TEIL A
HERAUSFORDERUNG DATENERSCHLIEßUNG

Teil A dient zur Einführung in die Thematik der Erschließung multidimensionaler heterogener Datenräume anhand ausgewählter Konzepte und Technologien. Ausgehend von einigen Vorklärungen (Kap. 1) werden in Kap. 2 Herausforderungen der Datenintegration vorgestellt. Neben dem Konzept des Data Warehousing wird hier auch auf die auf Analyse ausgerichtete Modellierung multidimensionaler Datenräume durch das Konzept des Online Analytical Processing (OLAP) sowie auf alternative Ansätze zur Datenintegration eingegangen. Kap. 3 behandelt die computergestützte Suche nach zuvor unbekannten, verwertbaren Mustern in Daten über Data Mining; Kap. 4 befasst sich mit dem Gebiet der Visualisierung von Daten und geht auch auf das Untergebiet der Informationsvisualisierung ein. Kap. 5 schließlich behandelt einige Entwicklungslinien von Internet und World Wide Web und stellt das Konzept des Grid-Computing vor.

1 Vorklärungen

In diesem Kapitel werden einige zentrale Begriffe eingeführt, die für das Verständnis der weiteren Arbeit hilfreich sind. In Kap. 1.1 wird der Begriff Daten konkretisiert und im Sinne der Informationswissenschaft gegen die Begriffe Wissen und Information abgegrenzt; ferner werden die Begriffe Metadaten und Datenraum eingeführt sowie Größenordnungen von Daten behandelt. In Kap. 1.2 erfolgt eine Klassifizierung von Daten nach ihrer Struktur sowie die Einführung der Begriffe Attribut, Datensatz und Werteausprägung. Kap. 1.3 wirft ein Schlaglicht auf in Statistik und Visualisierung vorgenommene Unterscheidungen von Daten anhand von Attributeigenschaften. Anmerkungen zum Begriff der Multidimensionalität schließen sich an (Kap. 1.4); Kap. 1.5 führt abschließend den Begriff der heterogenen Daten ein.

1.1 Zum Daten-Begriff

1.1.1 Abgrenzung von Wissen, Information und Daten

Zunächst soll der Begriff *Daten* gegen die Begriffe *Wissen* sowie *Information* abgegrenzt werden. Wersig verweist auf die Entwertung, die der Begriff Information durch „*seinen unterschiedlichen und inflationären Gebrauch*" erfahren hat [Wersig 1996, 9] und konstatiert, dass mit diesem Wort „*so etwas wie ein Mythos der Postmoderne verknüpft*" sei, den „*keiner gerne wissenschaftlich stringent angeht*" [Wersig 1993, 151f.]. Diese Kritik an einer unklaren Abgrenzung des Begriffes Information gegen Begriffe wie Daten oder Wissen ist nach wie vor aktuell; so sehen etwa Schumann und Müller die primäre Ursache für die uneinheitlich vorgenommene Abgrenzung der Gebiete Datenvisualisierung und Informationsvisualisierung (vgl. Kap. 4) im Fehlen einer „*allgemein anerkannten und vollständigen Definition des Informationsbegriffs*" [Schumann, Müller 2000, 341]. Ebenfalls im Kontext der Informationsvisualisierung betont Spence: „*It is important to make a clear distinction between data and information. The 'information explosion' so widely discussed is actually a data explosion: it is the derivation of information (or understanding, or insight) from the data that is difficult* […]" [Spence 2001, 4]. Ein aktuelles Beispiel für eine Vermischung der Begriffe Daten und Information sind die Studien der School of Information Management and Systems an der University of California, Berkeley [UC SIMS 2003ab] [UC SIMS 2000], in denen das weltweite Datenaufkommen abgeschätzt wird. Hier werden Datenvolumen und Information begrifflich gleichgesetzt: „*How much new information is created each year?*"

Die Informationswissenschaft hat die Begriffe Wissen, Information und Daten klar gegeneinander abgegrenzt; diese Arbeit folgt ihrer Aufteilung.

- **Wissen**

Hennings definiert *Wissen* als „*vorhandene Bestände an Modellen über konkrete und abstrakte Objekte, Ereignisse und Sachverhalte* [...]*, die partiell in einem Individuum (repräsentiert in seinem Gedächtnis), in einer gesellschaftlichen Gruppe, aber auch in einer Organisation, einem ganzen Kulturkreis oder in der Menschheit insgesamt vorhanden sind*" [Hennings 1991, 5]. Nach Wersig kann Wissen bezeichnet werden als „*Zunächst individuell in der Zeit (biografisch) erworbene und gespeicherte Modellierungen von Welt und Selbst*", die „*später dann gesellschaftlich verallgemeinert*" werden [Wersig 2000, 14].

