Informatik Spektrum (2020) 43:436–439
https://doi.org/10.1007/s00287-020-01320-0

AKTUELLES SCHLAGWORT

Datenkompetenz – Data Literacy

Thomas Ludwig[1,2] · **Hannes Thiemann**[1]

Online publiziert: 26. Oktober 2020

Zusammenfassung
Unsere Zeit der sprunghaft anwachsenden Datenmengen in allen Bereichen erfordert die Ausbildung von Datenkompetenz als Schlüsselkompetenz für das 21. Jahrhundert. Kenntnisse zur Datensammlung, zum Datenmanagement, zur Datenevaluation und zur Datenanwendung bilden die Grundlage für einen kompetenten Umgang mit Daten in Wissenschaft, Wirtschaft und Gesellschaft. Umfangreiche Datenmengen sind heute in allen Lebensbereichen in Wertschöpfungsketten eingebunden, die es zu gestalten und zu bewerten gilt. Insbesondere im Bereich der Wissenschaftsdaten wird dies auch institutionell unterstützt, um aus Daten neues Wissen und neue Einsichten generieren zu können.

Unsere moderne Welt produziert ein rasant steigendes Datenvolumen. Einem Bericht von IDC („International Data Corporation“) zufolge umfasst die *Global Datasphere* im Jahr 2019 bereits über 40 Zettabyte, das sind 40.000.000 Petabyte [1]. IDC versteht darunter die Daten, die weltweit neu erzeugt, verarbeitet und verteilt werden. Gleichzeitig wurden 2019 weltweit Speichermedien mit einer Gesamtkapazität von ca. 2 Zettabyte verkauft [1]. Das neue Jahrtausend wird zum Zeitalter der Daten. Fortschritte in der Halbleiter- und Vernetzungstechnik lieferten die Grundlage für ein exponentielles Wachstum an Leistung und Geschwindigkeit in allen IT-basierten Systemen. Man sieht das im privaten Bereich: Eine digitale Kamera Baujahr 1999 erstellte Bilder mit einer Auflösung von 640 × 480 Pixel, aktuell erfassen sie mehr als 20 Mio. Punkte pro Bild und im Laufe einer Urlaubsreise entstehen leicht 1000 Bilder und mehr. Bilder aus z. B. medizinischen Geräten oder Satelliten haben über die Jahre vergleichbare Dimensionssteigerung erfahren. Darüber hinaus gibt es zunehmend mehr Datenproduzenten jeglicher Art und die Heterogenität der Daten steigt an. Auch wird erwartet, auf Daten von jedem Ort der Welt über einen garantierten langen Zeitraum zugreifen zu können. Um diese wachsende Datenflut in ihrer Komplexität beherrschen zu können, bedarf es einer umfassenden Datenkompetenz (neudeutsch: *Data Literacy*).

✉ Thomas Ludwig
ludwig@dkrz.de

[1] Deutsches Klimarechenzentrum, Hamburg, Deutschland

[2] Universität Hamburg, Hamburg, Deutschland

Der Begriff der Datenkompetenz bzw. Data Literacy ist noch jüngeren Datums. Er umfasst die Fähigkeiten, Daten auf kritische Art und Weise zu sammeln, zu managen, zu bewerten und anzuwenden [2]. Datenkompetenz wird zu einer Schlüsselkompetenz für das 21. Jahrhundert, in dem Daten den Rohstoff zu Wissens- und Wertschöpfung in den unterschiedlichsten Zusammenhängen darstellen.

