der Forschenden zugreifen, sondern wollen angesichts der großen Datenmengen und der computationellen Kapazitäten auf den Daten selbst operieren und dabei eigene Analyseverfahren einsetzen.

Es wäre ein Irrglaube, dass Daten an sich bereits Wahrheiten enthalten und moderne KI-Techniken wie z. B. Maschinelles Lernen automatisch die wahren Interpretationen liefern würden. In diesem Sinne machen Begriffe wie z. B. „Open Science“ und „Open Data“ die Runde und werden bereits weitgehend akzeptiert. Dabei ist

3 Vgl. Intel n.d.
4 Vgl. Harari 2015, 497.
5 Vgl. Strawn 2019.

mit open keineswegs gemeint, dass z. B. auch personenbezogene oder Copyright-geschützte Daten prinzipiell offen und für alle einsehbar zur Verfügung stehen sollen. Wie vor mehreren Jahrhunderten nicht gefordert wurde, dass nun auch der gesamte Austausch zwischen den Wissenschaftlern und Wissenschaftlerinnen publik gemacht werden muss, geht es auch heute nicht um die Offenlegung aller durch Wissenschaftlerinnen und Wissenschaftler erzeugten Daten, sondern um eine dahingehende Änderung der Kultur, dass es eine prinzipielle Bereitschaft gibt, die für die Erkenntnisse relevanten Daten und Methoden, eventuell nach einer Karenzzeit, verfügbar zu machen.

Die Frage, die sich uns stellt, ist dann, ob wir auf diese Herausforderungen vorbereitet sind. Vinton G. Cerf, der gemeinsam mit Bob Kahn das Internet begründet hat, spricht davon, dass wir angesichts unserer Verfahren vor einem „Digital Dark Age" („dunklen Zeitalter der Digitalisierung"[6]) stehen, d. h. er befürchtet, dass wir nicht in der Lage sind, ein „wissenschaftliches Gedächtnis" für das digitale Zeitalter aufzubauen und zu verwalten.[7] Dabei spielt sicherlich eine große Rolle, dass wir noch nicht definiert haben, wer im digitalen Zeitalter die Nachfolger der Bibliotheken und der Verlage sein sollten und dass wir uns an das Internet als Basis des Informationsaustausches gewöhnt haben, dabei aber übersehen, dass es zum Aufbau eines Digitalen Gedächtnisses nicht konzipiert war und aufgrund seiner flüchtigen Natur auch vollkommen ungeeignet ist, um Datenmanagement erfolgreich über Zeitspannen von z. B. 100 Jahren zu betreiben.

Hinzu kommt, dass wir im Bereich des digitalen Datenmanagements seit Jahren eine Welle der „Kreolisierung"[8] in allen Aspekten (Datenformate, Organisationen, Werkzeuge, Dienste) erleben, in der sich viele intelligente Köpfe weltweit immer neue Lösungen für naheliegende Probleme ausdenken und diese auch implementieren, was letztlich zu einer enormen Fragmentierung des Datenraumes führt. Diese Fragmentierung sehen viele immer mehr als ein Hindernis, um Daten und Werkzeuge effizient und effektiv zusammenführen und analysieren zu können.[9] Verschiedene Untersuchungen haben gezeigt, dass etwa 80 Prozent der Zeit von Datenexpertinnen und -experten in Projekten mit Datenaufbereitung („Data Wrangling")[10] verlorengeht,[11] d. h. bevor die eigentliche Analysearbeit beginnen kann, sind bereits etwa 80 Prozent der Projekt-Personalkosten verbraucht. Somit werden viele For-

6 S. https://eandt.theiet.org/content/articles/2015/02/digital-data-storage-may-leave-future-in-dark-about-us-warns-cerf/ und https://cltc.berkeley.edu/2016/02/18/video-dr-vinton-cerf-safety-security-and-privacy-in-the-internet/.

7 Vgl. Ghosh 2015.

8 Dieser Begriff wird z. B. in der Linguistik verwendet, um den natürlichen Prozess des Auseinanderdriftens von Sprachen zu beschreiben.

9 Vgl. Wittenburg 2018.

10 S. https://en.wikipedia.org/wiki/Data_wrangling.

11 Vgl. CrowdFlower 2017.

schende von datenintensiver Wissenschaft ausgeschlossen, viele Projekte werden gar nicht erst begonnen und kostbare Ressourcen werden für an sich unnötige Arbeiten verbraucht. So fallen z. B. im US-Gesundheitssystem jährlich 400 Mrd. US-Dollar an unnötigen Kosten an, wobei einer der Hauptfaktoren die Nicht-FAIRness der Daten ist[12] (FAIR: Findable, Accessible, Interoperable, Reusable[13]). Ähnlich dramatisch ist die Situation bezüglich der Reproduzierbarkeit des wissenschaftlichen Erkenntnisgewinns im digitalen Zeitalter. Berichte z. B. aus dem Bereich der biomedizinischen Wissenschaften zeigen, dass sich nur ein kleiner Prozentsatz von Arbeiten reproduzieren lässt,[14] was Tür und Tor für alle möglichen Behauptungen öffnet, deren Substanz nicht überprüft werden kann.

Natürlich dürfen wir die Augen nicht vor grundlegenden Problemen verschließen, die am besten als Daten-Paradoxa umschrieben werden können.

> Data, Data Everywhere, Nor Any Drop to Drink.[15]

Das erste Paradoxon geht auf einen Beitrag von Christine Borgman zurück, in dem sie mittels einer Analogie zu einem Ausspruch von Samuel Taylor Coleridge („Water, water everywhere, nor any drop to drink") verdeutlicht, dass wir bereits über viele Daten verfügen, aber offensichtlich nicht in der Lage sind, diesen Reichtum sinnvoll zu verwenden. Wir sehen vor allem zwei Gründe für diese scheinbar größer werdende Lücke:

- Zum einen müssen wir das Argument von Dimitris Koureas berücksichtigen, dass Daten in einem lokalen Kontext von Experten und Expertinnen erzeugt werden, aber global von anderen verwendet werden sollen, die den Detail-Kontext der Datenerzeugung nicht kennen.[16] Diese Lücke lässt sich mit reichhaltigen Metadaten nur näherungsweise schließen. In den weitaus meisten Fällen werden gegenwärtig nicht einmal minimale Metadaten zur Verfügung gestellt.
- Der zweite Aspekt hat damit zu tun, dass man, um mit den Daten von anderen sinnvoll umgehen zu können, entweder selbst ein Mindestmaß an erforderlichen technischen Kenntnissen mitbringen muss, über das viele Forschende nicht verfügen, oder aber Zugriff auf einen teuren Personalstab haben muss, was ebenfalls in vielen Forschungsinstitutionen weit ab von der Realität ist. Mithin verlassen sich viele Forschende eben doch auf die in Publikationen beschriebenen Analyseresultate, für deren Verständnis man zunächst lediglich die Sprache als Basiswerkzeug beherrschen muss.

