Buch: Softwareentwicklung für die Naturwissenschaften

Dr. habil. Till Biskup

— Glossar zu Kapitel 30: „Datenaufnahme: Metadaten“ —

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. Englische Begriffe werden zwar nach Möglichkeit übersetzt, erscheinen aber ggf. unter ihrem englischen Namen in der Liste. Verweise untereinander sind durch ↑ gekennzeichnet.

Abhängigkeit *dependency*, im Quellcode durch explizite Nennung hervorgerufene ↑Kopplung von Programmteilen (↑Funktionen, ↑Objekte, …), die dazu führt, dass der aufgerufene Programmteil nicht mehr ohne Veränderung des aufrufenden Teils verändert werden kann.

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.

Abstraktionsebene Summe aller ↑Abstraktionen eines bestimmten Abstraktionsgrades.

Abwärtskompatibilität Kompatibilität einer Version einer Software mit einer früheren Version ihrer selbst. Von großer Bedeutung für die Nutzer einer solchen Software. Die Trennung zwischen öffentlichen und internen Schnittstellen ↑API ist hier von entscheidender Bedeutung. Nur erstere werden für die Abwärtskompatibilität berücksichtigt.

ANSI *American National Standards Institute*, private, gemeinnützige, amerikanische Organisation zur Koordinierung der Entwicklung freiwilliger Normen in den U.S.A.

API *application programming interface*, Programmierschnittstelle oder genauer ↑Schnittstelle zur Anwendungsprogrammierung

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).

Attribut im Kontext der ↑objektorientierten Programmierung eine Variable, die innerhalb einer ↑Klasse definiert wird. ↑Methoden operieren auf den Attributen einer ↑Klasse bzw. dem daraus erzeugten ↑Objekt.

Auszeichnungssprache *markup language* (ML) 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.

Clean Code „sauberer Code“, letztlich lesbarer Code, der insbesondere im Kontext der naturwissenschaftlichen Datenauswertung die essentiellen Kriterien von Wiederverwendbarkeit, Zuverlässigkeit und Überprüfbarkeit erfüllt.

CSV *comma-separated values*, ↑Datenformat für zeilenweise Speicherung zusammengehöriger Daten, ähnlich der Zeilen in einer Tabelle.

Die einzelnen Felder in einer Zeile werden oft durch Komma (daher der Name), ggf. aber auch durch Semikolon getrennt. Im Gegensatz zu ↑DSV wurde CSV nie standardisiert, weshalb es Tabellenkalkulationen großer Hersteller gibt, die zwar CSV exportieren, aber den eigenen Export nicht mehr importieren können. Insbesondere der Umgang mit Feldtrennern innerhalb eines Feldes ist nicht festgelegt.

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. Wichtige Vertreter binärer Datenformate in den Naturwissenschaften sind ↑HDF5 und ↑IEEE 754 (eigentlich ein Standard für die Darstellung von Gleitkommazahlen). Wichtige Vertreter von Textformaten sind ↑CSV, ↑DSV, ↑JSON, ↑Windows-INI, ↑XML und ↑YAML.

Datensatz Einheit aus (numerischen) Daten und Informationen über die Daten (↑Metadaten).

Dependency Inversion Umkehr der ↑Abhängigkeiten gegenüber der intuitiven Implementierung. Abhängigkeiten sollten häufig entgegen dem ↑Kontrollfluss verlaufen.

Dependency-Inversion-Prinzip (DIP) Anwendung der ↑Dependency Inversion: Abstraktionen sollten nicht von Details abhängen. Umgekehrt sollten Details auf Abstraktionen aufbauen.

DSV *delimiter-separated values*, ↑Datenformat für zeilenweise Speicherung zusammengehöriger Daten, ähnlich der Zeilen in einer Tabelle. Im Gegensatz zu ↑CSV ist das Format eindeutig definiert. Das Trennzeichen (*delimiter*) ist beliebig wählbar. Sollte es innerhalb eines Feldes auftauchen, wird es durch Backslash („\“) geschützt.

Funktion im Kontext der strukturierten Programmierung eine Liste von Anweisungen, die eine bestimmte Aufgabe erfüllt und der Programmiersprache unter einem festen Namen bekannt ist. Vgl. ↑Methode.

größeres Projekt hier: Alles, was mehr als zwei Wochen Arbeit kostet und deutlich mehr als zweihundert Zeilen (reinen) Quellcode bzw. mehr als eine Handvoll Unterfunktionen umfasst. Wichtig ist der Fokus: Sobald ein Programm über längere Zeit und/oder von anderen verwendet werden soll (was eher die Regel statt die Ausnahme ist), ist es ein größeres Projekt.