- **Information**

In Abgrenzung zu Wissen ist *Information* nach Wersig dasjenige „*Wissen, das für konkretes zielgerichtetes Handeln in der Welt benötigt wird, unter Berücksichtigung der Konditionen des Handelns (Zeitpunkt, erwarteter Handlungsgewinn etc.)*" [Wersig 2000, 14]. Hennings betont den Aspekt der Problemlösung durch Untermengen von Wissen, die hierzu in Information umgewandelt werden: Information ist „*eine Teilmenge von Wissen, die von be-*

stimmten Personen, Gruppen, Organisationen etc. in konkreten Situationen zur Durchführung von Handlungen, z.B. dem Lösen von Problemen, benötigt wird" [Hennings 1991, 6]. Er bringt den Zusammenhang der Begriffe Wissen, Problem und Information auf die Formel: „*Wissen* + *Problem* = *Information*" [Hennings 1991, 7].

- **Daten**

Daten können zunächst mit Wersig allgemein als „*Repräsentationen der Welt, die von einem Repräsentationssystem in ein anderes transformiert werden können und letztlich der Sinnenwelt von Menschen zugänglich gemacht werden müssen*" definiert werden [Wersig 2000, 14]. Im hier relevanten Sinn können Daten eingeschränkt werden auf alles, was sich auf einer Datenverarbeitungsanlage „*geeignet kodiert erfassen, speichern, bearbeiten, übertragen und wieder ausgeben läßt*"[2] [Hennings 1991, 5]. Damit können Daten „*auch die auf einer Datenverarbeitungsanlage repräsentierbaren Elemente von Wissen und / oder Informationen darstellen*" [Hennings 1991, 6, Fußnote 4].

1.1.2 Daten und Metadaten

Eine bestimmte Klasse von Daten, die in dieser Arbeit eine nicht unwesentliche Rolle spielen wird, wird als *Metadaten* bezeichnet. Es handelt sich hierbei um solche Daten, die notwendige Informationen über andere Daten bereitstellen; Metadaten werden deshalb auch als *Daten über Daten* bezeichnet. Metadaten beschreiben bspw. Inhalt und Qualität anderer Daten und dokumentieren ihre Verfügbarkeit, Zugriffsrechte etc. Sie bilden in vielen Anwendungsgebieten eine wichtige Voraussetzung für die Nutzung komplexer Datenbestände; Beispiele sind Bibliothekskataloge oder Daten zur Dokumentation der komplexen Datenressourcen von Firmen, wissenschaftlichen Einrichtungen etc. Metadaten sollen helfen, bspw. folgende Fragen zu beantworten (vgl. [ANZLIC 2001]):

- Um welche Arten von Daten handelt es sich?
- Wie wurden die Daten erzeugt?
- Wann wurden die Daten erzeugt?
- Wer hat die Daten erzeugt?
- Wie kann auf die Daten zugegriffen werden?

Je nach Anwendungskontext können Metadaten dabei zum Zugriff auf andere Daten verwendet werden oder als ihrerseits zu verarbeitende Daten auftreten: „*What is metadata to one application may be data to another*" [Jeffery 2003]. Ferner besitzt der Begriff im Bereich der Datenbankmanagementsysteme (DBMS) eine spezielle Bedeutung (*Datenbankmetadaten*) und bezeichnet diejenigen Daten, die Auskunft über den Aufbau einzelner Datenbank*strukturen* geben, etwa bei relationalen Datenbankmanagementsystemen (RDBMS) über Namen von Tabellen, ihre Zusammensetzung, verwendete Datentypen etc.

1.1.3 Datenräume

Der Begriff *Datenraum* wird eingeführt, um beliebige Datensammlungen auf abstraktem Niveau benennen zu können. Er soll im ersten Teil dieser Arbeit zunächst ganz allgemein für eine - bewusst nicht näher konkretisierte - Sammlung von Daten stehen; je nach Fokus des Interesses kann ein Datenraum also unterschiedliche Mengen von Daten enthalten oder sich auf verschiedene thematische Gebiete beziehen. Bei der Analyse der zu erschließenden Datenbestände des Potsdam-Institutes für Klimafolgenforschung (vgl. Kap. 6ff.) wird dieser Begriff auf spezifische Metadaten- und Datensammlungen aus dem Kontext der Erdsystemanalyse eingegrenzt.

