Physikalische Chemie, Universität Rostock

Vorlesung: Forschungsdatenmanagement
im Sommersemester 2025

Dr. habil. Till Biskup

— Glossar zu Vorlesung 18: „Metadaten während der Datenerhebung“ —

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.

assoziatives Datenfeld *map*, *dictionary* oder *associative array*, Datenstruktur, die – anders als ein gewöhnliches Feld (*array*) bzw. eine Liste (*list*) – nichtnumerische (oder nicht fortlaufende) Schlüssel (zumeist Zeichenketten) verwendet, um die enthaltenen Elemente zu adressieren. Die Elemente sind (meist) in keiner festgelegten Reihenfolge abgespeichert. Idealerweise werden die Schlüssel so gewählt, dass eine für die Programmierer nachvollziehbare Verbindung zwischen Schlüssel und Datenwert besteht (↑semantisches Verständnis).

Auszeichnungssprache *markup language* maschinenlesbare Sprache für die Gliederung und Formatierung von Texten und anderen Daten. Der bekannteste Vertreter ist die *Hypertext Markup Language* (HTML), die Kernsprache des World Wide Webs.

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.

elektronisches Laborbuch ELN, digitale Variante des klassischen Laborbuchs. Mitunter wird sehr viel mehr unter diesem Begriff zusammengefasst, bis hin zu ↑Repositorium für Rohdaten, ↑Katalog, Projektplanungswerkzeugen, Proben- und Geräteverwaltung, was aber mehr Probleme aufwirft als hilft. Für eine Diskussion vgl. [1]. Vgl. ↑Laborbuch

ELN ↑elektronisches Laborbuch

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].

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.

Heuristik die Kunst, mit begrenztem Wissen und wenig Zeit dennoch zu wahrscheinlichen Aussagen oder praktikablen Lösungen zu kommen; analytisches Vorgehen, bei dem mit begrenztem Wissen über ein System mithilfe von mutmaßenden Schlussfolgerungen Aussagen über das System getroffen werden. Die Aussagen können von der optimalen Lösung abweichen. Ist eine optimale Lösung bekannt, lässt sich durch Vergleich die Güte der Heuristik bestimmen.

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.

intellektuelle Beherrschbarkeit *intellectual manageability*, nach Edsger Dijkstra [5] 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.

JSON *JavaScript Object Notation*, hierarchisches Datenformat, das ursprünglich zur einfachen Persistierung (↑Persistenz) von JavaScript-Objekten entwickelt wurde. Heute erfreut es sich als Austauschformat für Objekte und Datenstrukturen großer Beliebtheit, wird aber gerade für die Ablage von Konfigurationen in menschenlesbaren (und schreibbaren) Dateien aufgrund dessen noch einfacherer und übersichtlicherer Syntax zunehmend von ↑YAML abgelöst.

Katalog Werkzeug zum Auffinden und Erschließen von Forschungsdaten. ↑Forschungsdaten können mit Hilfe eines Datenkatalogs gesucht, gefunden und erschlossen werden. 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

Laborbuch auch: Laborjournal, primäres Dokumentationswerkzeug in den empirischen Wissenschaften, meist gebundenes Buch mit durchnummerierten Seiten. Typische Inhalte sind Ideen, Planungen, Notizen während der Durchführung von Experimenten und zuweilen Ergebnisse. Für Details inkl. evtl. rechtlicher Aspekte vgl. u.a. [6, 7]. Vgl. ↑elektronisches Laborbuch

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 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.

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.

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

Parser (engl. *to parse*, analysieren) Programm, das für die Zerlegung und Umwandlung einer Eingabe in ein für die Weiterverarbeitung geeigneteres Format zuständig ist.

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.

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

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 [8]. 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.