HDF5 *Hierarchical Data Format*, ↑Datenformat, das insbesondere in wissenschaftlichen Anwendungen für die Speicherung großer Datenmengen verwendet wird. Optimierte Strukturen und Algorithmen erlauben das effiziente Speichern und Auslesen von ein- und mehrdimensionalen Tabellen, ohne dass jeweils der komplette Inhalt der Datei in den Speicher geladen werden muss. Tabellen und andere Daten können in ein und derselben Datei in einer beliebigen Verzeichnisstruktur abgelegt werden. Das ermöglicht u.a. die gleichzeitige Speicherung von Messwerten und zugehörigen ↑Metadaten. Das Format wurde vom *National Center for Supercomputing Applications* (NCSA) entwickelt und wird u.a. von der NASA für Missionen verwendet.

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 mit Hilfe 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.

HTML *Hypertext Markup Language*, textbasierte ↑Auszeichnungssprache zur Strukturierung elektronischer Dokumente wie Texte mit Hyperlinks, Bildern und anderen Inhalten. HTML-Dokumente sind die Grundlage des

World Wide Web und werden von Webbrowsern dargestellt. Neben den vom Browser angezeigten Inhalten können HTML-Dateien zusätzliche Angaben in Form von Metainformationen enthalten, z. B. über die im Text verwendeten Sprachen, den Autor oder den zusammengefassten Inhalt des Textes.

IEEE *Institute of Electrical and Electronics Engineers*, weltweiter Berufsverband von Ingenieuren hauptsächlich aus den Bereichen Elektrotechnik und Informationstechnik. Der Verband veranstaltet Fachtagungen, gibt diverse Fachzeitschriften heraus und bildet Gremien für die Standardisierung von Techniken, Hardware und Software.

IEEE 754 Norm des ↑IEEE, die die Standarddarstellungen für binäre Gleitkommazahlen in Computern definiert und genaue Verfahren für die Durchführung mathematischer Operationen, insbesondere für Rundungen, festlegt. In der Fassung ↑ANSI/IEEE 754-1985 ist nur der Standard für 64 Bit eindeutig.

Importroutine Softwareeinheit (oft eine ↑Funktion oder ↑Klasse), die ein ↑Datenformat einlesen und in eine für das jeweilige Programm intern sinnvolle Repräsentation umwandeln kann. Die Repräsentation von Daten innerhalb einer Anwendung sollte immer unabhängig vom jeweiligen externen Datenformat sein (eine Anwendung des ↑Dependency-Inversion-Prinzips).

Infrastruktur Personelle, sachliche und finanzielle Ausstattung, um ein angestrebtes Ziel zu erreichen. Im Kontext der Softwareentwicklung die Gesamtheit der Hilfsmittel, die (manche) Abläufe formalisieren und für Struktur und Überprüfbarkeit sorgen. Erleichtert die Arbeit des Programmierers, indem sie viele Aspekte festlegt, die so zur Routine werden (und keine Denkleistung absorbieren).

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.

Kapselung *encapsulation*, ein ↑Objekt enthält Daten (↑Attribute) *und* zugehöriges Verhalten (↑Methoden) und kann beides nach Belieben vor anderen Objekten verstecken.

Klasse *class*, im Kontext der ↑objektorientierten Programmierung die Blaupause für die Erzeugung eines ↑Objektes; Definition der Daten (↑Attribute) und des zugehörigen Verhaltens (↑Methoden).

Kontrollfluss *flow of control*, Reihenfolge des Aufrufs von Programmteilen (↑Funktionen, ↑Objekte, ...), um eine gegebene Aufgabe zu erfüllen.

Kopplung *coupling*, in Software der Grad der Verbindung zweier Komponenten; enge Bindung mehrerer Einheiten einer Software aneinander, so dass sie nicht unabhängig wiederverwendbar (bzw. ggf. auch nicht testbar) sind. Programmierkonzepte zielen generell auf eine lose Kopplung (*loose coupling*) einzelner Komponenten ab, da so die Wiederverwendbarkeit erleichtert wird.