[2] Schreibweise im Original.

1.1.4 Größenordnungen

Um die durch die Speicherung digitaler Daten anfallenden Volumina beschreiben zu können, werden in immer schneller Folge neue Größenordnungen der zugrundeliegenden Maßeinheit Byte notwendig. Zur Orientierung des Lesers sei hier ein kurzer Überblick über heute gebräuchliche Bezeichnungen gegeben (vgl. Tab. 1.1).

Bezeichnung	Bedeutung
Bit	erlaubt die Unterscheidung zweier Zustände (0 und 1)
Byte	8 Bit
Kilobyte	2^{10} Byte ≈ ca. 10^3 Byte (1024 Byte)
Megabyte	2^{20} Byte ≈ ca. 10^6 Byte (1024 Kilobyte)
Gigabyte	2^{30} Byte ≈ ca. 10^9 Byte (1024 Megabyte)
Terabyte	2^{40} Byte ≈ ca. 10^{12} Byte (1024 Gigabyte)
Petabyte	2^{50} Byte ≈ ca. 10^{15} Byte (1024 Terabyte)
Exabyte	2^{60} Byte ≈ ca. 10^{18} Byte (1024 Petabyte)

Tab. 1.1 - Maßeinheiten für Datenvolumen[3].

Ein Petabyte beispielsweise entspricht also 1024∗1024∗1024∗1024∗1024 Byte. Die Größenordnungen werden deutlich, wenn man sich vergegenwärtigt, *was* mit welchem Datenvolumen kodiert werden kann – so kann der Text von Goethes Faust II in 300 Kilobyte abgespeichert werden, das Hefegenom[4] benötigt rund 40 mal mehr, das Humangenom etwa 9.000 mal mehr Speicherplatz [GSF 2002]. Die in Data Warehouses (vgl. Kap. 2.2) zusammengeführten Datenbestände großer Unternehmen erreichen Volumen im Terabyte-Bereich; das Datenvolumen der statischen HTML-Seiten im World Wide Web wird für 2002 auf 167 Terabyte, das der aus Datenbankinhalten auf Anfrage dynamisch erzeugten HTML-Seiten auf 91.850 Terabyte geschätzt [UC SIMS 2003ab]. Der neue Teilchenbeschleuniger des CERN (Large Hadron Collider, LHC)[5], der 2007 in Betrieb gehen soll, wird hingegen alleine jährlich Daten in einer Größenordnung von 12 bis 14 Petabyte erzeugen (vgl. Kap. 5.3.4).

1.2 Klassifizierung nach Struktur und Auswertbarkeit

Daten im hier relevanten Sinn können anhand des Grades, in dem ihre jeweilige *Struktur* für automatisch ausführbare Operationen auf ihnen ausgenutzt werden kann, weiter in die drei Klassen strukturierte Daten, semistrukturierte Daten sowie unstrukturierte Daten unterschieden werden.

1.2.1 Strukturierte Daten

Zur Klasse der *strukturierten Daten* (vgl. [Dittrich, Domenig 1999] [Busse et al. 1999]) werden hier solche Daten gerechnet, die einem vorab definierten, rigidem und explizitem *Schema* gehorchen müssen. Das Schema legt fest, welche Datenelemente verwendet werden dürfen und welcher Datentyp jedem Datenelement zugrunde liegt; Datentypen können bspw. Zeichenketten (Strings), numerische Werte oder Datumswerte sein[6]. Das Schema strukturierter Daten ist ein rigides Schema in dem Sinne, dass alle derart abge-

[3] In Anlehnung an [UC SIMS 2003a].
[4] Das Gebiet der Genforschung ist ein aktuelles Beispiel für ein beständiges Anwachsen von Daten. So enthielten nach [Rötzer 2001] die Datenbanken des European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/) im Jahr 2001 bereits über 12,5 Milliarden Basenpaare; dem Autor zufolge verdoppelt sich der Umfang der Daten - mit steigender Geschwindigkeit - alle 8 Monate. Der Medizinnobelpreisträger Sidney Brenner kritisiert die Datenjagd in der Genforschung, die zu einer irrelevanten Datenflut führe, und fordert eine verstärkte Hinwendung zur Theoriebildung [Die Zeit 2002].
[5] http://lhc-new-homepage.web.cern.ch/lhc-new-homepage/
[6] Dies sind einige der typischen Datentypen in relationalen Datenbankmanagementsystemen.

speicherten Daten ihm folgen müssen – sie dürfen für die einzelnen Datenelemente nur Werte entsprechend der vorgegebenen Datentypen enthalten. Änderungen des Schemas sind zwar möglich, finden jedoch eher selten statt. Ferner handelt es sich bei dem Schema strukturierter Daten um ein explizites Schema in dem Sinne, dass dieses getrennt von den Daten gespeichert wird und abgefragt werden kann, so dass es als Grundlage für Operationen auf den Daten dienen kann.

