Physikalische Chemie, Universität Rostock

Vorlesung: Forschungsdatenmanagement
im Sommersemester 2024

Dr. habil. Till Biskup

— Glossar zu Vorlesung 11: „Eigenschaften und Konzepte“ —

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.

CERN Europäische Organisation für Kernforschung, Großforschungseinrichtung in der Nähe von Genf, in der physikalische Grundlagenforschung insbesondere auf dem Gebiet der Teilchenphysik betrieben wird.

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.

dauerhafte Kennung ↑PID

DOI *digital object identifier*, möglichst eindeutige (↑UID) und dauerhafte (↑PID) digitale Kennung für physische, digitale oder abstrakte Objekte. DOIs werden bislang vor allem für digital verfügbare wissenschaftliche Fachliteratur verwendet, zunehmend aber auch für andere digitale Artefakte.

eierlegende Wollmilchsau scherzhafte, umgangssprachliche Bezeichnung für eine Person oder Sache, die allen Ansprüchen genügt, alle Bedürfnisse befriedigt und keine Nachteile hat

eindeutige Kennung ↑UID

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. [1]; wesentliche Beiträge zur Erkenntnistheorie und ihrer Anwendung auf die Naturwissenschaft kommen von Kant [2, 3].

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.

Hash ↑Prüfsumme

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

Hypertext ursprünglich von Ted Nelson 1965 [4] eingeführtes Konzept für einen Text, der nicht linear sein muss, sondern Verweise auf andere Texte oder Textstellen enthält. Wurde durch die Entwicklung des ↑WWW durch Berners-Lee [5] weltweit bekannt.

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

intellektuelle Beherrschbarkeit *intellectual manageability*, nach Edsger Dijkstra [6] 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.

Kollisionsresistenz zu einer aus einer gegebenen Zeichenkette A errechneten ↑Prüfsumme existiert keine von A verschiedene Zeichenkette B, die zur gleichen Prüfsumme führt. Die Kollisionsresistenz einer ↑kryptographischen Hashfunktion ist in der Praxis nie vollständig, aber ggf. hinreichend unwahrscheinlich.

Kontrolliertes Vokabular Sammlung von Begriffen mit dem Ziel, die Beschreibung von Objekten zu vereinheitlichen. Innerhalb des kontrollierten Vokabulars sind die Begriffe eindeutig identifiziert.

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

Kopplung *coupling*, Grad der Verbindung zweier Gegenstände oder Systeme. Systemarchitektur zielt generell auf eine lose Kopplung (*loose coupling*) einzelner Komponenten ab, um die ↑Modularität zu erhalten.

Kristallkugel Nur in der Theorie funktionierendes Hilfsmittel für den Blick in die Zukunft, das u.a. hilfreich wäre, um Software bereits in ihrer Entstehung auf künftige Anforderungen hin auszulegen. Aufgrund anderer damit einhergehender Probleme ist die reale Funktionalität einer Kristallkugel nicht wünschenswert.

Kryptographie ursprünglich die Wissenschaft der Verschlüsselung von Informationen. Heute befasst sie sich auch allgemein mit dem Thema Informationssicherheit, also der Konzeption, Definition und Konstruktion von Informationssystemen, die widerstandsfähig gegen Manipulation und unbefugtes Lesen sind.

kryptographische Hashfunktion Funktion, die eine Zeichenfolge beliebiger Länge auf eine solche fester Länge abbildet, ↑kollisionsresistent

sein sollte und immer eine Einwegfunktion (↑unumkehrbar) ist.

kryptographischer Hash ↑Prüfsumme, die kryptographischen (↑Kryptographie) Gesichtspunkten entspricht bzw. sich Methoden der Kryptographie bedient und über eine ↑kryptographische Hashfunktion berechnet wird.

Mandantenfähigkeit Fähigkeit eines Systems, mehrere Nutzer(gruppen) gleichzeitig so zu bedienen, dass sie sich nicht wechselseitig beeinflussen und weder Zugriff auf die Daten der jeweils anderen Nutzer haben noch deren Aktivität beobachten können.

Medienbruch Übertragung von Daten bzw. Informationen von einem auf ein anderes Daten- bzw. Informationsmedium. Beispiele sind das manuelle Abtippen von Informationen, aber auch der Wechsel von Programmen oder Rechnern. Medienbrüche verlangsamen und verschlechtern Datenverarbeitungsprozesse und können zu Übertragungsfehlern führen. Selbst wenn Daten bzw. Informationen zwischen Prozessen in digitalen Formaten ausgetauscht werden können, ist ein Informationsverlust nur dann gewährleistet, wenn das Datenmodell des empfangenden Prozesses mindestens gleich mächtig ist wie das des sendenen Prozesses und es eine (eindeutige) Abbildung des einen auf das andere Datenmodell gibt.

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.

monolithisch aus einem Stück bestehend; zusammenhängend und fugenlos

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 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 Darstellung der Eigenschaften eines Fachgebiets und ihre Beziehungen zueinander, indem eine Reihe von Konzepten und Kategorien definiert wird, die das Fachgebiet repräsentieren.