Mehrfachvererbung *multiple inheritance*, eine ↑Klasse erbt (↑Vererbung) von mehr als einer ↑Superklasse. Wird von den wenigsten Programmiersprachen unterstützt, oftmals behilft man sich hier aber des Konzeptes einer ↑Schnittstelle (*interface*) (3.) und kann dann mehr als eine solche implementieren (bzw. davon erben). Konzeptionell lassen sich diese beiden Ansätze quasi identisch einsetzen.

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

Methode im Kontext der ↑objektorientierten Programmierung eine ↑Funktion, die innerhalb einer ↑Klasse definiert wird und auf den

↑Attributen einer ↑Klasse bzw. dem daraus erzeugten ↑Objekt operiert.

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.

monolithisch aus einem Stück bestehend; zusammenhängend und fugenlos

Objekt *object*, im Kontext der ↑objektorientierten Programmierung der grundlegende Baustein eines Programms, bestehend aus den Daten (↑Attribute) und dem zugehörigen Verhalten (↑Methoden). Ein Objekt ist in diesem Kontext immer die Instanz einer ↑Klasse.

objektorientierte Programmierung (OOP) ein ↑Programmierparadigma, bei dem Daten (Variablen zugewiesene Werte, als ↑Attribute bezeichnet) und Funktionen (↑Methoden), die auf diesen Daten (Attributen) operieren, eine Einheit bilden. Die in den ↑Attributen gespeicherten Daten lassen sich i.d.R. nur vermittelt durch (öffentlich zugängliche) ↑Methoden der ↑Klasse bzw. des daraus erzeugten ↑Objektes ansprechen. Es gibt eine klare Trennung zwischen öffentlicher ↑Schnittstelle und internen Verarbeitungsroutinen. Wichtige Vertreter objektorientierter Programmiersprachen sind Smalltalk, C++ und Java, aber auch Python.

Open Closed Offenheit einer Software-Einheit für Erweiterungen bei gleichzeitiger Abgeschlossenheit gegenüber Abänderung

Open-Closed-Prinzip Anwendung von ↑Open Closed: Software-Einheiten (↑Klassen, ↑Module, ↑Funktionen etc.) sollten offen für Erweiterung, aber verschlossen gegenüber Abänderung sein.

Paradigma nach Thomas S. Kuhn [2] ein Satz allgemein anerkannter wissenschaftlicher Leistungen, der für eine gewisse Zeit einer Gemeinschaft von Fachleuten maßgebende Probleme und Lösungen liefert

parsbar von einem ↑Parser verarbeitbar, d.h. insbesondere, dass die zu parsende Eingabe in ein für die elektronische/digitale Weiterverarbeitung geeignetes Format umgewandelt werden kann.

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 ↑Objekte) oder logische Verbindungen über lange Zeit (insbesondere über einen Programmabbruch hinaus) bereitzuhalten; benötigt ein nichtflüchtiges Speichermedium.

Polymorphie *polymorphism*, „Vielgestaltigkeit“, ähnliche ↑Objekte können auf die gleiche Botschaft (den Aufruf einer gleichnamigen ↑Methode) in unterschiedlicher Weise reagieren.

Prinzip der geringsten Überraschung *principle of least surprise*, Regel für die Programmentwicklung, soweit möglich auf in einem gegebenen Kontext etablierte Regeln und Standards zurückzugreifen. Baustein für möglichst intuitive Bedienbarkeit.

Programmierparadigma ein ↑Paradigma der Art zu programmieren. Wichtige Beispiele sind strukturierte Programmierung, ↑objektorientierte Programmierung und funktionale Programmierung.

reproduzierbare Wissenschaft *reproducible science*, seit der Etablierung rechnergestützter Datenauswertung eigentlich nie mehr erreichter, aber für die Wissenschaft konstituierender Aspekt, dass sich Ergebnisse und Auswertungen unabhängig reproduzieren lassen, weil alle dazu notwendigen Aspekte vollständig und ausreichend beschrieben wurden. Motivation für die Vorlesung, deren Ziel es ist, die Hörer mit Konzepten vertraut zu machen, die letztlich eine ernstzunehmende reproduzierbare Wissenschaft ermöglichen.

Schnittstelle *interface*, Begriff mit mehreren leicht unterschiedlichen Bedeutungen; (1.) ↑Signatur einer ↑Funktion oder ↑Methode. (2.)

