Physikalische Chemie, Universität Rostock

Vorlesung: Forschungsdatenmanagement
im Sommersemester 2025

Dr. habil. Till Biskup

— Glossar zu Vorlesung 10: „Wiederverwenden“ —

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.

Datenformat digitales Speicherformat für Daten jeglicher Form. Grundsätzlich werden binäre und Textformate unterschieden. Während erstere meist mit deutlich geringerem Speicherbedarf auskommen, sind sie im Gegensatz zu letzteren nicht ohne Hilfsmittel lesbar. Textformate hingegen sind, ein beliebiger Texteditor vorausgesetzt, prinzipiell menschenlesbar.

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.

Denial of Service (DoS), Begriff aus der Informationstechnik für die Nichtverfügbarkeit eines Dienstes, meist durch Überlastung aufgrund zu vieler Zugriffe. Im übertragenen Sinn die Überflutung mit Informationsmengen, so dass ihre Verarbeitung schlicht nicht mehr möglich ist.

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

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.* [5] 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.

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 Regel 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

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

Katalog Werkzeug zum Auffinden und Erschließen von Forschungsdaten. ↑Forschungsdaten können mit Hilfe eines Datenkatalogs gesucht, gefunden und erschlossen werden (vgl. die ↑FAIR-Prinzipien). Ein Datenkatalog enthält vergleichbar zu einem Bibliothekskatalog verschiedene ↑Metadaten, die die Grundlage für die Suche und Filterung darstellen, aber nicht (notwendigerweise) die ↑Forschungsdaten selbst – im Falle der Bibliothek die Bücher. Typischerweise bieten auch ↑Repositorien grundständige Katalogfunktionen, so dass die Unterscheidung zwischen Katalog und Repositorium in der Praxis mitunter verschwimmt. Ein Katalog als Sammlung von ↑Metadaten zu bestimmten Objekten erweist sich insbesondere dann als sinnvoll, wenn die Menge der Objekte eine gewisse Schwelle überschreitet, die ein Auffinden und Abrufen über die einzelnen Objekte selbst unmöglich macht oder zumindest massiv erschwert.

Konsistenz hier: logische Widerspruchsfreiheit; Zusammenhang der Gedankenführung

Konvention innerhalb einer Gruppe oder einem (lokalen) Kontext getroffene (temporäre) Festlegung. Ziel von Konventionen ist die Vereinheitlichung und damit einhergehend die Befreiung von der Notwendigkeit, jedesmal aufs Neue nachdenken zu müssen, wie z.B. gewisse Prozesse durchgeführt oder Objekte benannt werden sollen. Konventionen sind im Gegensatz zu ↑Standards weniger verbindlich und deutlich flexibler sowie *ad hoc* innerhalb einer Gruppe einführbar. Vgl. ↑Standard

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.

Lizenz *license*, Nutzungsrecht; u.a. Software ist *per se* vom Urheberrecht geschützt, unabhängig von ihrer Funktionalität. Lizenzen übertragen Nutzungsrechte vom Urheber der Software an ihren Nutzer. Inwieweit ↑Forschungsdaten urheberrechtlich geschützt sind, ist eine in der Rechtsprechung noch nicht abschließend beantwortete Frage. Tendenziell sind Daten, die nicht weiter kuratiert wurden, nicht urheberrechtlich geschützt, da ihnen die nötige Schöpfungshöhe fehlt.

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

Modell vereinfachende Annäherung an die Wirklichkeit bzw. einen Forschungsgegenstand, die nur die als wesentlich erachteten Phänomene als Aspekte berücksichtigt. Ein Modell entsteht oft durch ↑Abstraktion und Verallgemeinerung von Beobachtungen. Mathematisch formulierte Modelle zeichnen sich häufig durch Parameter aus, durch die sie sich charakterisieren lassen und die ggf. variiert werden können, um das Modell an die Realität bzw. erhobene Daten anzupassen. Während Modelle der einzige Weg sind, die Realität zu erklären und so ggf. zu ↑Erkenntnis zu gelangen, sind sie immer vorläufig und beschränkt. Von George E. P. Box stammt der Satz „Alle Modelle sind falsch, aber manche sind nützlich" [6].

monolithisch aus einem Stück bestehend; zusammenhängend und fugenlos

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.

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

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

proprietär auf herstellerspezifischen, (meist) nicht veröffentlichten Standards basierend

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 Charakeristika 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.

Repositorium Publikationsplattform (u.a.) für ↑Forschungsdaten. Repositorien sind Publikationsplattformen (u.a.) für Forschungsdaten. Als IT-Dienst werden sie i.d.R. von Institutionen, Organisationen oder Firmen bereitgestellt und speichern die Forschungsdaten i.d.R. langfristig, dokumentieren die Forschungsdaten mit ↑Metadaten, regeln den Zugang (inkl. ↑Lizenz) zu den Forschungsdaten und vergeben einen ↑PID. Die dort publizierten Forschungsdaten sind meist über eine Metadatensuche und -filterung für Nutzerinnen und Nutzer auffindbar und erschließbar (Datenkatalog). Vgl. ↑Katalog

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 ↑Sysstem zur Datenverarbeitung voraus. Notwendiges, aber nicht ↑hinreichendes Kriterium für die Datenqualität.

Standard von einem oft internationalen und anerkannten Gremium definierte Festlegung. Standards sind im Gegensatz zu ↑Konvention sehr viel starrer und nicht *ad hoc* von einer Gruppe einführbar. Vgl. ↑Konvention

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.

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 [7] als essentiell für die Wissenschaftlichkeit hervorgehoben.

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 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. [8, 9].

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] Immanuel Kant. *Kritik der reinen Vernunft*. Herausgegeben von Wilhelm Weischedel. Frankfurt am Main: Suhrkamp, 1974.

[4] Immanuel Kant. *Metaphysische Anfangsgründe der Naturwissenschaft*. Mit einer Einleitung herausgegeben von Konstantin Pollok. Hamburg: Felix Meiner Verlag, 1997.

[5] 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.

[6] G.E.P. Box. „Robustness in the Strategy of Scientific Model Building“. In: *Robustness in Statistics*. Hrsg. von Robert L. Launer und Graham N. Wilkinson. Academic Press, 1979, S. 201–236. ISBN: 978-0-12-438150-6. DOI: https://doi.org/10.1016/B978-0-12-438150-6.50018-2.

[7] Richard P. Feynman. Cargo cult science. 37.7 (1974), S. 10–13. URL: https://resolver.caltech.edu/CaltechES:37.7.CargoCult.

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

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