Kategorien der Datenkompetenz

Das Hochschulforum Digitalisierung beschreibt in seinem Bericht „Future Skills: Ansätze zur Vermittlung von Data Literacy in der Hochschulbildung“ [3–5] fünf Kompetenzbereiche. Sie orientieren sich am Lebenszyklus von Daten von der Erzeugung bis zur Nutzung. Zunächst soll in einem **konzeptionellen Rahmen** das Wissen über und das Verständnis für Daten aufgebaut werden, um deren Nutzung und Anwendung verstehen zu können. Der zweite Bereich umfasst die **Datensammlung**. Wir erfassen Daten aus den unterschiedlichsten Quellen, z. B. aus Messgeräten und Sensoren, in Form von Berechnungsergebnissen aus wissenschaftlich-technischen Simulationen, aber auch z. B. aus Nachrichten in sozialen Medien. Diese Datenquellen müssen bezüglich ihrer Zuverlässigkeit und der Qualität der Daten kritisch bewertet werden. Im Kompetenzbereich **Datenmanagement** befasst man sich mit der Qualität der Daten. Daten werden kuratiert, d. h. Anomalien werden beseitigt, Ausreißer entfernt. Datenformate werden konsolidiert und dabei Daten gegebenenfalls konvertiert. Den Schwerpunkt stellt die Annotierung der Daten mit Metainformationen dar. Nur ausreichend annotierte Daten können zu einem

späteren Zeitpunkt weiterverwertet werden. Das Datenmanagement befasst sich auch mit der Datenspeicherung und ggf. einer Langzeitarchivierung. Im Folgenden werden die Daten ausgewertet. Der Kompetenzbereich **Datenevaluation** erfasst die numerische und grafische Auswertung von Daten mittels geeigneter Methoden und Werkzeuge. Die Daten werden interpretiert und präsentiert und im Rahmen von Entscheidungsfindungsprozessen verwertet. Der Kompetenzbereich **Datenanwendung** schließlich befasst sich schwerpunktmäßig mit Fragen der Datenethik, der Datenzitierung, der Datenverteilung und der Evaluierung von datenbasierten Entscheidungen.

Die Kategorisierung von Datenkompetenz ist unabhängig vom konkreten Anwendungsgebiet. Für jeden Kompetenzbereich lassen sich Lehrinhalte und Lernziele definieren, um bereits im Schulunterricht, spätestens an der Hochschule Grundkenntnisse und ggf. erweiterte Kompetenzen aufzubauen. Auch innerhalb der Gesellschaft für Informatik befasst man sich intensiv mit diesen Fragestellungen [6].

Warum ist es wichtig?

Datenkompetenz ist in vielen Bereichen des modernen Lebens unerlässlich. Wir sind heute beständig von Verfahren der Datensammlung, des Datenmanagements, der Datenevaluierung und der Datenanwendung umgeben und sollten diese Vorgänge beurteilen und bewerten können. Wenn wir selber diese Verfahren gestalten, sollten wir über detaillierte Kenntnisse in diesen Kompetenzbereichen verfügen.

Betrachten wir das Beispiel der intelligenten Stromzähler in Privathaushalten, sogenannte Smart Meter. Sie gestatten die Erfassung der Stromverbrauchsdaten in verschiedenen Zeitrastern. Die Daten des einzelnen Kunden werden zu einer Auswertestelle übertragen und dort gespeichert. Der Stromanbieter kuratiert die Daten, z. B. werden Übertragungsfehler korrigiert oder automatische Metadaten mit den Verbrauchsdaten verknüpft. Danach werden die Daten evaluiert. Der Anbieter nutzt sie sowohl zur Optimierung der Bereitstellung seines Angebots als auch potenziell zur Klassifizierung der Verbraucher. Eventuell teilt er die Verbraucher in Kategorien ein, die unterschiedliche Preise zahlen müssen. Im Bereich Datenanwendung könnte der Stromanbieter z. B. erwägen, die Kundeninformationen an Versicherungen zu verkaufen, die ihrerseits eine Wertschöpfung mit diesen Daten betreiben. Der Stromzähler in unserem Haus ist nur eine Komponente, die Daten erfasst. Jeder Nutzer von Diensten wie Alexa u. a. gibt freiwillig eine Vielzahl von Daten über sein Verhalten preis, die alle dem oben angeführten Lebenszyklus unterliegen. Diese Daten sind die Geschäftsgrundlage von Dienstanbietern, und jeder sollte über ausreichend Datenkompetenz verfügen, um die Facetten dieses Geschäfts verstehen und beurteilen zu können.

In Wissenschaft und Industrie ist heute in sehr vielen Bereichen eine vertiefte Datenkompetenz unerlässlich. Dies soll nun näher beleuchtet werden.

