



Physikalische Chemie, Universität Rostock

**Vorlesung: Forschungsdatenmanagement
im Sommersemester 2024**

Dr. habil. Till Biskup

— Glossar zu Vorlesung 13: „Notwendige Kompetenzen“ —

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 Glossars sind durch das vorangestellte Symbol ↑ gekennzeichnet.

Abstraktion Nach Edsger Dijkstra [1] das einzige mentale Werkzeug, das es erlaubt, eine große Vielzahl von Fällen abzudecken. Zweck der Abstraktion ist es nicht, vage zu sein, sondern im Gegenteil ein neues Bedeutungsniveau zu schaffen, das präzise Beschreibungen erlaubt.

Analyse Systematische Zerlegung eines zu untersuchenden Gegenstands oder Sachverhalts in seine Bestandteile, die auf Grundlage vorab festgelegter Kriterien erfasst, geordnet, untersucht und ausgewertet werden.

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.

datengetriebene Wissenschaft „viertes Paradigma“, von Jim Gray [2] 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.

Digitalkompetenz Beherrschung digitaler Werkzeuge zur Automatisierung von Abläufen

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. [3]; wesentliche Beiträge zur Erkenntnistheorie und ihrer Anwendung auf die Naturwissenschaft kommen von Kant [4, 5].

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.

Fachkompetenz Gemäß der KMK die „Bereitschaft und Fähigkeit, auf der Grundlage fachlichen Wissens und Könnens Aufgaben und Probleme zielorientiert, sachgerecht, methodengeleitet und selbstständig zu lösen und das Ergebnis zu beurteilen.“ [6, S. 15]

Forschungsdaten zunächst einmal Daten, die im Zuge wissenschaftlicher Vorhaben im Rahmen von Forschung z.B. durch Digitalisierung, Quellenforschungen, Experimente, Messungen, Erhebungen oder Befragungen entstehen. Forschungsdaten im weiteren Sinn umfassen darüber hinaus (physische) Objekte und Werkzeuge (z.B. Fragebögen, Software und Simulationen). Forschungsdaten können grundsätzlich analog oder digital vorliegen. Sie sind Ausgangspunkt der (empirischen) Wissenschaft.

Forschungsdatenmanagement Umgang mit ↑Forschungsdaten über ihren gesamten Lebenszyklus hinweg mit dem Fokus auf ↑Nachvollziehbarkeit und Nachnutzbarkeit; wird meist auf die digitale Welt bezogen, ist letztlich aber nichts anderes als sauberes wissenschaftliches Arbeiten; notwendige, aber nicht hinreichende Bedingung für den wissenschaftlichen Erkenntnisgewinn.

GIGO *garbage in garbage out*, griffig formuliertes „Prinzip“, nach dem die Ergebnisse in der Re-

gel nicht von besserer Qualität sein können als die Daten, auf denen sie basieren. Insbesondere im Kontext der Weiterverwendung von Daten relevant. Vgl. ↑Qualitätskontrolle, ↑Qualitätssicherung

hinreichend mathematisches Konzept, das eine Bedingung beschreibt, deren Erfüllung ausreicht, um ein gegebenes Ziel zu erreichen. Vgl. ↑notwendig

intellektuelle Beherrschbarkeit *intellectual manageability*, nach Edsger Dijkstra [1] das Hauptziel der Softwaretechnik (*software engineering*) – und letztlich des Projektmanagements. Unterschiedliche Lösungsansätze für ein Problem sind unterschiedlich gut intellektuell beherrschbar. Entsprechend ist die intellektuelle Beherrschbarkeit das zentrale Kriterium für die Entscheidung, welche Lösung für ein Problem bevorzugt wird.

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

Modularisierung Aufteilung der Gesamtaufgabe in kleinere Abschnitte. Die Aufteilung wird so lange fortgesetzt, bis die Lösung für den aktuellen Abschnitt unmittelbar in Form von Quellcode offensichtlich ist. Setzt die Definition von ↑Schnittstellen voraus.