Jedes einzelne Datenelement mit zugehörigem Datentyp wird hier als *Attribut* bezeichnet; ein Attribut besitzt innerhalb eines Schemas einen eindeutigen Namen und einen Datentyp. Eine Untermenge strukturierter Daten, die durch eine Kombination unterschiedlicher Attribute definiert ist, wird hier als *Datensatz* bezeichnet. Jeder Datensatz besteht aus jeweils einem Wert für jedes vorgegebene Attribut, ein konkreter Wert wird hier als *Werteausprägung* des Datensatzes für dieses Attribut bezeichnet. Typische Beispiele für strukturierte Daten sind Daten, die in Datenbankmanagementsystemen (DBMS) gehalten werden.

Strukturierte Daten bilden diejenige Klasse von Daten, deren Struktur am weitesten für eine Auswertung herangezogen werden kann. Solche Daten erlauben beispielsweise eine präzise Suche durch strukturierte Anfragen, die auf der durch die Attribute vorgegebenen Struktur und dem zugrundeliegenden Typsystem basieren.

(a) **Schema:**

Attributname	Datentyp
Autor	String
Titel	String
Verlag	String
Jahr	Number

(b) **Datensätze:**

Autor	Titel	Verlag	Jahr
Müller	Rosenzucht	Prentice Hall	1999
Meyer	Backen heute	Springer	1977
Müller	Tulpenzucht	MIT Press	2003

Abb. 1.1 - Strukturierte Daten am Beispiel einer einfachen Datenbanktabelle: (a) vorab definiertes, rigides, explizites Schema; (b) entsprechend abgespeicherte Datensätze.

Dies soll an einem einfachen Beispiel verdeutlicht werden: In einem relationalen Datenbankmanagementsystem (RDBMS) sei eine Tabelle zur Speicherung von Literaturangaben definiert. Das gewählte Schema (vgl. Abb. 1.1a) legt fest, dass diese Tabelle durch genau vier Attribute definiert ist; drei der Attribute (Autor, Titel, Verlag) besitzen jeweils den Datentyp String, das vierte Attribut (Jahr, das Erscheinungsjahr eines Buches) hingegen einen numerischen Datentyp. Jede Literaturangabe, die in diese Tabelle eingefügt werden soll, ist ein Datensatz, der dem rigiden Schema folgen muss (vgl. Abb. 1.1b). Das RDBMS stellt auf Anfrage Datenbankmetadaten über den Aufbau dieser Tabelle, die Namen ihrer Attribute und die jeweils zulässigen Datentypen zur Verfügung – das Schema ist explizit. Datensätze können nun basierend auf den Namen der Attribute und Wertevorgaben entsprechend des Typsystems ausgewählt werden: *Wähle alle Datensätze aus, in den Autor = „Müller" und Jahr > 2000.*

1.2.2 Semistrukturierte Daten

Als *semistrukturierte Daten* (vgl. [Dittrich, Domenig 1999] [Busse et al. 1999]) werden hier solche Daten bezeichnet, die zwar ebenfalls Struktur besitzen, sich jedoch von strukturierten Daten in wesentlichen Punkten unterscheiden. Zunächst ist ihre Struktur nicht rigide, also nicht durch ein striktes Schema vordefiniert; ferner kann sie implizit sein, so dass jedes Datenelement seine eigene semantische Definition bspw. in Form eines Labels mit sich führt. Damit kann zu einem gegebenen Zeitpunkt die Summe aller Label einer semistrukturierten Datenquelle als ihr Schema angesehen werden; dieses Schema kann sich jedoch potentiell jedes Mal ändern, wenn neue Daten hinzugefügt werden. Ein Beispiel für semistrukturierte Daten sind HTML-Seiten[7] – hier werden über unterschiedliche Label

[7] Mit dem Aufkommen des World Wide Web ist das Interesse an der Erforschung semistrukturierter

