Physikalische Chemie, Universität Rostock

Vorlesung: Forschungsdatenmanagement
im Sommersemester 2024

Dr. habil. Till Biskup

— Glossar zu Vorlesung 02: „Warum Forschungsdatenmanagement?
(II) Zunahme von Datenmenge und Digitalität “ —

Hinweis: Die nachfolgend genannten Begriffe und Definitionen erheben keinen Anspruch auf formale Korrektheit, sondern dienen lediglich dem besseren Verständnis der in der Vorlesung behandelten Themen und sind im jeweiligen Kontext zu sehen. Mehrfache, voneinander abweichende Definitionen in unterschiedlichen Kontexten sind daher möglich. Fremdsprachige Begriffe werden nach Möglichkeit übersetzt, erscheinen aber ggf. unter ihrem ursprünglichen Namen in der Liste. Verweise auf andere Begriffe innerhalb des Glosssars sind durch das vorangestellte Symbol ↑ gekennzeichnet.

Automatisierung *automation* Strategie, um sich die manuelle Durchführung repetitiver und meist langweiliger Prozesse zu ersparen, indem sie an Maschinen ausgelagert wird. Automatisierung sorgt für Konsistenz (aber nicht Fehlerfreiheit) und ermöglicht es den menschlichen Akteuren, ihre dadurch freiwerdende Kapazität auf die eigentlichen intellektuellen Aufgaben, die weder automatisiert noch von Algorithmen übernommen werden können, zu verwenden.

Big Data nach Gartner Informationen mit großem Umfang, hoher Geschwindigkeit und/oder großer Vielfalt, deren Verarbeitung kosteneffiziente und innovative Werkzeuge erfordert, die sich positiv auf Einblicke, Entscheidungsfindungen und die Automatisierung von Prozessen auswirken. Ein Wesensmerkmal von Big Data ist die Verwendung der Daten in einem anderen Kontext als jenem, in dem sie ursprünglich erhoben wurden. Das führt meist zu einer ganzen Reihe von Problemen, die aber den Datennutzenden nicht unbedingt bewusst sind. Insbesondere ist der Kontext der Datenerhebung (fast) nie ausreichend dokumentiert (und dokumentierbar), um einer fachfremden Person die Einschätzung zu erlauben, ob die Verwendung der Daten im gegebenen Kontext zulässig ist.

Big Science siehe ↑Großforschung

Datenbank *database* (DB) auch Datenbanksystem genannt; System zur elektronischen Datenverwaltung. Die wesentliche Aufgabe einer Datenbank besteht darin, große Datenmengen effizient, widerspruchsfrei und dauerhaft zu speichern und benötigte Teilmengen in unterschiedlichen, bedarfsgerechten Darstellungsformen für Benutzer und Anwendungsprogramme bereitzustellen. Eine Datenbank besteht aus zwei Teilen: der Verwaltungssoftware (Datenbankmanagementsystem, DBMS), und der Menge der zu verwaltenden Daten, der Datenbank (DB) im engeren Sinn. Die Verwaltungssoftware organisiert intern die strukturierte Speicherung der Daten und kontrolliert alle lesenden und schreibenden Zugriffe auf die Datenbank. Zur Abfrage und Verwaltung der Daten bietet ein Datenbanksystem eine Datenbanksprache (eine eigenständige Programmiersprache, häufig SQL) an.

datengetriebene Wissenschaft „viertes Paradigma“, von Jim Gray [1] maßgeblich geprägter Begriff; beschreibt das Betreiben von Wissenschaft ausgehend von verfügbaren Daten. Die Fragestellung wird durch die Daten und deren Verfügbarkeit bestimmt, nicht umgekehrt.

Nur möglich durch die unter dem Begriff ↑e-Science zusammengefassten Werkzeuge und Infrastrukturen.

Erkenntnis Aneignung des Sinngehalts von erlebten bzw. erfahrenen Sachverhalten, Zuständen oder Vorgängen, Ergebnis des Vorgangs des Erkennens. Erkenntnis beinhaltet immer eine auf die Erfahrung gestützte Beurteilung und setzt notwendiger Weise ein Subjekt voraus, das erkennt. Neue Erkenntnisse, die von innerer und äußerer Erfahrung unabhängig sind, sind immer Ergebnis einer schöpferischen Phantasie. Bei der Erkenntnis stehen sich Subjekt und Objekt als Erkennendes und Erkanntes gegenüber. Die Erkenntnis führt zu einem Abbild des Objekts im Subjekt. Die grundsätzliche Unvollständigkeit dieses Abbilds ist die Triebkraft hinter dem Erkenntnisgewinn und letztlich der ↑Wissenschaft. Vgl. [2]

e-Science Summe der digitalen Werkzeuge und der notwendigen digitalen Infrastruktur, um mit großen Datenmengen umzugehen; Voraussetzung für die ↑datengetriebene Wissenschaft, aber von dieser unabhängig.