Modularität Eigenschaft eines Systems, aus lauter separaten, durch ↑Schnittstellen miteinander verbundenen Teilen zu bestehen. I.d.R. Folge der ↑Modularisierung und einzig erfolgversprechende Strategie für die ↑intellektuelle Beherrschbarkeit komplexer Systeme.

Muster pattern, nach Christopher Alexander [7] abstrakte Beschreibung eines wiederkehrenden Problems sowie einer generellen Lösung für dieses Problem, deren konkrete Ausgestaltung meist hochgradig individuell ist. Ein wichtiger Teil der Beschreibung von Mustern ist eine Kosten–Nutzen-Analyse, die bei der Entscheidung über ihren Einsatz hilft. Muster wurden später mit explizitem Bezug auf C. Alexander in die Softwareentwicklung als Entwurfsmuster (*design patterns*) eingeführt [8].

nachvollziehbare Wissenschaft *reproducible science*, seit der Etablierung rechnergestützter Datenauswertung eigentlich nie mehr erreicht, 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.

notwendig mathematisches Konzept, das eine Bedingung beschreibt, die zwar erfüllt sein muss, um ein bestimmtes Ergebnis zu bekommen, aber für die Erfüllung nicht ausreicht. Vgl. ↑hinreichend

Plausibilität (kontextabhängiges) Beurteilungskriterium: etwas ist plausibel, wenn es möglich und wahrscheinlich erscheint.

Qualitätskontrolle Überprüfung der Qualität von Dingen oder Prozessen anhand vorher festgelegter Kriterien. Allgemeine Kriterien sind

↑Konsistenz und ↑Plausibilität. Wenn sich die Kriterien formal definieren und die relevanten Charakteristika der zu überprüfenden Dinge oder Prozesse ohne direkte menschliche Interaktion bestimmen lassen, ist eine Automatisierung möglich. Vgl. ↑Qualitätssicherung

Qualitätssicherung Sicherstellung der Qualität von Dingen oder Prozessen. Vgl. ↑Qualitätskontrolle

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

Repräsentativität Eigenschaft der ausgewählten Daten, die Variabilität der Grundgesamtheit wiederzugeben und damit Fehlschlüssen aufgrund nicht berücksichtigter Fälle tendenziell vorzubeugen. Voraussetzung für die ↑Verallgemeinerbarkeit und die Bildung wissenschaftlicher ↑Modelle. Gerade im Kontext von ↑Big Data oft unzulässig vernachlässigtes Kriterium mit entscheidenden Konsequenzen für die Resultate, vgl. ↑GIGO.

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

Rückführbarkeit hier: Möglichkeit, ein Ergebnis auf seine Quelle bzw. die zugrundeliegenden Daten und Beobachtungen zurückführen zu können. Umfasst auch die hinreichende Dokumentation des gesamten Weges von den (Roh-)Daten zum finalen Ergebnis und setzt i.d.R. ein ↑System zur Datenverarbeitung voraus. Notwendiges, aber nicht ↑hinreichendes Kriterium für die Datenqualität.

Schnittstelle der Teil eines Systems, der der Kommunikation und dem Austausch z.B. von Information dient. Systeme werden von außen als abgeschlossen (*black box*) betrachtet und kommunizieren ausschließlich über ihre Schnittstelle(n). Die explizite Definition, Dokumentation und Implementation von Schnittstellen sind wesentliche Voraussetzungen für ↑modulare ↑Systemarchitekturen. Schnittstellen ermöglichen die ↑Trennung der Belange. Oft genug stimmen Schnittstellen in Systemen mit Organisationsgrenzen beteiligter Gruppen überein [9]. In jedem Fall ist es essentiell, mit Systemen nur über deren Schnittstellen zu kommunizieren und *keine* Annahmen über die innere Organisation dieser Systeme zu treffen.

Systemarchitektur Summe der während der Entwicklung eines Systems getroffenen und in der Umsetzung manifestierten Entscheidungen. Nach [10] minimieren gute Architekturen die Zahl getroffener Entscheidungen.

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.