12 Vgl. World Economic Forum n.d.
13 Vgl. FORCE11 2016; Kraft 2017; Beitrag von Linne et al., Kap. 3.2 in diesem Praxishandbuch.
14 S. https://en.wikipedia.org/wiki/Replication_crisis.
15 Borgman 2014, 1–2.
16 Vgl. Koureas 2018.

Ein zweites, verwandtes Paradoxon hat mit der Realität von „Dark Data“ zu tun, wie Ryan Heidorn es beschrieben hat.[17] Etwa 80 Prozent der wissenschaftlichen Daten sind demzufolge „Dark Data“,[18] also nicht sichtbar und verfügbar, von denen die weitaus meisten in etwa 20 Jahren verloren gegangen sein dürften. Es werden sehr viele Mittel in Dienste investiert, die auf Daten aufbauen sollen, wie z. B. die Verlinkung von Daten mit Publikationen, Erzeugern, Institutionen, Projektförderungen und anderem, aber eigentlich fehlen in großem Maßstab die Daten selbst, die verlinkt werden können. Momentan ist noch nicht geklärt, wie und von wem die Mittel dafür aufgebracht werden können, solche Daten sichtbar und verfügbar zu machen, und wer letztlich die Rolle für die gewaltige Kurationsaufgabe übernehmen soll. Vielen Forschenden fehlt daher auch das Vertrauen, sich auf die Verfügbarkeit der Daten anderer zu verlassen und Zeit in das Erlernen neuer Methoden zu investieren.

Peter Wittenburg und George Strawn sprechen von einer Notwendigkeit der Konvergenz im Bereich der Daten, um die Phase der Kreolisierung zumindest auf einem bestimmten Niveau zu beenden und dadurch Energie zur Lösung der primären wissenschaftlichen Probleme freizusetzen.[19] Sie sehen momentan zwei wesentliche Ansätze:

- Die nach langen Diskussionen erfolgreiche Etablierung der nunmehr weltweit anerkannten FAIR-Prinzipien[20] kann als Maßstab für alle gesehen werden, ihre Daten so zu gestalten, dass das Umgehen mit diesen effizienter wird.
- Die Definition des Konzeptes der FAIR Digital Objects, die auf langjährige Diskussionen in der Research Data Alliance (RDA)[21] über Disziplingrenzen hinweg basieren und auf frühe Publikationen von Robert Kahn zurückgehen,[22] stellt einen Weg dar, um die FAIR-Prinzipien praktisch umzusetzen.

In dieser Verbindung sehen Wittenburg und Strawn die Chance, eine neue Ebene zu definieren, auf die sich alle einigen können und die – ähnlich wie bei der weltweiten Einigung auf TCP/IP als Internet Protokoll – ungeahnte Kräfte freisetzen könnte, um die oben genannten Probleme anzupacken.

17 Vgl. Heidorn 2008.

18 S. https://de.wikipedia.org/wiki/Dark_Data. Viele Experten halten die Schätzung von Heidorn noch für weit untertrieben.

19 Vgl. Wittenburg und Strawn 2018.

20 Vgl. Wilkinson 2016.

21 S. https://www.rd-alliance.org.

22 Vgl. Kahn und Wilensky 1995; Kahn und Wilensky 2006.

1 Gesellschaftliche Treiber

Nicht zuletzt die Diskussionen um die „Grand Challenges“[23] und die 17 Ziele für eine nachhaltige Entwicklung der UNO[24] haben uns vor Augen geführt, wie stark die Einflüsse unserer Entscheidungen auf die Gestaltung von Natur und Gesellschaft und wie komplex zugleich die Zusammenhänge sind. Im Allgemeinen haben wir auch verstanden, dass angesichts der Komplexität der Herausforderungen nur multikausale, nationale Grenzen übergreifende Betrachtungen zu Lösungen führen werden. Vor allem der zusätzliche Einsatz von datenbasierten Methoden gepaart mit neuartigen Analysemethoden und Simulationen von Modellen, wie sie von Jim Gray beschrieben wurden,[25] werden neue Einsichten vermitteln. Für den Erfolg dieses Weges lassen sich bereits sehr gute Beispiele auch aus dem deutschen Raum nennen. So ist im Bereich der Umweltwissenschaften das Deutsche Klimarechenzentrum (DKRZ) führend in der Erzeugung der Berichte zur Entwicklung des Klimas an die UNO beteiligt, wobei immer umfangreichere, auf Standards basierende Basisdaten[26] und iterativ ergänzte Modelle die Präzision der Vorhersagen kontinuierlich verbessern. Im Bereich der Materialwissenschaften ist es dem EU-Projekt NOMAD[27] gelungen, Millionen von Simulationsergebnissen von Laboren aus vielen Ländern zusammenzubringen und zu normalisieren, sodass die Wissenschaft nunmehr über einen kohärenten Datenraum verfügt, der geeignet ist, Deskriptoren zu berechnen, mit denen sich verschiedene Kategorien von Verbundmaterialien mit spezifischen Eigenschaften klassifizieren lassen.

Im Bereich der Geisteswissenschaften hat z. B. das von der Volkswagen-Stiftung finanzierte DOBES-Projekt[28] Sprachdaten von bedrohten Sprachen aus aller Welt zusammengetragen, an dem Teams von Forschenden aus vielen Ländern mitgewirkt haben. Diese Daten und diejenigen vergleichbarer Projekte ermöglichen es, z. B. Theorien über die Evolution von Sprachen und Kulturen zu präzisieren oder auch vergleichende Untersuchungen z. B. über die Funktionen der Intonation in verschiedenen Sprachen vorzunehmen.

Wie bereits erwähnt, betreffen die Fragmentierung der Daten und daraus folgend deren ineffiziente Weiterverarbeitung auch andere Gesellschaftsbereiche. Bezüglich der Durchdringung durch die Digitalisierung im öffentlichen Dienst hat Deutschland offensichtlich einen Nachholbedarf, dessen sich die Politik zunehmend bewusst wird.[29] Im kommerziellen Sektor werden große Anstrengungen unter-

23 S. https://en.wikipedia.org/wiki/Grand_Challenges.
24 Vgl. United Nations n.d.
25 Vgl. Hey, Tansley und Tolle 2009, xvii-xxxi.
26 Vgl. World Climate Research Programme 2017.
27 S. https://www.nomad-coe.eu.
28 S. https://tla.mpi.nl/project/dobes und http://dobes.mpi.nl.
29 Vgl. Skala 2018.

nommen, dass sich die Vormachtstellung der technologischen Großkonzerne im Bereich der Informationsverwertung nicht auch noch auf den Bereich der Daten ausdehnt. Konsortien wie die von der Fraunhofer Gesellschaft angestoßene International Data Space[30] oder die von der EU finanzierte Big Data Value Association[31] machen deutlich, dass sich die europäische und auch die deutsche Industrie der Herausforderungen annehmen und nach gemeinsamen Lösungen suchen.