Datenkompetenz für Data Science

Im Jahr 2009 veröffentlichten Tony Hey et al. das Buch *The Fourth Paradigm – Data-Intensive Scientific Discovery* [7]. Das Buch ist dem 2007 auf See verschollenen Turing-Award-Gewinner Jim Gray gewidmet, der diese Ideen bereits Ende der 1990er-Jahre in den USA und in Zusammenhang mit den Ansätzen zu E-Science in England entwickelte [8]. Die Autoren zeigen an vielen Beispielen eindrucksvoll auf, wie moderne datengetriebene Wissenschaft funktioniert. Die stetig anwachsenden Datenmengen, die sowohl die Voraussetzung als auch die Folge dieses neuen Wissenschaftsparadigmas sind, werden als Big Data und Data Deluge bezeichnet. Der Begriff der Data Science entwickelt sich und findet zunächst Anwendung in den Natur- und Ingenieurwissenschaften und in Wirtschaft und Industrie, wenn große Datenmengen in die Wissens- und Wertschöpfung einbezogen werden. Die Anzahl der Stellenangebote für Data Scientists steigt rasant an. Gesucht werden Personen mit Kompetenzen im Bereich der Informatik (Datenbanken, Datenspeicherung, ...), der Mathematik und Statistik und mit Domänenwissen. Dies alleine ist jedoch nicht ausreichend. Bereits 2012 charakterisieren Davenport/Patil „The Sexiest Job of the 21st Century“ im Wesentlichen durch die oben genannten Kategorien der Datenkompetenz [9].

Am Beispiel der Klimamodellierung soll dies für einen Wissenschaftsbereich etwas detaillierter dargestellt werden. Die Phase der Datenerzeugung, wir sprechen hier von Forschungsdaten, findet auf einem Hochleistungsrechnersystem statt. Klimamodelle sind komplexe Computerprogramme, die numerische Simulationen durchführen. Sie weisen lange Laufzeiten auf und erzeugen Datenmengen im Bereich vieler Tera- und Petabyte. Die rohen Ergebnisdaten werden kuratiert, Lücken werden geschlossen, Fehler korrigiert, und es werden umfangreiche Metadaten hinzugefügt. Diese Annotationen ermöglichen die spätere Weiterverwertung. Die Ergebnisdaten werden aggregiert und numerisch und visuell ausgewertet. Im Rahmen der Datenanwendung werden die Datensätze für Zitationen aufbereitet und in E-Science-Umgebungen für eine potenziell weltweite Weiternutzung bereitgestellt. Das neue Forschungsparadigma, die vierte Säule der Erkenntnisgewinnung neben Theorie, Experiment und Simulation, besteht in der datengetriebenen Wissenschaft, bei der disziplinübergreifend und disziplinzusammenführend Daten z. B. aus der Klimaforschung in der Sozialforschung oder Versicherungswirtschaft weiter-

genutzt werden können. Dies exemplifiziert nochmals die Aussage „Daten sind der Rohstoff des 21. Jahrhunderts“.

Das Beispiel der Klimamodellierung ist typisch für viele Natur- und Ingenieurwissenschaften. Wir finden diese Methoden in der Biologie, Chemie, Physik, im Automobil- und Flugzeugbau und anderen Bereichen. Data Science ist die Basis der Wissensgewinnung. Im Zug der Digitalisierung hat dieses Vorgehen auch in anderen Wissenschaftsbereichen Einzug gehalten. Im geisteswissenschaftlichen Bereich findet wir mit Digital Humanities sogar eine neue Unterdisziplin.