Semantik 1. Bedeutung bzw. Inhalt eines Wortes, Satzes oder Textes; 2. Teilgebiet der Linguistik, dessen Untersuchungsobjekt die Bedeutung sprachlicher Zeichen und Zeichenfolgen ist.

semantische Information Bedeutungsebene einer Information (vgl. ↑Semantik).

semantisches Verständnis Verständnis der ↑semantischen Information, also der Bedeutung bzw. des Inhaltes. Im Kontext von Auswertungsroutinen kann durch aussagekräftige Namensgebung der Schlüssel (Felder) eines ↑assoziativen Datenfeldes, in dem die ↑Metadaten abgelegt sind, eine entsprechende Ausdrucksstärke des Quellcodes erreicht werden. Außerdem „weiß“ die Auswertungsroutine dann immer, was sich in einem gewissen Feld verbirgt (z.B. Größe und Einheit für die Achsenbeschriftung).

Serialisierung in der Informatik eine Abbildung von strukturierten Daten auf eine sequenzielle Darstellungsform. Serialisierung wird hauptsächlich für die ↑Persistierung von Objekten in Dateien und für die Übertragung von Objekten über das Netzwerk bei verteilten Softwaresystemen verwendet.

Systemarchitektur Summe der während der Entwicklung eines Systems getroffenen undin der Umsetzung manifestierten Entscheidungen. Nach [9] 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.

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 [10] als essentiell für die Wissenschaftlichkeit hervorgehoben.

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

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

Version eindeutiger Zustand einer Software oder eines Dokuments. Zur ↑Nachvollziehbarkeit bedarf es einer strukturierten Versionierung mit Hilfe einer ↑Versionsverwaltung und zum Verweis auf eine Version typischerweise einer ↑Versionsnummer.

Versionsnummer hier: eindeutige Bezeichnung einer ↑Version einer Software oder einer Prozessbeschreibung, deren Kenntnis es erlaubt, auf genau diese Version der Software/Prozessbeschreibung Bezug zu nehmen.

Versionsverwaltung *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. [12, 13].

XML *eXtensible Markup Language*, „erweiterbare Auszeichnungssprache“, ↑Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten im Format einer Textdatei, die sowohl von Menschen als auch von Maschinen lesbar ist. Der große Vorteil von XML ist seine syntaktische Validierbarkeit, der Nachteil das Verhältnis von beschreibender Syntax zu eigentlichem Inhalt, das sowohl die Dateigröße erheblich erhöhen kann als auch die Lesbarkeit durch Menschen einschränkt. Alternativen, die von Menschen einfacher les- und insbesondere schreibbar sind, sind ↑JSON und ↑YAML.

YAML *YAML Ain't Markup Language* (rekursives Akronym, ursprünglich *Yet Another Markup Language*), vereinfachte ↑Auszeichnungssprache (*markup language*) zur Datenserialisierung (↑Serialisierung). Die grundsätzliche Annahme von YAML ist, dass sich jede beliebige Datenstruktur nur mit assoziativen Listen, Listen (Arrays) und Einzelwerten (Skalaren) darstellen lässt. Durch dieses einfache Konzept ist YAML wesentlich leichter von Menschen zu lesen und zu schreiben als beispielsweise ↑XML, außerdem vereinfacht es die Weiterverarbeitung der Daten, da die meisten Sprachen solche Konstrukte bereits integriert haben. Durch weitgehenden Verzicht auf Klammern ist YAML für Menschen

les- und insbesondere schreibbarer als ↑JSON und eignet sich daher gut für die Ablage von Metadaten und Konfigurationen. Seit Version YAML 1.2 ist ↑JSON eine vollständige Untermenge von YAML.

Literatur

[1] Mirjam Schröder und Till Biskup. LabInform ELN: A lightweight and flexible electronic laboratory notebook for academic research based on the open-source software DokuWiki. *ChemRxiv* (2023). DOI: 10.26434/chemrxiv-2023-2tvct.

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

[6] Howard M. Kanare. *Writing the Laboratory Notebook*. Washington, D.C.: American Chemical Society, 1985.

[7] Hans F. Ebel, Claus Bliefert und William E. Russey. *The Art of Scientific Writing*. Weinheim: Wiley-VCH, 2004.

[8] Melvin E. Conway. How do committees invent? *Datamation* 14.4 (1968), S. 28–31.

[9] Robert C. Martin. *Clean Architecture. A Craftman's Guide to Software Structure and Design*. Boston: Prentice Hall, 2018.

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

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

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