Die gesellschaftlichen Treiber für ein verbessertes Datenmanagement lassen sich wie folgt zusammenfassen:

- Die Erkenntnis, dass datenintensive Forschung eine Notwendigkeit ist, um verborgene Muster in komplexen Zusammenhängen zu identifizieren und somit zu neuen Einsichten zu kommen, die uns bei der Bewältigung der „Großen Herausforderungen" helfen können, und um international konkurrenzfähige Forschungsresultate zu liefern.
- Die Erkenntnis, dass Daten ein kostbares Gut sind, um deren Auswertung ein internationaler Wettbewerb entbrannt ist, in dem es letztlich in allen Bereichen darum geht, Zugang zu bekommen bzw. die Hoheit über die Daten nicht zu verlieren.
- Die Erkenntnis, dass drei große Problemstellungen zu bewältigen sind: 1. Wie kann aus Daten Wissen extrahiert werden? 2. Wie kann das Wissen, das in immer mehr Studien gewonnen wird, sinnvoll repräsentiert und auch kombiniert werden, um daraus verwertbare Erkenntnisse abzuleiten? 3. Welche Art von Dateninfrastruktur muss zur Verfügung gestellt werden, um die ersten beiden Problemstellungen nachhaltig und im Sinne hoher Effizienz und Effektivität zu unterstützen?
- Die Erkenntnis, dass Regierungen bezüglich der ersten zwei Punkte dieser Liste nur stimulierend einwirken können, aber bezüglich des dritten Punktes, wie auch bei früheren Infrastrukturen, die Verantwortung übernehmen und entsprechende Mittel bereitstellen müssen, wenn zumindest der Wille vorhanden ist, an dem Reichtum, der den Daten innewohnt, teilhaben zu wollen.

Der Bereich der Wissensextraktion ist gekennzeichnet durch statistische Methoden, die immer weniger Vorannahmen benötigen und auf der Basis von Beispielen lernen, wie z. B. Machine Learning. Die Frage, wie man das aus den Unmengen von Experimenten und Simulationen extrahierte Einzel-Wissen repräsentieren kann, um es in kombinierter Form auswerten zu können, wird weiterhin heftig diskutiert. Nano-Publikationen, die Wissen in Form von erweiterten Resource-Description-Framework-Aussagen[32] (RDF-Aussagen) darstellen, scheinen an Popularität zu ge-

30 S. https://www.fraunhofer.de/de/forschung/fraunhofer-initiativen/international-data-spaces.html und https://www.internationaldataspaces.org.

31 S. http://www.bdva.eu.

32 S. https://de.wikipedia.org/wiki/Resource_Description_Framework und https://www.w3.org/RDF/.

winnen, stellen sie doch eine Form dar, Wissen hochkonzentriert und formal derart zu repräsentieren, dass weitergehende Operationen ermöglicht werden.[33]

Bereits im Jahre 2002 wurde der ESFRI-Prozess (European Strategy Forum on Research Infrastructures)[34] gestartet, um die Gestaltung von Forschungsinfrastrukturen in Europa systematischer anzugehen und Absprachen über Standards zu erzielen. Seit 2006 wurden in mehreren Runden ESFRI-Roadmaps für den Aufbau derartiger Forschungsinfrastrukturen in verschiedenen Disziplinen aufgestellt mit der Konsequenz, dass

- in mehr als 50 Bereichen derartige Infrastrukturen durch europäische und nationale Mittel gefördert wurden, die auf breiter Basis ein höheres Bewusstsein für Daten und neue Technologien erzeugten und auch zu einem großen Teil für verbesserte Methoden sorgten;
- einige dieser Infrastrukturen in ERICs (European Research Infrastructure Consortium)[35] mit der Zielsetzung einer verstetigten Förderung umgewandelt wurden;
- mittels des ESFRI-Prozesses traditionelle Vorstellungen von Wissenschaftsinfrastrukturen überwunden und nunmehr auch die virtuelle Zusammenführung verteilter Datenbanken als essentielle Forschungsinfrastrukturen angesehen werden.

Diese Konzepte wurden in vielen Staaten Europas aufgegriffen und parallele Programme gestartet. Hunderte derartiger virtueller Infrastrukturprojekte wurden in Europa finanziert, was bereits in vielen Bereichen zu einem Aufbruch führte und die Kultur des Datenaustausches in den Disziplinen beeinflusste. Diese Förderungen führten einerseits innerhalb enger Disziplingrenzen zu einer Reduzierung der Fragmentierung, aber andererseits auch zu einer Verfestigung von Silo-Lösungen.

Somit können wir die wesentlichen Treiber hin zu besseren FDM-Lösungen benennen:

- Wissenschaftlerinnen und Wissenschaftler sind daran interessiert, an den bestmöglichen Forschungseinrichtungen, die nunmehr auch die datenintensive Forschung unterstützen müssen, zu arbeiten, um sowohl zum Erkenntnisgewinn beizutragen als auch um ihre Karriere im Rahmen des globalen Wettbewerbs absichern zu können.
- Forschungsorganisationen benötigen eine Basis, die es ihnen erlaubt, einerseits relevante Daten unter Wahrung der zugrundeliegenden Rechte sicher und dauerhaft zu speichern und es andererseits ihren Forschenden zu ermöglichen, relevante datenintensive Forschung (DIF) zu betreiben. Dabei müssen in Zukunft die Effizienz und die Effektivität der DIF gesteigert werden, um die momentan

33 Vgl. Mons und Velterop 2009.

34 S. https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/esfri_en und https://www.esfri.eu.

35 S. https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eric_en.

zu hohen Kosten merkbar zu senken und um die flexible Kombination von Daten verschiedener Herkunft zu vereinfachen.
- Besonders im medizinischen Bereich müssen Wege gefunden werden, um die Verwendung von Daten über das ursprüngliche Forschungsprojekt hinaus zur Erzielung neuer Einsichten über häufigere Krankheiten[36] verwenden zu können, ohne die Persönlichkeitsrechte der Patientinnen und Patienten zu verletzen.
- Die Industrie möchte die Hoheit über die von ihr erzeugten und gesammelten Daten behalten und die möglichen Wertschöpfungsketten in vertragsbasierter Kollaboration mit anderen selbst mitgestalten. Hierbei spielt in Deutschland vor allem die Produktionsindustrie und der Mittelstand eine große Rolle. Sie sind in großem Maße die Datenerzeuger und befürchten, dass andere das Wissen aus diesen Daten extrahieren könnten, ohne dass sie als Erzeuger davon profitieren.
- Die Bürgerinnen und Bürger wollen ebenfalls teilnehmen, wobei sie als Akteurinnen und Akteure mit verschiedenen Rollen auftreten. Sie erzeugen vielfältige Daten und haben ein genuines Interesse, diese auch in Kombination mit anderen Daten zu analysieren, z. B. über Smart Watches oder als Citizen Scientists. Ein demokratisches Verständnis der Gesellschaft legt nahe, dass auch der Bürgerin bzw. dem Bürger, wann immer möglich, Zugang zu Daten und Analysetools gegeben wird, insbesondere zu ihren oder seinen eigenen personenbezogenen Daten. Sie sind jedoch auch daran interessiert, dass ein gewisser Wohlstand und Arbeitsplätze dadurch geschaffen werden, dass die entsprechenden Akteure an den Wertschöpfungsketten bezüglich ihrer Daten teilhaben.
- Die Regierungen müssen sich darum kümmern, dass die gesellschaftlichen Akteure die besten Voraussetzungen haben, um in den beschriebenen Rollen aktiv werden zu können. Dies betrifft dann vor allem den Ausbau einer entsprechenden Dateninfrastruktur, die ein nachhaltiges und effizientes Engagement ermöglicht.