Datenkompetenz und Institutionen

Die Durchdringung der Wissenschaft mit einem neuen Forschungsparadigma erfordert natürlich auch institutionelle Maßnahmen. Die Gemeinsame Wissenschaftskonferenz (GWK) beschloss im November 2013 die Einrichtung eines Rats für Informationsinfrastrukturen (RfII). Aufgabe des Rates ist es, „die Transparenz der Entwicklungen und Prozesse auf dem Gebiet der Informationsinfrastrukturen [zu] erhöhen sowie die Entwicklung und Vermittlung deutscher Positionen in europäischen und internationalen Debatten zu unterstützen“ [10]. In seiner ersten Mandatsphase hat sich der Rat insbesondere auf die Fragen zu Forschungsdaten, Nachhaltigkeit und Internationalität konzentriert. Aus einer umfassenden Länderanalyse aus dem Jahr 2017 leitet der RfII Empfehlungen für den Aufbau einer Nationalen Forschungsdateninfrastruktur für Deutschland ab. „Die gegenwärtig schwach koordinierte und nicht nachhaltig förderbare Landschaft der Dateninfrastrukturen in der Wissenschaft wird so in eine effizientere und kooperativere Richtung gelenkt. Eine Systematisierung der Datenbestände, gute Zugänglichkeit der Forschungsdaten und kontinuierliche Weiterentwicklung der Dienste stärken die Forschung in Deutschland und ihre globale Wettbewerbsposition“ [11].

Die Umsetzung der NFDI wird seit 2018 vorangetrieben. Die ersten neun ausgewählten Konsortien starten im Oktober 2020. Als Ziele sind hierbei definiert:

> „* Nachhaltige, qualitative und systematische Sicherung, Erschließung und Nutzbarmachung von Forschungsdaten über regionale und vernetzte Wissensspeicher
> * Etablierung eines Forschungsdatenmanagements nach den FAIR-Prinzipien
> * Anbindung und Vernetzung zu internationalen Initiativen wie der European Open Science Cloud“ [12, 13].

Die genannten FAIR-Prinzipien definieren vier Grundprinzipien für den Umgang mit Forschungsdaten [14, 15]:

- F („findability“) – die Daten müssen auffindbar sein
- A („accessibility“) – die Daten sind über ihren Identifikator zugreifbar
- I („interoperability“) – die Daten können mit anderen Daten verknüpft werden und in Anwendungen und Workflows genutzt werden
- R („reusability“) – die Daten und Metadaten sollten in anderen Zusammenhängen nutzbar sein

Die vier Prinzipien sind in weitere Unterprinzipien aufgeschlüsselt, um eine Operationalisierbarkeit zu ermöglichen. Ein wichtiger Teilaspekt der Qualität von Forschungsdaten bemisst sich aktuell am Umsetzungsgrad der FAIR-Prinzipien. Selbstverständlich müssen Forscherinnen und Forscher über die nötigen Datenkompetenzen verfügen, um dies zu realisieren.

Da die Förderung der NFDI keinen Ausbau von Forschungsdateninfrastrukturen umfasst, muss dieser in einem getrennten Förderrahmen stattfinden. Eine Maßnahme auf

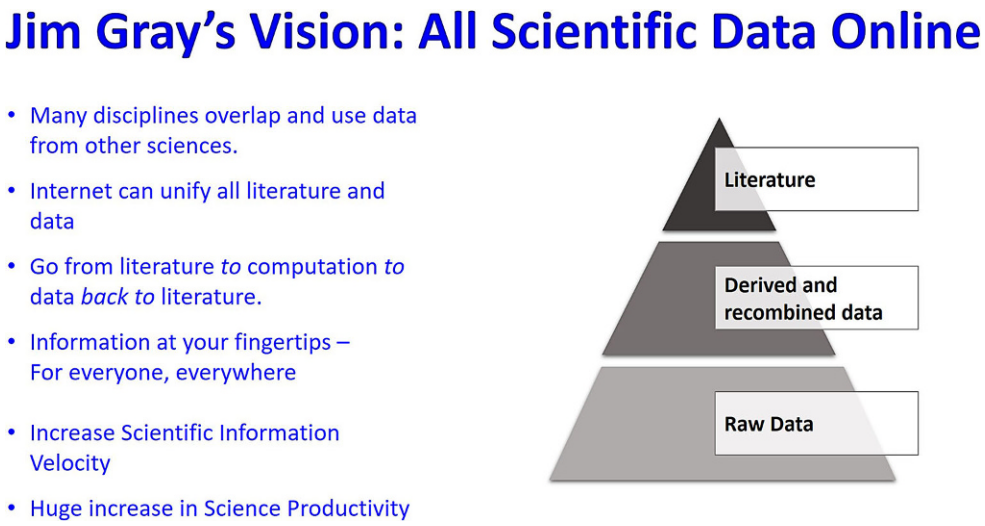

Abb. 1 Jim Grays Vision der vereinheitlichten Forschungsdaten. Folie aus seinem letzten Vortrag

