Physikalische Chemie, Universität Rostock

Vorlesung: Forschungsdatenmanagement
im Sommersemester 2024

Dr. habil. Till Biskup

— Glossar zu Vorlesung 04: „Forschungsdatenlebenszyklus“ —

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.

Cargo Cult Phänomen, dass Indigene auf Südseeinseln, die im Zweiten Weltkrieg als US-amerikanische Stützpunkte dienten und über denen in dieser Zeit aus Flugzeugen große Mengen Nahrungsmittel und andere Güter abgeworfen wurden, nach dem Krieg, als die Lieferungen ausblieben, kultartig das Verhalten der Soldaten in der Hoffnung nachstellten, wieder Güter zu erhalten. So wurden Landebahnen und Flughafentower genauso imitiert wie das Verhalten der Soldaten auf den vorigen Flugplätzen.

Cargo Cult Science von Richard Feynman [1] eingeführter Begriff für eine Form der (vermeintlichen) Wissenschaft, die allen offensichtlichen Vorzeichen und Formen wissenschaftlicher Forschung folgt, der aber etwas Wesentliches fehlt: wissenschaftliche Integrität, Ehrlichkeit, rigoroses Hinterfragen der eigenen Ergebnisse und Erklärungen und Offenlegung aller relevanten Informationen nach bestem Wissen und Gewissen. Vgl. ↑Cargo Cult

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.

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.

FAIR Akronym für die vier Begriffe *findable* (auffindbar), *accessible* (zugreifbar), *interoperable* (interoperabel) und *reusable* (wiederverwendbar); von Wilkinson *et al.* [6] 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.

Forschung Systematisches Vorgehen, um einer bestimmten Fragestellung nachzugehen oder Phänomene zu erklären oder Experimente unter kontrollierten Bedingungen durchzuführen, das der wissenschaftlichen Methodik folgt. Wissenschaft setzt Forschung voraus. Allerdings kann Forschung ohne Beitrag zur Wissenschaft (ohne Erkenntnisgewinn) bleiben, vgl. Feynmans Begriff ↑Cargo Cult Science [1].

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.

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.

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.

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.

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

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

Wissenschaft Auf den Erkenntnisgewinn ausgerichtetes, systematisches menschliches Unterfangen, das in der Regel eine Reihe von Kri-

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

[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] 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.

[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.