Alle genannten Akteure scheinen sich der enormen Herausforderungen bewusst zu sein und auch in der Bevölkerung ist der Begriff der „Digitalisierung“ jetzt derart mental verankert, dass hohe staatliche Ausgaben breit unterstützt werden. Die Industrie versucht, im Bereich der Infrastruktur mit Initiativen wie dem „International Data Space“[37] Felder zu besetzen und die Daten nicht den technologischen Großkonzernen zu überlassen. Die Regierungen in Europa reagieren mit einer zweiten Welle von Initiativen, wobei die Europäische Kommission (European Commission,

36 Ein Beispiel sind Hirnkrankheiten, deren Verständnis z. B. über Korrelationen zwischen Phänomenen und Mustern in umfangreichen Daten verschiedenster Quellen (Hirnscans, Gensequenzen, psychologische Experimente etc.) vertieft werden sollen. Dies sind Methoden, die die Verfügbarkeit umfangreicher Datenbestände aus verschiedenen Laboren und Kliniken erfordern.

37 S. https://www.internationaldataspaces.org/.

EC) mit der European Open Science Cloud[38] (EOSC) und Deutschland mit der Nationalen Forschungsdateninfrastruktur[39] (NFDI) am weitesten mit ihren Planungen sind. Weitere Staaten und auch Regionen wie z. B. Frankreich, die Niederlande und die nordischen Staaten werden folgen. Dabei sind die Ansätze für die Programme durchaus unterschiedlich.

Laut Thomas P. Hughes besteht die erfolgreiche Umsetzung von großen Infrastrukturprojekten aus einem Zusammenspiel von drei wesentlichen Faktoren:[40]

- technologische Innovation,
- ökonomische/wissenschaftliche Anforderungen und
- geeignete organisatorische und politische Formen.

Das EOSC-Programm der EC hat der Schaffung einer für die Mitgliedstaaten überzeugenden organisatorischen Struktur den Vorrang gegeben und ist nunmehr in der zweiten Projektphase bemüht, den technologischen Kern zu definieren. Der Anspruch ist derartig umfassend, dass viele Expertinnen und Experten ein Scheitern der Pläne befürchten. Die EC setzt jedoch auf verschiedene Arbeitsgruppen, die in Zusammenarbeit mit der neuen Allianz aus RDA, Committee on Data of the International Science Council[41] (CODATA) und GO FAIR[42] konkrete Vorschläge ausarbeiten sollen, die die EOSC als eine FAIR-basierte, distribuierte Infrastruktur-Landschaft entstehen lassen können. Das deutsche NFDI-Programm verfolgt einen anderen Ansatz, indem es zunächst eindeutig den wissenschaftlichen Motivationen und Planungen eine höhere Priorität gibt und die Planung einer konvergenten technologischen Komponente in den Hintergrund stellt. Damit wird im Prinzip der frühere Ansatz des ESFRI-Programms weiterverfolgt, der jedoch in einer Verfestigung der Silo-Mentalität enden könnte. Allein durch EOSC und NFDI werden jährlich ca. 90 Mio. Euro für die Entwicklung eines Daten-Infrastruktur-Ökosystems ausgegeben.

In den USA, wo die Entwicklung bisher durch die großen Informationskonzerne, wie Google, Facebook etc., vorangetrieben wird, verhalten sich die staatlichen Akteure, die ein nationales Programm für eine US-Forschungsdateninfrastruktur einfordern, noch zurückhaltend.[43] Bisherige Programme waren konzipiert, um Pilotprojekte mit dem Ziel zu unterstützen, ein größeres Verständnis darüber zu bekommen, was Infrastrukturen leisten können und wie man sie organisieren kann. Insbesondere kann hier das Programm zu „Research Data Commons“ von den National Institutes of Health[44] genannt werden. Bisher gab es in den USA keine einheitliche

38 S. https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud; https://eosc-portal.eu; Beitrag von Streit und van Wezel, Kap. 1.2 in diesem Praxishandbuch.
39 S. https://www.dfg.de/foerderung/programme/nfdi/index.html.
40 Vgl. Hughes 1983, 461-465.
41 S. http://www.codata.org.
42 S. https://www.go-fair.org; Beitrag von Linne et al., Kap. 3.2 in diesem Praxishandbuch.
43 Vgl. Wittenburg und Strawn 2019.

Überzeugung, wie man eine umfassende Dateninfrastruktur aufbauen soll. Mit der breiten Akzeptanz der FAIR-Prinzipien und dem Ausformulieren der FAIR Digital Objects[45] scheint sich eine Änderung der Einschätzungen abzuzeichnen.

In China verfolgt man die Diskussionen über die FAIR-Prinzipien und den FAIR Digital Objects engagiert und organisiert entsprechende Konferenzen,[46] um sich gegebenenfalls mit großem Mittelaufwand an den Entwicklungen führend beteiligen zu können.

2 Technologische Treiber

Die technologische Innovation wird immer wieder neue Anziehungspunkte für die Wissenschaft und darüber hinaus definieren, wobei die Wissenschaft oftmals eine Vorreiterrolle einnimmt, ist sie doch prinzipiell zu größeren Risiken bereit. Dabei gilt jedoch, dass Standards gut für die Wissenschaft insgesamt sind, zunächst nicht jedoch für die individuellen Wissenschaftlerinnen und Wissenschaftler, die Produktivitätseinbußen befürchten. Dennoch wollen einige die neuesten technologischen Entwicklungen für ihre Zwecke so früh wie möglich einsetzen, da sie neue Möglichkeiten antizipieren und bereit sind, mit Technologen zusammenzuarbeiten, und sich trauen, in neuen Bereichen zu publizieren. Die wesentlichen gegenwärtigen technologischen Trends, die relevant für den Bereich der datenintensiven Wissenschaft[47] sind, lassen sich in einigen Kernaussagen zusammenfassen:

Tab. 1: Die von G. Strawn übernommene Tabelle über die Entwicklung der IT-Kapazitäten schaut vom Stand 2000 aus 30 Jahre zurück und wagt Prognosen für 30 Jahre in die Zukunft. Es gibt keine Gründe anzunehmen, dass die dynamische Entwicklung mit der Einführung von Post-Chip und anderen innovativen Technologien abnehmen wird.