europäischer Ebene ist die Konzeption und Implementierung einer offenen Cloud-Umgebung für Wissenschaftsdaten, der European Open Science Cloud (EOSC) [16]. Auf nationaler Ebene übernehmen aktuell vielfach die Hochleistungsrechenzentren diese Aufgabe.

Die Verbindung all dieser Ansätze wird dazu beitragen, die von Jim Gray in seinem letzten Vortrag formulierten Visionen zur Umsetzung zu bringen (Abb. 1; [8]) und sie mit moderner Informationstechnologie noch zu übertreffen. Die Ausbildung einer umfassenden Datenkompetenz ist hierbei eine der dringendsten Herausforderungen, mit der wir im digitalen Zeitalter konfrontiert sind. Sie stellt eine Schlüsselkompetenz für die Aufgaben des 21. Jahrhunderts in der Wissenschaft und darüber hinaus dar.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

1. Reinsel D, Gantz J, Rydning J The digitization of the world – from edge to core, November 2018 (data refreshed May 2020), IDC white paper #US44413318. https://www.seagate.com/files/www-content/our-story/trends/files/dataage-idc-report-final.pdf. Zugegriffen: 26. Sept. 2020
2. Datenkompetenz. https://de.wikipedia.org/wiki/Datenkompetenz. Zugegriffen: 26. Sept. 2020
3. Hochschulforums Digitalisierung. https://hochschulforumdigitalisierung.de/. Zugegriffen: 26. Sept. 2020
4. Heidrich J, Bauer P, Krupka D (2018) Ansätze zur Vermittlung von Data-Literacy-Kompetenzen, Hochschulform Digitalisierung, Nr. 47. https://gi.de/fileadmin/GI/Hauptseite/Aktuelles/Aktionen/Data_Literacy/HFD_AP37_DALI_Studie_2018-09.pdf. Zugegriffen: 26. Sept. 2020
5. Schüller K, Busch P, Hindinger C Hochschulforum Digitalisierung NR. 47/August 2019 Future Skills: Ein Framework für Data Literacy. https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_AP_Nr_47_DALI_Kompetenzrahmen_WEB.pdf. Zugegriffen: 26. Sept. 2020
6. Data Literacy: Digitale Kompetenzen in der Hochschule. https://gi.de/dataliteracy. Zugegriffen: 26. Sept. 2020
7. Hey T et al (Hrsg) (2009) The fourth paradigm – data-intensive scientific discovery. Microsoft Research, Washington
8. Hey T, Trefethen A (2019) The fourth paradigm 10 years on. Informatik Spektrum. https://doi.org/10.1007/s00287-019-01215-9
9. Davenport TH, Patil DJ (2012) Data scientist: the sexiest job of the 21st century. Harward Business Review. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century. Zugegriffen: 26. Sept. 2020
10. Rat für Informationsinfrastrukturen http://www.rfii.de/. Zugegriffen: 26. Sept. 2020
11. Rat für Informationsinfrastrukturen Zur Nationalen Forschungsdateninfrastruktur. http://www.rfii.de/de/themen/. Zugegriffen: 26. Sept. 2020
12. Nationale Forschungsdateninfrastruktur https://www.nfdi.de/. Zugegriffen: 26. Sept. 2020
13. Nationale Forschungsdateninfrastruktur Entstehung, Struktur und Aufgaben. https://www.nfdi.de/informationen. Zugegriffen: 26. Sept. 2020
14. GO FAIR https://www.go-fair.org/. Zugegriffen: 26. Sept. 2020
15. Wilkinson MD et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18
16. European Open Science Cloud (EOSC). https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud. Zugegriffen: 26. Sept. 2020