Im weiteren Sinne die Gesamtheit der öffentlichen ↑Attribute und ↑Methoden einer ↑Klasse bzw. eines ↑Objekts. Der Nutzer kennt nur die Schnittstelle, die Implementierung ist irrelevant und kann sich problemlos jederzeit ändern, solange die Funktionalität erhalten bleibt. Das dient der Trennung von Verantwortlichkeiten und ermöglicht ↑Modularisierung und ist in der Folge ein wesentlicher Aspekt der ↑Softwarearchitektur. (3.) In einer weiteren Bedeutung wird der Begriff (auch im Deutschen dann häufig mit seinem englischen Pendant) für (abstrakte) Klassen verwendet, die lediglich eine Schnittstelle (im Sinne von 2.) definieren. Das ist hauptsächlich dann von Bedeutung, wenn die Programmiersprache keine ↑Mehrfachvererbung unterstützt, aber das Implementieren von „*Interfaces*“.

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.

Signatur hier: Name und Parameter einer ↑Funktion bzw. ↑Methode, also alles, was ein Nutzer braucht, um diese Funktion oder Methode verwenden zu können.

Softwarearchitektur Aufteilung eines größeren Projektes in einzelne kleinere Projekte bzw. Aufgaben (↑Modularisierung), Definition klarer ↑Schnittstellen und Anforderungen sowie der Interaktion der einzelnen Teile miteinander. Nach Robert C. Martin die Gestalt eines Systems, die ihm von seinen Entwicklern gegeben wird: Unterteilung des Systems in Komponenten, ihre Anordnung, und die Art ihrer Interaktion miteinander. [3, S. 136]

Subklasse ↑Klasse, die von einer anderen Klasse (der ↑Superklasse) ↑Attribute und ↑Methoden erbt. Die ↑Vererbung geht dabei i.d.R. über die nach außen hin sichtbare ↑Schnittstelle der Superklasse hinaus. Die Subklasse erbt von der ↑Superklasse häufig nur den „kleinsten gemeinsamen Nenner“ und implementiert die spezifische Funktionalität.

Superklasse ↑Klasse, von der andere Klassen (↑Subklassen) ↑Attribute und ↑Methoden erben. Die ↑Vererbung geht dabei i.d.R. über die nach außen hin sichtbare ↑Schnittstelle der Superklasse hinaus. Superklassen implementieren bzw. definieren normalerweise nur das Notwendigste, sozusagen den „kleinsten gemeinsamen Nenner“. Alle spezifische Funktionalität wird in der ↑Subklasse implementiert.

System zur Datenverarbeitung hier: Gesamtsystem für wissenschaftliche Datenverarbeitung von der Datenaufnahme bis zur fertigen Publikation, das alle Aspekte umfasst und das ↑reproduzierbare 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 (↑Clean Code) und eine solide ↑Softwarearchitektur voraus.

Typisierung *typing*, Zuweisung eines Typs zu einem Objekt (im abstrakten Sinne) einer Pro-

grammiersprache, z.B. Ganzzahl (*integer*) oder Zeichenkette (*string*) im Fall einer Variable. ↑Abstraktion, die die Ausdrucksstärke von Programmiersprachen und Programmen deutlich erhöht, und die Überprüfung der Korrektheit erleichtert sowie Optimierungen ermöglicht. Typisierung kann explizit und implizit erfolgen. Darüber hinaus wird zwischen starker und schwacher Typisierung sowie zwischen statischer und dynamischer Typisierung unterschieden. Jede Art der Typisierung hat ihre Vor- und Nachteile, und unterschiedliche Programmiersprachen verwenden unterschiedliche Arten der Typisierung.

Vererbung *inheritance*, Weitergabe aller Eigenschaften (↑Attribute, ↑Methoden) von einer ↑Superklasse an eine ↑Subklasse. Die Subklasse ist vom gleichen Typ (↑Typisierung) wie die Superklasse, was wiederum die Grundlage der ↑Polymorphie ist. Änderungen der Superklasse wirken sich allerdings auf die Subklasse aus, die ↑Kapselung wird entsprechend geschwächt.

Windows-INI blockweise strukturiertes ↑Datenformat, das ursprünglich für die Ablage von Konfigurationsoptionen für das Windows-Betriebssystem und darauf laufende Programme entwickelt wurde. Die blockweise Struktur erlaubt zwei Hierarchieebenen: Blöcke und Schlüssel-Wert-Paare. Für weitere Hierarchieebenen und die Ablage beliebig verschachtelter Datenstrukturen müssen hierarchische Datenformate wie ↑XML, ↑JSON oder ↑YAML herangezogen werden.

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.

Literatur

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

[2] Thomas S. Kuhn. *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt am Main: Suhrkamp, 1976.

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