	1970	2000	2030
Technology	pre-chip	Chip	post-chip
US $	1.000.000	1.000	1
CPU	1 mips	1 gips	1 tips
Disk	$ 1/kB	$ 1/gB	$ 1/pB
Net	10 kbps	10gbps	10pbps

44 S. https://commonfund.nih.gov/commons.

45 Vgl. Wittenburg et al 2019; Schultes und Wittenburg 2019 und RDA GEDE group 2019.

46 Vgl. FAIR DO Session 2019.

47 Wie auch bereits Jim Gray bei seiner Einführung des Begriffes der Data-Intensive Science betonte, wird es in der Wissenschaft auch weiterhin traditionelle Methoden geben, deren Bedeutung nicht in Frage gestellt wird.

- Die *Kapazitäten in der IT* (CPU, Speicher, Netzwerk) nehmen weiter zu und neue technologische Ansätze wie z. B. Quantencomputing lassen für die Zukunft enorme Sprünge erwarten wie in Tab. 1 dargestellt wird. Wir kennen die optimalen Einsatzmöglichkeiten dieser neuen Ansätze noch nicht genau, aber es besteht kein Zweifel, dass die Wissenschaft darauf wartet, sie einsetzen zu können.
- Neuartige mathematische Verfahren werden entwickelt, um die riesigen, virtuell integrierten *Datenmengen analysieren* zu können. Hier sei nur auf die Deep-Learning-Ansätze verwiesen, die noch weniger Vorannahmen erfordern als frühere Ansätze und daher noch abhängiger von großen Datenmengen und geeigneten Lernstrategien sind. Wir beschreiben diese Ebene mit dem Begriff der „Extraktion von Wissen aus Daten".
- Angesichts der großen Datenmengen ist der Einsatz *automatischer Verfahren mittels Workflow-Werkzeuge* eine zunehmende Notwendigkeit. Dabei wird zu erwarten sein, dass Wissenschaftlerinnen und Wissenschaftler Daten-Profile definieren und es Crawlern überlassen, geeignete Daten zu finden und das Ausführen der Workflows zu starten.
- Die große Anzahl der aktiven Wissenschaftlerinnen und Wissenschaftler und der Einsatz automatischer Verfahren werden es erforderlich machen, nach *neuen Methoden der Präsentation von Wissen* zusätzlich zu den etablierten wissenschaftlichen Publikationen zu suchen. Vorschläge wie Nano-Publikationen,[48] im Wesentlichen augmentierte RDF-Aussagen über wesentliche Resultate, werden gegenwärtig diskutiert.
- Eine weitere große Herausforderung wird sein, wie wir die zunehmende Zahl wissenschaftlicher Resultate (Detailwissen) zu *Erkenntnissen* zusammenbringen können, die für die Gesellschaft nutzbringend sind. Wenn wir über geeignete formale Verfahren verfügen, um Wissen darzustellen, werden uns intelligente AI-Methoden (Artificial Intelligence) helfen, zu Erkenntnissen zu kommen.
- Die bisher genannten Ebenen werden nur dann erfolgreich und effizient umgesetzt werden können, wenn wir über geeignete *Dateninfrastrukturen* verfügen, die es unter anderem erlauben, inkrementell ein digitales Gedächtnis aufzubauen, sodass eine Abkehr von flüchtigen Methoden z. B. des Internets möglich ist.

Während der in den ersten fünf Punkten beschriebene Fortschritt vom Erkenntnisdrang der Wissenschaft und dem Marktstreben der Industrie vorangetrieben wird und auch bereits ziemlich große Schritte gemacht wurden, folgt der letztgenannte Bereich der Infrastrukturen gänzlich anderen Gesetzen. Das Entwickeln von Infrastrukturen ist wissenschaftlich wenig attraktiv und für die Industrie ambivalent.

48 Vgl. Mons und Velterop 2009.

Proprietäre Infrastrukturen können einen Marktvorteil bieten, sind aber für die Allgemeinheit nicht akzeptabel. Offene Infrastrukturen eröffnen für alle, also auch neuen innovativen Firmen, die gleichen Einstiegschancen. Offene Infrastrukturen können mithin auch nur von der Allgemeinheit finanziert werden.

IT-geprägte Infrastrukturen müssen global geplant und umgesetzt werden, da die intensive internationale Vernetzung einheitliche Strukturen verlangt und auf Dauer keine Infrastruktur-Inseln überlebensfähig sind. Auch proprietäre Strukturen, wie sie von großen IT-Konzernen etabliert werden, werden sich gegen allgemeine Trends nicht durchsetzen können, sowie sich eine weitgehende, globale Übereinkunft auf bestimmte Standards abzeichnet. Diese Übereinstimmung zu erzielen, stellt allerdings eine große technologische und vor allem soziologische Herausforderung dar – sie ist ungleich schwerer zu erreichen, als es im Falle des Internets der Fall war, da der Bereich des FDM sehr viel vielschichtiger ist. Die Durchsetzung großer Infrastrukturen in der Vergangenheit basierte jeweils auf sehr einfachen minimalen Spezifikationen (z.B. 50 Hz/220 V, TCP/IP, HTTP), um ein Momentum hin zur Reduktion der Fragmentierung zu erzeugen, ohne Innovationen auf anderen Ebenen zu blockieren.

Eine ganze Reihe von Initiativen hat sich gebildet, um zu Übereinkünften zu kommen, die die Fragmentierung verringern können. Die RDA, die etwa 9 000 Expertinnen und Experten aus derzeit 137 Ländern umfasst, arbeitet aktuell in 86 Gruppen an Spezifikationen von Komponenten sowie an Prozeduren. Bemängelt wird oftmals, dass es der RDA an einem großen übergeordneten Konzept fehlt und somit keine Richtung erkennbar ist. CODATA ist eine internationale Organisation, die vornehmlich an politischen Richtlinien arbeitet und sich mit verschiedenen Netzwerk-Methoden insbesondere auch an Entwicklungsländer richtet. World Data Systems (WDS) hat sich insbesondere der Qualität und Persistenz von Repositorien gewidmet und dann unter dem Dach der RDA zusammen mit der Data-Seal-of-Approval-Initiative den neuen gemeinsamen Standard, CoreTrustSeal, für die Zertifizierung von Repositorien ausgearbeitet. Etwas neueren Datums ist die GO FAIR Initiative,[49] die Impulse setzen will, indem sie über die Spezifikation hinausgeht und Standards implementieren will. Die Formulierung der FAIR-Prinzipien geht auf die Gründer der GO-FAIR-Initiative zurück, die es verstanden, längere Diskussionen zu prägnanten Aussagen zu bündeln. Unter dem Mantel einer RDA-Arbeitsgruppe wird momentan an FAIR-Maturity-Indikatoren[50] gearbeitet, wobei es das vordringliche Ziel ist, auch Software bereitzustellen, die automatische Tests der FAIRness von Datensätzen erlaubt.