FAIR Akronym für die vier Begriffe *findable* (auffindbar), *accessible* (zugreifbar), *interoperable* (interoperabel) und *reusable* (wiederverwendbar); von Wilkinson *et al.* [3] unter dem vollständigen Titel „The FAIR Guiding Principles for scientific data management and stewardship“ berühmt gemachte Prinzipien, die aus der ↑datengetriebenen Wissenschaft und der Verwendung von ↑künstlicher Intelligenz zur Verarbeitung großer Datenmengen kommen. Oft missverstanden als tragfähiges Grundkonzept für Forschungsdatenmanagement. Für die meisten Forschenden in ihrer originalen Form eher irrelevant, aber für die Wissenschaft und den Erkenntnisgewinn tendenziell gefährlich.

Großforschung *Big Science*, außeruniversitär, meist in großen Forschungseinrichtungen teils quasi-industriell betriebene Form der Wissenschaft, die seit Mitte des 20. Jh. verstärkt auftritt. Meist wird der Beginn mit dem Manhattan-Projekt (Los Alamos National Laboratory) gleichgesetzt. Als Begriff von Alvin Weinberg [4] eingeführt und von Derek J. de Solla Price [5] berühmt gemacht. Nach dem zweiten Weltkrieg einsetzendes Phänomen, dass sich die Politik stark für die Wissenschaft interessierte und große Forschungsstandorte mit sehr viel Geld förderte. Die wissenschaftlichen Fragestellungen werden von den Großforschungsgeräten und -Einrichtungen bestimmt, nicht umgekehrt. Nach Weinberg besteht die Gefahr, dass das nun vorhandene Geld ausgegeben wird, anstatt erst sinnvoll nachzudenken.

Infrastruktur personelle, sachliche und finanzielle Ausstattung, um ein angestrebtes Ziel zu erreichen.

Kausalität Ursache-Wirkungs-Beziehung zwischen Ereignissen und Zuständen: A ist die Ursache von B, wenn B von A erzeugt wird. Vgl. ↑Korrelation, ↑Koinzidenz

Koinzidenz zeitliches oder/und räumliches Zusammentreffen von Ereignissen oder Objekten; deskriptiver Begriff ohne Implikation von Zusammenhängen. Vgl. ↑Kausalität, ↑Korrelation

kontrolliertes Vokabular *controlled vocabulary*, Sammlung von Begriffen mit dem Ziel, die Beschreibung von Objekten zu vereinheitlichen. Innerhalb des kontrollierten Vokabulars sind die Begriffe eindeutig identifiziert. Ein Beispiel aus der Informatik für kontrollierte Vokabulare wäre der Aufzählungstyp (*enumeration type*). Vgl. ↑Ontologie

Korrelation Maß für den Zusammenhang zweier Größen; mathematisch i.d. R. auf $|[0..1]|$ beschränkt, wobei negative Werte eine Antikorrelation anzeigen. Vgl. ↑Kausalität, ↑Koinzidenz

künstliche Intelligenz (KI), meist besser beschrieben als „maschinelles Lernen“ (ML); aktuell wieder einmal sehr populär und als Heilsversprechen gehandelt. Letztlich in seiner momentanen Ausprägung die Anwendung (komplexerer) statistischer Algorithmen auf große Datenmengen.

Metadaten Informationen zu den numerischen Daten, notwendige Voraussetzung für eine sinnvolle Verarbeitung der Daten im Kontext eines ↑Systems zur Datenverarbeitung und für ↑nachvollziehbare Wissenschaft.

nachvollziehbare Wissenschaft *reproducible science*, seit der Etablierung rechnergestützter Datenauswertung eigentlich nie mehr erreichter, aber für die Wissenschaft konstituierender Aspekt, dass sich Ergebnisse und Auswertungen unabhängig nachvollziehen lassen, weil alle dazu notwendigen Aspekte vollständig und ausreichend beschrieben wurden (↑Nachvollziehbarkeit). Motivation für die Vorlesung, deren Ziel es ist, die Hörer mit Konzepten vertraut zu machen, die letztlich eine ernstzunehmende nachvollziehbare Wissenschaft ermöglichen. Die ↑Nachvollziehbarkeit geht dabei weit über ↑Replizierbarkeit und ↑Reproduzierbarkeit hinaus.

Nachvollziehbarkeit zentraler Aspekt der ↑Wissenschaft und der wissenschaftlichen Methode, die die Intersubjektivität ihrer Aussagen ermöglicht. Setzt in der Regel eine ↑hinreichende Beschreibung (und Dokumentation) der einzelnen Schritte voraus, die von einem gegebenen Ausgangspunkt zu einem (neuen) Ergebnis oder auch einer Erkenntnis kommt.