(Tags) einzelne Bereiche bspw. als Überschriften unterschiedlicher Ebenen, als Absatz etc. gekennzeichnet (vgl. Abb. 1.2). Diese Struktur ist dabei nicht fest vorgeschrieben, sie ist also nicht rigide; und zugleich ist sie implizit. Sie kann dennoch für eine automatische Auswertung herangezogen werden; so kann etwa eine Suchmaschine HTML-Seiten für eine Suchanfrage unterschiedlich einstufen, je nachdem, ob ein Suchbegriff in einer Überschrift oder in einem Absatz vorkommt[8]. Die Suchmaschine Google berücksichtigt für die Erstellung einer Treffer-Rangfolge unter anderem auch, ob ein Suchbegriff im Titel einer HTML-Seite, einem Hyperlink, einem Textabschnitt mit großer oder kleiner Schrift etc. gefunden wird [Brin, Page 1998].

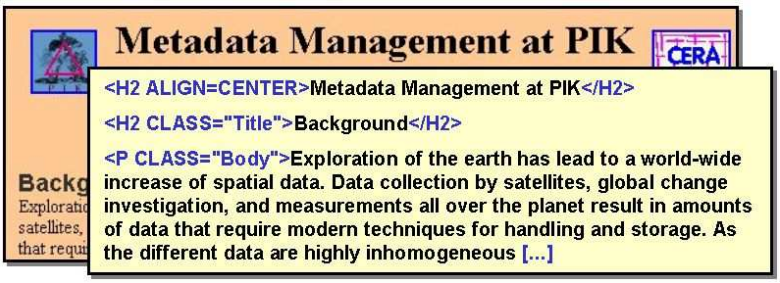

Abb. 1.2 - Beispiel für semistrukturierte Daten: HTML-Seite.

1.2.3 Unstrukturierte Daten

Das andere Ende der Skala bildet die Klasse der *unstrukturierten Daten* (vgl. [Dittrich, Domenig 1999] [Busse et al. 1999]). Darunter werden hier solche Daten verstanden, die nicht weiter in Datenelemente mit jeweils spezifischen Datentypen untergliedert, sondern nur in Zeichenketten oder Bytesequenzen zerlegt werden können. Beispiele für solche unstrukturierten Daten sind Text, Video und Audio. Unstrukturierte Daten besitzen keine Struktur in dem Sinne, dass diese wie bei strukturierten Daten direkt für Anfragen ausgenutzt werden kann. Um unstrukturierte Daten auswerten zu können, muss entsprechend auf andere Wege zurückgegriffen werden.

So können auf unstrukturierten Daten bspw. sog. unscharfe Suchen durch unstrukturierte Anfragen durchgeführt werden, d.h. durch Anfragen, die aus Zeichen- oder Bytesequenzen sowie Operationen auf diesen (wie and, or, near) bestehen [Dittrich, Domenig 1999]. Der klassische Ansatz zur automatischen Bewertung von Ähnlichkeiten zwischen Texten stammt aus dem Kontext des Information Retrieval und basiert auf sog. Vektorraum-Analyse [Salton et al. 1995]. Dabei werden sowohl gespeicherte Texte wie Anfragen gegen diese jeweils als hochdimensionale Vektoren von Worten, Wortstämmen oder Phrasen repräsentiert, die nachfolgend ausgewertet werden. Dies kann u.a. dazu herangezogen werden, um die Treffergenauigkeit von Anfragen zu erhöhen, Themen in Dokumenten zu identifizieren oder durch sog. selektive Texttraversierung nur solche Passagen aus Dokumenten zu extrahieren, die als relevant beurteilt werden. Ein weiteres Beispiel für die Auswertung unstrukturierter Daten ist das Gebiet der Computer Vision (vgl. [Fischler, Firschein 1987] [Jähne 1993]), bei dem ein automatisches Erkennen und Verstehen von Bildern angestrebt wird. Hier können bspw. durch Auswertung der Bilddaten zunächst Umrisse (Kanten) identifiziert werden, anhand derer dann auf dargestellte Objekte gefolgert wird.

Daten gestiegen (vgl. die Arbeiten von [Abiteboul 1997] sowie [Buneman 1997]).

[8] Zur automatischen Auswertung von HTML-Seiten vgl. auch die diesbezüglichen Untergebiete des Data Mining (Kap. 3.4.2, Web Content Mining sowie Kap. 3.4.3, Web Structure Mining).