Gegenwärtig zeichnet sich eine breite internationale Einigkeit über die FAIR-Prinzipien ab. In der RDA-Maturity-FAIR-Indicator-Gruppe wird intensiv an Regeln

49 S. https://www.coretrustseal.org.

50 S. https://www.rd-alliance.org/groups/fair-data-maturity-model-wg.

gearbeitet, um die FAIRness festzustellen, und es soll auch auf nutzbare Software hingewiesen werden. Ebenfalls arbeitet eine breite Gruppe von Expertinnen und Experten an der Umsetzung der FAIR-Prinzipien mittels des Konzepts der FAIR Digital Objects. Eine Reihe von Kernkomponenten, wie z. B. das Handle PID System, das Digital Object Interface Protokoll, die Data Type Registry und die Kernel-Attribute, wurden größtenteils in RDA-Gruppen spezifiziert und anschließend implementiert. Ebenfalls erfolgen in verschiedenen Projekten Implementierungen und Testbett-Entwicklungen.

Von großer Bedeutung für die Wiederverwendung von Daten ist auch das Vorhandensein von „rich" Metadaten, wobei der in FAIR verwendete Begriff „rich" bewusst vage gehalten ist. Letztlich geht es darum, die Wissensdifferenz zwischen den lokal arbeitenden Erzeugerinnen und Erzeugern von Daten und den global arbeitenden Benutzerinnen und Benutzern zu überbrücken. Die Art der benötigten Metadaten hängt allerdings sehr stark vom Verwendungszweck ab. Für allgemeine Suchen reichen typischerweise Attribute wie „Autor, Titel, grobe Disziplinklassifizierung, Institution" u. ä. aus. Für das gezielte wissenschaftliche Suchen zur Auswahl von Daten für spezifische Operationen reicht das nicht aus und disziplinspezifische Attribute sind erforderlich. Ebenso sind für die Orchestrierung automatischer Workflows sehr spezifische Beschreibungen des Datentyps erforderlich. Der Bereich der Metadaten ist bisher keineswegs vernünftig im Sinne von maschinenlesbaren Verfahren gelöst. So fehlen z. B. klare Kategorisierungen von Metadatentypen, auffindbare und harmonisierte Registraturen bzw. aktuell gehaltene Mappings für Schemas und Vokabulare sowie einfache Ontologie-unterstützte Editoren. Große Einigkeit besteht darin, dass Metadaten unabhängig von der Art der internen Handhabung als RDF-Aussagen exportiert werden sollten, um mittels Linked-Data-Methoden Inferenzen und anderes bilden zu können.

Die dargestellten Verfahren basieren allesamt darauf, dass eine funktionierende und ständig erweiterte Basisinfrastruktur vorhanden ist. Netzwerk-, Speicher- und CPU-Kapazitäten müssen ständig erweitert werden, um die höheren Bedarfe abzusichern. Cloud-Systeme stellen dabei einen neuen Ansatz dar, der es erlaubt, schnell mit großen Mengen an Objekten zu arbeiten und auch effizient mit großen Rechnerkapazitäten (Virtual Machines) umzugehen. Insbesondere die großen IT-Firmen bieten verlockende Dienste an, wobei allerdings große Fragen hinsichtlich der Nutzung und Sicherheit der Daten aufgeworfen werden. Die Regeln der europäischen General Data Protection Regulation (GDPR) stellen dabei einen sehr strikten Rahmen für die Verwendung personenbezogener Daten dar.

3 Nationalstaatliche Treiber

Wie bereits dargestellt, muss die Entwicklung neuartiger Dateninfrastrukturen von den Staaten gefördert werden, um ihren verschiedenen Akteurinnen und Akteuren die Mittel zu geben, unnötige Ausgaben zu vermeiden und neue Erkenntnisse zu ermöglichen bzw. neue Wertschöpfungsketten und Jobs zu realisieren. Dies alles erfolgt unter den Rahmenbedingungen eines harten internationalen Konkurrenzwettbewerbs.

Daher haben sich vor allem die nord- und westeuropäischen Staaten frühzeitig finanziell engagiert. Erste große Programme wurden gemeinsam mit der Grid-Initiative[51] gestartet, die jedoch sehr schnell von IT-Aspekten geleitet wurde und trotz eines hohen Wissenszuwachses bei den direkt Beteiligten zu keinen wesentlichen Impulsen führte, sieht man einmal von den positiven Folgen z.B. für die Hochenergiephysik und der Wegbereiter-Funktion für das Cloud-Computing ab. In einer zweiten Welle beteiligten sich im Wesentlichen die meisten europäischen Staaten am ESFRI-Prozess und finanzierten auch selbst umfangreiche Infrastrukturprojekte und Projekte, in denen die Digitalisierung und Aufbereitung von Datensammlungen im Mittelpunkt standen. Dabei wurden verschiedenste Ansätze gefördert mit dem Ergebnis, dass sich in vielen Sektoren und auch Disziplinen ein klareres Bild davon abzeichnete, was denn nun Dateninfrastrukturen ausmacht, was generisch und was sektor- bzw. disziplinspezifisch angegangen werden muss.

Gleichzeitig wurden Initiativen gebildet, um Diskussionsprozesse zu starten, die Beiträge in Richtung einer höheren Kohärenz der Datenlandschaft liefern und Brücken bilden sollen. Im Bereich der Wissenschaft sind in Deutschland vor allem die Allianz-Initiative[52] „Digitale Information“, die 2008 von der Allianz der deutschen Wissenschaftsorganisationen gegründet wurde, und der Rat für Informationsinfrastrukturen (RfII)[53] zu nennen. Während Erstere für die Datenpraxis wenig sichtbare Resultate brachte, formulierte Letztere die Rahmenbedingungen für die NFDI, die jetzt mit der Bildung von breiten und vernetzten Konsortien eine konkrete Form angenommen haben.

Deutschland hat mit der NFDI einen umfassenden neuen Anstoß gegeben, der parallel zur europäischen EOSC Beiträge liefern soll, und ist gleichzeitig Vorreiter für weitere nationale und regionale Programme in Europa. Wie bereits angedeutet, setzt die NFDI-Initiative auf ein Primat der wissenschaftsgetriebenen Ansätze. Sogenannte Querschnittsthemen sollen in einem zweiten Ansatz behandelt werden, was das Risiko in sich birgt, dass technologisch innovative Konzepte nicht verfolgt werden und somit anderen das Feld für Innovation überlassen wird.