Trennung der Belange *separation of concerns*, grundlegendes Prinzip für ↑Modularisierung, nach Edsger Dijkstra [11] die einzig effektive Möglichkeit, seine Gedanken zu ordnen, indem man sich auf einen Aspekt eines ↑komplexen Problems fokussiert, ohne dabei zu vergessen, dass es lediglich ein Teilaspekt

ist.

Transparenz über die ↑Nachvollziehbarkeit hinausgehendes Konzept, das die Wege der Entscheidungsfindung inklusive verworfener oder nicht beschrittener Alternativen nach bestem Wissen und Gewissen umfassend dokumentiert. Von R. Feynman [12] als essentiell für die Wissenschaftlichkeit hervorgehoben.

Unix-Philosophie griffige Formulierung dreier Prinzipien, die wesentlich zum Erfolg des Unix-Betriebssystems und seiner Nachfolger (u.a. Linux, macOS, Android, iOS) beigetragen haben. Nach Salus [13, S. 53]: „Write programs that do one thing and do it well. Write programs to work together. Write programs that handle text streams, because that is a universal interface.“ Diese Prinzipien gelten für die Softwareentwicklung genauso wie für die analoge Welt: ↑Modularität und Interoperabilität sind entscheidend, um mit begrenzten Ressourcen Werkzeuge zu entwickeln und zu etablieren.

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

Wissenschaft Auf den Erkenntnisgewinn ausgerichteter, 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. [14, 15].

Literatur

- [1] Edsger W. Dijkstra. The humble programmer. *Communications of the ACM* 15 (1972), S. 859–865.
- [2] Tony Hey, Stewart Tansley und Kristin Tolle, Hrsg. *The Fourth Paradigm*. Redmont, Washington: Microsoft Research, 2009.
- [3] Heinrich Schmidt. *Philosophisches Wörterbuch*. 22. Aufl. Neu bearbeitet von Georgi Schischkoff. Stuttgart: Kröner, 1991.
- [4] Immanuel Kant. *Kritik der reinen Vernunft*. Herausgegeben von Wilhelm Weischedel. Frankfurt am Main: Suhrkamp, 1974.
- [5] Immanuel Kant. *Metaphysische Anfangsgründe der Naturwissenschaft*. Mit einer Einleitung herausgegeben von Konstantin Pollok. Hamburg: Felix Meiner Verlag, 1997.
- [6] Kultusministerkonferenz, Hrsg. *Handreichung für die Erarbeitung von Rahmenlehrplänen der Kultusministerkonferenz für den berufsbezogenen Unterricht in der Berufsschule und ihre Abstimmung mit Ausbildungsordnungen des Bundes für anerkannte Ausbildungsberufe*. 2021. URL: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2021/2021_06_17_-_GEP_-_Handreichung.pdf.
- [7] Christopher Alexander, Sara Ishikawa und Murray Silverstein. *A Pattern Language*. New York: Oxford University Press, 1977.
- [8] Erich Gamma u. a. *Design Patterns. Elements of Reusable Object-Oriented Software*. Boston: Addison-Wesley, 1995.
- [9] Melvin E. Conway. How do committees invent? *Datamation* 14.4 (1968), S. 28–31.
- [10] Robert C. Martin. *Clean Architecture. A Craftman's Guide to Software Structure and Design*. Boston: Prentice Hall, 2018.
- [11] Edsger W. Dijkstra. „On the Role of Scientific Thought (EWD447)“. In: *Selected Writings on Computing: A Personal Perspective*. New York: Springer-Verlag, 1982, S. 60–66.
- [12] Richard P. Feynman. Cargo cult science. *Engineering and Science* 37.7 (1974), S. 10–13. URL: <https://resolver.caltech.edu/CaltechES:37.7.CargoCult>.
- [13] Peter H. Salus. *A Quarter Century of UNIX*. Reading, MA: Addison-Wesley, 1994.
- [14] Alan F. Chalmers. *What is this thing called Science?* Third edition. Berkshire, UK: Open University Press, 1999.
- [15] Hans Poser. *Wissenschaftstheorie*. Stuttgart: Reclam, 2001.