1.2.4 Zuordnung zu Speicherformen

Typischerweise handelt es sich bei strukturierten Daten um Daten in Datenbankmanagementsystemen (DBMS), während semi- oder unstrukturierte Daten in Dateien vorliegen. Busse et al. verweisen allerdings darauf, dass hier keine klare Grenze zwischen datenbankbasierten und nicht-datenbankbasierten Systemen existiert. So erlauben einige DBMS die Speicherung bspw. von Bild- oder Multimediadaten in Form sog. Binary Large Objects (BLOBs), deren interne Struktur dem DBMS explizit nicht zur Auswertung zur Verfügung steht. Auf der anderen Seite wiederum können den Autoren zufolge Sammlungen wohlstrukturierter Dateien, für die ein striktes Format durchgesetzt und eine deklarative Sprache zum Zugriff bereitgestellt wird, ähnlich wie ein DBMS behandelt werden [Busse et al. 1999].

1.3 Klassifizierung anhand von Attributeigenschaften

Den Eigenschaften der Werte in einzelnen Attributen kommt für Verfahren, die diese zu einer automatischen Auswertung bzw. zur Transformation in geeignete visuelle Repräsentationen heranziehen, eine zentrale Bedeutung zu. Hier sollen kurz typische Aufteilungen aus den Bereichen Statistik sowie Visualisierung vorgestellt werden.

1.3.1 Unterscheidungen in der Statistik

Gegenstand der *Statistik* ist „[...] *die Entwicklung und Anwendung formaler Methoden zur Gewinnung, Beschreibung und Analyse sowie zur Beurteilung quantitativer Beobachtungen (Daten)* [...]" [Vogel 1995, 3]. Sie kann die Bereiche der *deskriptiven Statistik* (eine reine Beschreibung von Daten bspw. anhand von Häufigkeiten oder Verteilungen) sowie der *analytischen Statistik* (Ableitung allgemeingültiger Aussagen aus Daten) unterschieden werden. Die vielfältigen statistischen Verfahren sollen hier nicht näher ausgeführt werden (für eine anschauliche Einführung vgl. [Zöfel 2001]). An dieser Stelle soll kurz in Anlehnung an [Zöfel 2001] die in der Statistik verwendete Klassifizierung von Attributen[9] skizziert werden.

Diese Klassifizierung berücksichtigt die Art der statistischen Auswertungen, die sinnvoll auf den jeweiligen Werteausprägungen einzelner Attribute durchgeführt werden können. So werden textuelle Werteausprägungen zur Auswertung zunächst in eine numerische Form überführt und anhand entsprechender Zahlenwerte repräsentiert – etwa durch Abbildung der drei Kategorien *Nichtraucher*, *mäßiger Raucher* und *starker Raucher* auf die Zahlenwerte 1, 2 und 3. Die Eignung einzelner statistischer Verfahren für konkrete Attribute hängt dabei von der empirischen Bedeutung, die ihrer Zahlenrepräsentation zugemessen werden darf. Um einzelne Attribute diesbezüglich zu unterscheiden, werden sie einem von vier unterschiedlichen sog. Skalenniveaus (nominalskaliert, ordinalskaliert, intervallskaliert oder verhältnisskaliert) zugeordnet (vgl. Tab. 1.2).

Skalenniveau	empirische Relevanz
Nominal	Keine
Ordinal	Ordnung der Zahlen
Intervall	Differenzen der Zahlen
Verhältnis	Verhältnisse der Zahlen

Tab. 1.2 - Die empirische Relevanz unterschiedlicher Skalenniveaus[10].

- **Nominalskalierung**

Als *nominalskaliert* werden solche Attribute bezeichnet, deren Wertevorrat, auch wenn er über Zahlen repräsentiert ist, keinerlei sinnvolle Auswertung der Zahlenwerte - auch nicht

[9] Zöfel verwendet den Begriff *Variable* anstelle von *Attribut.*
[10] Tabelle übernommen aus [Zöfel 2001, Tab. 2.1].

bezüglich der Rangfolge - zulässt. Ein typisches Beispiel ist das Attribut *Familienstand* mit den möglichen Ausprägungen *ledig*, *verheiratet*, *verwitwet* und *geschieden* – es ist irrelevant, welche Zahlenwerte zur Codierung der Ausprägungen verwendet werden; ebenso lässt sich keine sinnvolle Anordnung der Ausprägungen vornehmen. Als *dichotom* bezeichnet man solche nominalskalierten Attribute, die genau zwei Kategorien besitzen (etwa das Attribut *Geschlecht* mit den möglichen Ausprägungen *männlich* oder *weiblich*). Dichotome Attribute stellen insofern einen Spezialfall dar, als sie - anders als nominalskalierte Variablen mit mehr als zwei Kategorien - eine Ordnungsrelation beinhalten. Nominalskalierte Attribute sind in den Möglichkeiten ihrer statistischen Auswertung in der Regel auf Häufigkeitsauszählungen eingeschränkt; so ist etwa (zumindest bei nicht-dichotomen) nominalskalierten Attributen die Berechnung eines Mittelwertes nicht sinnvoll.