51 S. https://gauss-allianz.de/de/network/NGI-DE.
52 S. https://www.allianzinitiative.de.
53 S. http://www.rfii.de.

Erhebliche nationale (öffentliche und private) Mittel werden in den Ausbau der AI investiert, wobei sich vor allem auch der Bitkom[54] engagiert. Es bedarf der Ergänzung des Methodenkanons, um Wissen zu extrahieren, des Verfügbarmachens dieser Methoden in einfacher Weise und vor allem auch der Ausbildung einer Generation von Expertinnen und Experten, die verstehen, mit diesen Methoden umzugehen. Auch bezüglich der Ausbildung von Datenmanagerinnen und Datenmanagern sowie Data Stewards ist von den Ausbildungseinrichtungen ein dringender Nachholbedarf erkannt worden. An verschiedenen Universitäten und Fachhochschulen werden Curricula entworfen und auch schon angeboten.[55] Dies sind Maßnahmen, die sich in ein paar Jahren auszahlen werden.

4 Bundeslandspezifische Treiber

Viele Bundesländer haben in den letzten Jahren eigene Digitalstrategien entwickelt und in ihren Bildungsministerien oder zentralen Forschungseinrichtungen verankert und kommen somit ihrer Verantwortung für die Weiterentwicklung der Hochschulen nach, die durch die zunehmende Bedeutung der Daten erforderlich ist. Einige Beispiele hierfür sind:

Das Land *Baden-Württemberg* hat ein Fachkonzept von fünf zentralen Handlungsfeldern publiziert: Lizenzierung elektronischer Informationsmedien, Digitalisierung, Open Access, Forschungsdatenmanagement, Virtuelle Forschungsumgebungen[56]. In vier zentralen Forschungsdatenzentren (Science Data Centers, SDC) werden Forschung und Ausbildung bezüglich Datenwissenschaft und -management verschiedener Fachbereiche vorangetrieben.[57] Des Weiteren wurde und wird eine Bandbreite von datenbezogenen Diensten und Projekten zu verschiedensten Bereichen der Lehre und Forschung entwickelt und vom Arbeitskreis der Leiterinnen und Leiter der wissenschaftlichen Rechenzentren in Baden-Württemberg[58] bereitgestellt.

In *Bayern* wird an der Plattform „Forschungsdatenmanagement“[59] gearbeitet, um die bayerischen Akteure und Projekte zu vernetzen.

Im Land *Berlin* wurde 2015 ein Open-Access-Büro[60] gegründet, das die beteiligten Akteure[61] koordiniert. Des Weiteren ist ein regionales Datenzentrum Digital Humanities geplant.

54 S. https://www.bitkom.org/Bitkom/Organisation/Gremien/Big-Data-und-Advanced-Analytics.html.

55 Z. B. https://www.ddm-master.de/.

56 Vgl. Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg n.d.

57 Vgl. Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg 2019.

58 S. https://www.alwr-bw.de/kooperationen.

59 S. https://www.fdm-bayern.org.

In *Brandenburg* wurde im Rahmen eines Forschungsprojekts eine Open-Access-Strategie[62] entwickelt, die 2019 veröffentlicht wurde.

Die *Hamburger* Bildungs- und Wissenschaftseinrichtungen haben sich zu einer hochschulübergreifenden Strategie „Hamburg Open Science“[63] zusammengeschlossen. Darüber hinaus bietet z. B. die Universität Hamburg mit ihrem Zentrum für nachhaltiges Forschungsdatenmanagement[64] Beratungsangebote und ein Repositorium an. Die Stadt Hamburg veröffentlicht im Rahmen eines Open-Data-Aktionsplans ihre Verwaltungsdaten im Transparenzportal Hamburg[65].

In *Hessen* haben sich elf hessische Hochschulen in der Landesinitiative „Hessische Forschungsdateninfrastrukturen“ (HeFDI) zusammengeschlossen, um mittels eines Repositoriums, Beratungs- und Service-Leistungen ihr Forschungsdatenmanagement zu verbessern.

In *Niedersachsen* wurde 2017 eine Digitalisierungsoffensive gestartet[66] und das Zentrum für digitale Innovationen Niedersachsen (ZDIN)[67] gegründet.

In *Nordrhein-Westfalen* wurde die zentrale Koordinierungsstelle „fdm.nrw“[68] aufgebaut, das die Hochschul- und Landesaktivitäten koordiniert, auch im Hinblick auf Verknüpfung mit der NFDI und anderen bundesweiten Aktivitäten.

In *Schleswig-Holstein* wurde der Schwerpunkt der Digitalisierung auf Open Access und Open Data gelegt. So hat die Landesregierung eine „Strategie 2020 für Open Access“[69] initiiert, an der die Hochschulen sowie die Wissenschaftlerinnen und Wissenschaftler die Hauptakteurinnen und -akteure sind. Zusätzlich werden Daten der öffentlichen Einrichtungen über Repositorien[70] zugänglich gemacht.

Die *Thüringer* Strategie für die Digitale Gesellschaft[71] umfasst verschiedene Bereiche der Gesellschaft: Wirtschaft 4.0 wurde 2016 gestartet. Danach folgten „Mittelstand 4.0“, „Digitale Landesentwicklung für den städtischen und ländlichen

60 S. http://www.open-access-berlin.de/strategie.

61 S. http://www.open-access-berlin.de/akteure/index.html.

62 Vgl. Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg n. d. Hier werden die Themenfelder des Ministeriums beschrieben und die Open-Access-Strategie verlinkt.

63 S. https://openscience.hamburg.de/de/ueber-uns/beteiligte-institutionen.

64 S. https://www.fdm.uni-hamburg.de.

65 S. http://transparenz.hamburg.de/open-data.

66 S. https://www.lhk-niedersachsen.de/positionen/digitalisierung und https://www.niedersachsen.de/startseite/themen/digitales_niedersachsen.

67 S. http://www.zdin.de.

68 S. https://www.fdm.nrw.

69 S. https://www.schleswig-holstein.de/DE/Fachinhalte/H/hochschule_allgemein/OpenAccess.html.

70 S. https://www.schleswig-holstein.de/DE/Landesregierung/Themen/Digitalisierung/Transparenzportal/transparenzportal.html.

71 S. https://www.digital-thueringen.de.

Raum“, „Bildung und Forschung digital“ sowie Querschnittsthemen; 2019 wurde die Digitalstrategie aktualisiert.

Der Schwerpunkt der Landesaktivitäten liegt in der Bereitstellung von Repositorien, der Definition von Rahmenrichtlinien sowie insbesondere auch der Vernetzung der Expertinnen und Experten. Außerdem bieten sie Schulungen des Fachpersonals, um möglichst frühzeitig Trends zu identifizieren und darauf reagieren zu können. Hinzu kommt natürlich, dass es einige Hochschulen Bildungsangebote im Bereich des FDM entwickelt haben.