Ontologie *ontology*, in der Informatik die Darstellung der Eigenschaften eines Fachgebiets und ihre Beziehungen zueinander, indem eine Reihe von Konzepten und Kategorien definiert wird, die das Fachgebiet repräsentieren. Formale Ontologien spielen im *Semantic Web* und in der Anwendung von ↑Künstlicher Intelligenz eine große Rolle, weil sie ermöglichen (sollen), unterschiedlich formuliertes Wissen automatisiert verarbeitbar zu machen und implizite ↑Metadaten und Beziehungen von Begriffen und Konzepten untereinander für die Maschine explizit zu machen. Vgl. ↑kontrolliertes Vokabular

Paradigma nach Thomas S. Kuhn [6] ein Satz allgemein anerkannter wissenschaftlicher Leistungen, der für eine gewisse Zeit einer Gemeinschaft von Fachleuten maßgebende Probleme und Lösungen liefert

Persistenz Fähigkeit, Daten oder logische Verbindungen über lange Zeit (insbesondere über einen Programmabbruch hinaus) bereitzuhalten; benötigt ein nichtflüchtiges Speichermedium.

PID *persistent identifier*, dt. dauerhafte Kennung, i.d.R. eineindeutige und langzeitstabile Kennung für physische oder digitale Objekte. Bekannte und weit verbreitete PIDs sind z.B. der *digital object identifier* (DOI), aber auch die *International Standard Book Number* (ISBN) oder die *Open Researcher and Contributor ID* (ORCID). Vgl. ↑Persistenz

Replizierbarkeit *replicability*, unabhängige Wiederholbarkeit der (Roh-)Datenerhebung, meist in Form von Experimenten und Beobachtungen, entsprechend nicht in jedem Fall durchführbar. Vgl. ↑Reproduzierbarkeit, ↑Robustheit, ↑Verallgemeinerbarkeit.

Reproduzierbarkeit *reproducibility*, vollständige Wiederholbarkeit einer beschriebenen Datenverarbeitung und -Analyse. Ausgangspunkt sind existierende Daten, entsprechend sollte sie in jedem Fall möglich sein. Vgl. ↑Replizierbarkeit.

Robustheit *robustness*, im Kontext der Datenverarbeitung die Tatsache, dass unterschiedliche, unabhängige Analysen derselben Daten zum gleichen Ergebnis führen. Vgl. ↑Reproduzierbarkeit, ↑Replizierbarkeit, ↑Verallgemeinerbarkeit

System zur Datenverarbeitung hier: Gesamtsystem für wissenschaftliche Datenverarbeitung von der Datenaufnahme bis zur fertigen Publikation, das alle Aspekte umfasst und das ↑nachvollziehbare Wissenschaft möglich macht und gewährleistet. Definitiv ein größeres Projekt, das nicht nur eine ↑monolithische Anwendung umfasst, sondern viele Aspekte darüber hinaus. Setzt entsprechende ↑Infrastruktur und in der Umsetzung der einzelnen Komponenten sauberen Code und eine solide Softwarearchitektur voraus.

Verallgemeinerbarkeit auch: Generalisierbarkeit, *generalisability*, im Kontext der Datenverarbeitung die Tatsache, dass sowohl unabhängig erhobene Daten als auch voneinander unabhängige Analysemethoden zum gleichen Ergebnis führen. Baustein zur unabhängigen Bestätigung wissenschaftlicher Hypothesen. Vgl. ↑Reproduzierbarkeit, ↑Replizierbarkeit, ↑Robustheit

viertes Paradigma *fourth paradigm*, siehe ↑datengetriebene Wissenschaft

Wissenschaft Auf den Erkenntnisgewinn ausgerichtetes, systematisches menschliches Unterfangen, das in der Regel eine Reihe von Kriterien erfüllt bzw. erfüllen sollte: Unabhängigkeit vom Beobachtenden bzw. Durchführenden, gegründet auf den Erkenntnissen früherer Generationen, sowie überprüfbar, nachvollziehbar und ggf. reproduzierbar. Für Einführungen vgl. u.a. [7, 8].

Literatur

[1] Tony Hey, Stewart Tansley und Kristin Tolle, Hrsg. *The Fourth Paradigm*. Redmont, Washington: Microsoft Research, 2009.

[2] Heinrich Schmidt. *Philosophisches Wörterbuch*. 22. Aufl. Neu bearbeitet von Georgi Schischkoff. Stuttgart: Kröner, 1991.

[3] Mark D. Wilkinson u. a. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (2016), S. 160018. DOI: 10.1038/sdata.2016.18.

[4] Alvin Weinberg. Impact of large-scale science on the United States. *Science* 134 (1961), S. 161–164. DOI: 10.1126/science.134.3473.161.

[5] Derek John Price de Solla. *Little science, big science*. New York: Columbia University Press, 1963.

[6] Thomas S. Kuhn. *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt am Main: Suhrkamp, 1976.

[7] Alan F. Chalmers. *What is this thing called Science?* Third edition. Berkshire, UK: Open University Press, 1999.

[8] Hans Poser. *Wissenschaftstheorie*. Stuttgart: Reclam, 2001.