- **Ordinalskalierung**

Als *ordinalskaliert* werden Attribute bezeichnet, deren Repräsentation über Zahlenwerte zwar eine auswertbare Ordnungsrelation enthält, aber keine Auswertung der Abstände zwischen den Zahlenwerten zulässt. Die Repräsentation der drei Kategorien *Nichtraucher*, *mäßiger Raucher* und *starker Raucher* über die Zahlenwerte 1, 2 und 3 ist ein Beispiel für das Vorliegen einer Ordinalskalierung: Zwar kann eine gültige Rangfolge aus den Zahlenwerten abgeleitet werden (Kategorie 2 raucht mehr als Kategorie 1 und weniger als Kategorie 3); die Differenz zwischen den Zahlenwerten hingegen kann nicht als empirisch relevant angesehen und entsprechend nicht zur Auswertung herangezogen werden. So beträgt der Abstand zwischen den Zahlenrepräsentationen für *Nichtraucher* und *mäßiger Raucher* einerseits und zwischen *mäßiger Raucher* und *starker Raucher* andererseits jeweils den Wert 1, dennoch ist der reale Unterschied zwischen diesen Ausprägungen nicht unbedingt als gleich anzusehen.

- **Intervallskalierung**

Intervallskalierte Attribute liegen vor, wenn auch den Differenzen zwischen den einzelnen Zahlenwerten eine auswertbare empirische Relevanz zukommt. Attribute wie Gewicht, Körpergröße, Einkommen, Temperatur etc. erlauben auswertbare Aussagen über den Abstand einzelner Ausprägungen zueinander – so beträgt der Abstand zwischen den Temperaturwerten 15°C und 17°C ebenso wie der Abstand zwischen den Temperaturwerten 33°C und 35°C jeweils 2°C. Die statistische Auswertung intervallskalierter Attribute unterliegt nach [Zöfel 2001] keiner Einschränkung.

- **Verhältnisskalierung**

Als *verhältnisskaliert* werden schließlich solche intervallskalierten Attribute bezeichnet, die den Zahlenwert 0 annehmen können und für die 0 zugleich auch den kleinsten möglichen Wert darstellt. Verhältnisskalierte Variablen erlauben Aussagen über das Verhältnis von Zahlenwerten zueinander. So können Aussagen wie „*Das Verhältnis eines Einkommens von 2000 Euro zu einem Einkommen von 1000 Euro ist dasselbe wie das eines Einkommens von 1000 Euro zu 500 Euro (nämlich jeweils doppelt so groß)*" getroffen werden, nicht aber bspw. für Temperaturen in °Celsius, die auch negative Werte zulassen, oder den Intelligenzquotienten, der nicht den Wert 0 annehmen kann.

1.3.2 Unterscheidungen in der Visualisierung

Auch bei der Erzeugung visueller Repräsentationen von Daten kommt deren Eigenschaften eine zentrale Bedeutung zu – so soll ein aus vorliegenden Daten generiertes Bild nicht etwa dadurch zu falschen Schlussfolgerungen verleiten, indem es bspw. eine Ordnungsrelation suggeriert, die in den Daten überhaupt nicht vorhanden ist (vgl. dazu Kap. 4, Visualisierung). In diesem Bereich werden unterschiedliche Begriffe verwendet, die im Prinzip auf eine Klassifizierung in numerische Werte (mit impliziter Ordnungsrelation), nichtnumerische

Werte mit Ordnungsrelation und nichtnumerische Werte ohne Ordnungsrelation hinauslaufen. So unterscheidet etwa [Spence 2001, 4f.] zwischen

- *numerical data*,
- *ordinal data* (Vorliegen einer natürlichen Ordnung, Beispiel: die Wochentage) und
- *categorical data* (es existiert keine Ordnung, Beispiel: Unterscheidung von Tieren in Pferd, Zebra, Antilope).