Fazit

Die Vorstellung der FAIR-Prinzipien hat allen Akteurinnen und Akteuren bis hin zu den Entscheidungstragenden verdeutlicht, dass das FDM bereits jetzt nicht optimal erfolgt und dass die Ineffizienzen und Verluste sich angesichts der zunehmenden Datenvolumina und vor allem der Komplexität noch potenzieren würden, wenn die Wissenschaftsgemeinde nicht entschieden gegensteuern würde. Dabei ist seitens der politischen Ebene erkannt worden, dass große Investitionen erforderlich sein werden, um wirklich eine Open-Science-Landschaft aufzubauen und sich somit auch dem kommerziellen Druck entgegenzustellen.

Auf der Ebene der Expertinnen und Experten sind Europa und insbesondere auch Deutschland gut aufgestellt. Es gibt ein breites Wissen durch den ESFRI-Prozess und viele andere Maßnahmen auch auf nationalem Niveau. Es waren europäische und zum großen Teil deutsche Expertinnen und Experten, die die RDA vorangetrieben haben aus dem Wissen heraus, dass nur globale Standards helfen werden. Es waren vor allem europäische Expertinnen und Experten, die die FAIR-Prinzipien formuliert haben.

Das Beispiel der Diskussion um die FAIR-Digitalen-Objekte zeigt aber auch, dass es in Europa wiederum zu wenig Bereitschaft gibt, neue integrative Technologien auszutesten und damit den konzeptionellen Vorsprung auch in einen Implementationsvorsprung umzusetzen.

Literatur

Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

Borgman, Christine. 2014. „Data, data, everywhere, nor any drop to drink.“ Amsterdam: 4th RDA Plenary. https://www.slideshare.net/ResearchDataAlliance/christine-borgman-keynote.

CrowdFlower. 2017. „2017 Data Scientist Report.“ https://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf.

FAIR DO Session at CODATA Conference. 2019. http://codata2019.csp.escience.cn/dct/page/70006.
FORCE11. 2016. „The FAIR Data Principles“ https://www.force11.org/group/fairgroup/fairprinciples.
Ghosh, Pallab. 2015. „Google's Vint Cerf warns of ‚digital Dark Age'.“ *BBC News*. https://www.bbc.com/news/science-environment-31450389.
Harari, Yuval N. 2017. *Homo Deus: Eine Geschichte von Morgen*. München: C. H.Beck.
Heidorn, Bryan. 2008. „Shedding Light on the Dark Data in the Long Tail of Science.“ https://www.academia.edu/23517673/Shedding_Light_on_the_Dark_Data_in_the_Long_Tail_of_Science.
Hey, Tony, Stewart Tansley und Kristin Tolle, Hg. 2009. „The Fourth Paradigm: Data-Intensive Scientific Discovery.“ Microsoft Research. https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/M091000H.pdf.
Hughes, Thomas. 1983. *Networks of Power*. Baltimore: Johns Hopkins University Press.
Intel. n. d. „A Guide to the Internet of Things Infographic.“ https://www.intel.com/content/www/us/en/internet-of-things/infographics/guide-to-iot.html.
Kahn, Robert und Robert Wilensky. 2006. „A Framework for Distributed Digital Object Services.“ https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf.
Kahn, Robert und Robert Wilensky. 1995. „A Framework for Distributed Digital Object Services.“ https://www.cnri.reston.va.us/home/cstr/arch/k-w.html.
Koureas, Dimitris. 2018. „Digital Objects – The Science Case.“ https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/DO-Workshops/workshop-September-18/6-koureas-intro-talk.pdf.
Kraft, Angelina. 2017. „Die FAIR Data Prinzipien für Forschungsdaten.“ TIB Blog. https://blogs.tib.eu/wp/tib/2017/09/12/die-fair-data-prinzipien-fuer-forschungsdaten/.
Max-Planck-Institut für Plasmaphysik. 1998. „Professor Friedrich Hertweck emeritiert. Wegbereiter des Supercomputing in Deutschland.“ https://www.ipp.mpg.de/ippcms/de/presse/archiv/10_98_pi.
Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg. n. d. „E-Science.“ https://mwk.baden-wuerttemberg.de/de/forschung/forschungslandschaft/e-science/.
Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg. 2019. „Vier Science Data Centers in Baden-Württemberg.“ https://mwk.baden-wuerttemberg.de/de/service/presse/pressemitteilung/pid/vier-science-data-centers-in-baden-wuerttemberg/.
Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg. n. d. „Digitalisierung.“ https://mwfk.brandenburg.de/mwfk/de/wissenschaft/digitalisierung/.
Mons, Barend und Jan Velterop. 2009. „Nano-Publication in the e-science era.“ Semantic Web Applications in Scientific Discourse (SWASD). http://ceur-ws.org/Vol-523/Mons.pdf.
RDA GEDE Group members. 2019. „Moving Forward on Data Infrastructure Technology Convergence.“ https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop.
Schultes, Erik und Peter Wittenburg. 2019. „FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure.“ doi:10.23728/b2share.166a074bff614a31-b05e9df5bfd9809d.
Skala, Fridolin. 2018. „Darum liegt Deutschland bei der Digitalisierung hinten.“ faz.net. https://www.faz.net/aktuell/politik/inland/digitalisierung-darum-liegt-deutschland-im-eu-vergleich-hinten-15480625.html.
Strawn, George. 2019. „Open Science, Business Analytics, and FAIR Digital Objects.“ doi:10.23728/b2share.6ceeed13eb6340fcb132bcb5b5e3d69a.
United Nations. n. d. „Sustainable Development Goals.“ https://www.un.org/sustainabledevelopment/sustainable-development-goals/.

Wilkinson, Mark et al. 2016. „The FAIR Guiding Principles for scientific data management and stewardship.“ https://www.nature.com/articles/sdata201618.

Wittenburg, Peter und George Strawn. 2019. „About Building Data Infrastructures.“ https://b2share.eudat.eu/records/6b596f01bc224ff284f80a057212e07f.

Wittenburg, Peter, George Strawn, Barend Mons, Luiz Boninho und Erik Schultes. 2019. „Digital Objects as Drivers towards Convergence in Data Infrastructures.“ doi:10.23728/b2share.b605d85809ca45679b110719b6c6cb11.

Wittenburg, Peter und George Strawn. 2018. „Common Patterns in Revolutionary Infrastructures and Data.“ doi:10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0.

World Climate Research Programme. 2017. „CMIP Phase 6 (CMIP6).“ https://www.wcrp-climate.org/wgcm-cmip/.

World Economic Forum. n. d. „A Global Roadmap for Health Informatics Standardization. Proposal prepared by the World Economic Forum, in collaboration with Boston Consulting Group.“ http://www3.weforum.org/docs/WEF_Global_Roadmap_for_Health_Informatics_Standardization.pdf.