Open–Closed-Prinzip Offenheit für Erweiterungen bei gleichzeitiger Abgeschlossenheit gegenüber (inkompatiblen) Abänderungen

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

Prüfsumme *checksum*, *hash*, in der Informationstechnik ein Wert, der aus den Ausgangsdaten berechnet wurde und in der Lage ist, mindestens einen Bitfehler in den Daten zu erkennen. Ein einfaches Beispiel für eine Prüfsumme ist die Quersumme. Prüfsummen werden in Fehlerkorrekturmechanismen verwendet und lassen sich dazu verwenden, zufällige Veränderungen an Daten zu erkennen. Einfache Prüfsummen bieten aber keinerlei Schutz gegenüber absichtlichen Veränderungen. Dazu bedarf es ↑kryptographischer Hashes.

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

Schema formales Modell der Struktur von Daten bzw. Informationen

Schlüssel-Wert-Paar Kombination einer benannten Variable und ihres zugewiesenen Wertes. Wird häufig in Datenstrukturen abgelegt, die dann über den Schlüssel einen Zugriff auf den damit assoziierten Wert erlauben. Grundlegender Baustein der ↑Schemata von ↑Metadaten.

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.

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

Systemarchitektur Summe der während der Entwicklung eines Systems getroffenen undin 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.

UID *unique identifier*, dt. eindeutige Kennung, (in einem gegebenen Kontext) eindeutiger Verweis auf eine beliebige Ressource. Vgl. ↑PID

Unix-Philosophie griffige Formuierung dreier Prinzipien, die wesentlich zum Erfolg des Unix-Betriebssystems und seiner Nachfolger (u.a. Linux, macOS, Android, iOS) beigetragen haben. Nach Salus [12, 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.

unumkehrbar im Kontext ↑kryptographischer Hashes die Tatsache, dass sich aus der Prüfsumme (*hash*) nicht die ursprüngliche Zeichenkette wiederherstellen lässt.

unvermeidliche Komplexität *essential complexity*, nach Fred Brooks [13] jener Teil der ↑Komplexität eines Systems, der in der Komplexität der Fragestellung begründet ist und der sich nicht verkleinern lässt. Eine gute Systemarchitektur zielt auf die Beherrschung dieser unvermeidlichen Komplexität u.a. durch Einsatz von ↑Abstraktion und ↑Modularisierung. Vgl. ↑vermeidbare Komplexität.

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

vermeidbare Komplexität *accidential complexity*, nach Fred Brooks [13] jener Teil der ↑Komplexität eines Systems, der *nicht* in der Komplexität der Fragestellung begründet ist und der sich durch geschickten Einsatz von (etablierten) Strategien beheben lässt. Ein wesentlicher Baustein zur Verringerung dieser vermeidbaren Komplexität ist die Verwendung guter ↑Abstraktionen. Vgl. ↑unvermeidbare Komplexität.

Versionsverwaltungssystem *version control system*, VCS; Software zur Verwaltung unterschiedlicher Versionen von Dateien und Programmen, die den Zugriff auf beliebige ältere als Versionen (↑Revision) gespeicherte Zustände ermöglicht. Gleichzeitig ein wichtiges Werkzeug für die Softwareentwicklung und wesentlicher Aspekt einer Projektinfrastruktur.

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. [14, 15].

World Wide Web WWW, „weltweites Netz", ursprünglich Anfang der 1990er Jahre am ↑CERN von Tim Berners-Lee [5] entwickeltes über das Internet aufrufbares System elektronischer ↑Hypertext-Dokumente (Webseiten), die durch Querverweise (Links) untereinander verbunden sind.

WWW ↑World Wide Web

YAGNI „*You ain't gonna need it*", (nicht nur) in der angelsächsischen Programmierwelt verbreite-

tes Akronym und wichtige Regel für die Programmierung. Zielt darauf ab, jeweils nur das zu implementieren, was im Augenblick wichtig ist oder von dem zweifellos klar ist, dass es in Kürze gebraucht wird. Versuch, durch einen pragmatischen Ansatz ein „zu viel“ an ↑Abstraktion zu vermeiden. Das zugrundeliegende Problem, das damit angegangen werden soll: Prognosen (hier: bzgl. der zukünftigen Anforderungen an eine bestimmte Software) sind schwierig, insbesondere wenn sie die Zukunft betreffen (vgl. ↑Kristallkugel).

Literatur

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

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

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

[4] T. H. Nelson. „Complex information processing: a file structure for the complex, the changing and the indeterminate“. In: *Proceedings of the 1965 20th National Conference*. ACM '65. Cleveland, Ohio, USA: Association for Computing Machinery, 1965, S. 84–100. ISBN: 9781450374958. DOI: 10.1145/800197.806036. URL: https://doi.org/10.1145/800197.806036.

[5] Tim Berners-Lee. *Weaving the Web : the original design an ultimate destiny of the World Wide Web by its inventor*. New York: HarperSanFrancisco, 1999.

[6] Edsger W. Dijkstra. The humble programmer. *Communications of the ACM* 15 (1972), S. 859–865.

[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] Peter H. Salus. *A Quarter Century of UNIX*. Reading, MA: Addison-Wesley, 1994.

[13] Frederick P. Brooks. *The Mythical Man Month*. Anniversary edition with four new chapters. Boston: Addison Wesley Longman, 1995.

[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.