[Schumann, Müller 2000, 29ff.] unterteilen zu visualisierende Daten zunächst in solche, die den Raum beschreiben, in dem die Daten erhoben wurden (Beobachtungsraum), und solche, die die dort beobachteten Werte beschreiben (Merkmale)[11]. Für die Daten des Beobachtungsraumes wird dabei vorausgesetzt, dass es sich stets um metrische Daten handelt, die kontinuierlich oder diskret sein können. Die einzelnen Merkmale werden nach der Skalierung ihres Wertebereichs zunächst unterschieden in

- *quantitative Merkmale* (verwenden metrische Skalen, die diskret oder kontinuierlich sein können), sowie
- *qualitative Merkmale* (verwenden nichtmetrische Skalen).

Qualitative Merkmale werden dann weiter unterschieden in

- *ordinale Merkmale* (auf ihrer Skala ist eine Ordnungsrelation definiert), sowie
- *nominale Merkmale* (auf ihrer Skala ist keine Ordnungsrelation definiert).

Die Autoren verweisen darauf, dass diese Begriffe nicht immer einheitlich verwendet werden und dass bspw. auch eine Unterscheidung zwischen nominalen, diskret skalierten ordinalen und kontinuierlich skalierten ordinalen Daten vorgeschlagen wurde [Schumann, Müller 2000, 37, Fußnote 2].

1.4 Multidimensionalität

Da die einzelnen Attribute eines Datensatzes die unterschiedlichen Dimensionen darstellen, anhand derer ein Datensatz charakterisiert werden kann, werden sie hier auch als seine *Dimensionen* bezeichnet. Die Zahl der Dimensionen hat direkte Auswirkungen auf die Herausforderungen, die bei der Erschließung der in einem Datenraum enthaltenen Informationen auftreten: Je mehr Dimensionen ein Datenraum enthält, desto schwieriger ist es, alle in ihm repräsentierten Informationen und ihre Beziehungen zueinander zu erfassen.

Herausforderungen durch Datenräume mit einer hohen Zahl von Dimensionen können je nach Anwendungskontext und eingesetzter Technologie auf verschiedenen Ebenen entstehen. So sind Daten, die in relationalen Datenbankmanagementsystemen gehalten werden, meist für eine Betrachtung unter spezifischen Blickwinkeln modelliert; flexiblere Zugänge zur Auswertung von Datenräumen mit vielen Dimensionen werden in diesem Zusammenhang mit dem Konzept des Online Analytical Processing (OLAP, vgl. Kap. 2.3) angestrebt. Ferner erschwert eine hohe Zahl von Dimensionen eine manuelle Auswertung von Datenräumen [Fayyad et al. 1996c]; in diesem Kontext sind Entwicklungen aus dem Bereich des Data Mining zu sehen (vgl. Kap. 3). Bei der Generierung von graphischen Repräsentationen aus Daten wiederum stellt eine hohe Zahl von Dimensionen eine grundsätzliche Herausforderung dar (vgl. Kap. 4): „*With up to three rows* [im Sinne von Attributen bzw. Dimensionen, MW]*, a data table can be constructed directly as a single image* [...] *However, an image has only three dimensions. And this barriere is impassable*" [Bertin 1977/1981].

Unter einem *multidimensionalen Datenraum* (von lat. multus, viel) wird hier ein Datenraum verstanden, dessen Anzahl an Dimensionen groß genug ist, um spezifische Herausforderungen an seine Auswertung zu implizieren. Im Kontext der Visualisierung von Daten sind

[11] Vgl. dazu ausführlicher Kap. 4.4

hingegen feinere Unterscheidungen sinnvoll, die bei der Behandlung dieses Themengebietes eingeführt werden (vgl. Kap. 4.4).

1.5 Heterogenität

Heterogen bedeutet - im Gegensatz zu *homogen* - andersartig (von gr. hetero..., anders..., verschieden...); als *heterogene Daten* sollen hier solche Daten bezeichnet werden, die aufgrund von Verschiedenartigkeiten nicht ohne vorherige Prozesse, die diese Verschiedenartigkeit aufheben, sinnvoll verarbeitet oder ausgewertet werden können. Busse et al. stufen Heterogenität als eine natürliche Folge der Tatsache ein, dass autonome Entwicklungen von Systemen offenbar immer in *unterschiedlichen Lösungen* resultieren. Gründe hierfür sehen sie bspw. im unterschiedlichem Verständnis oder in unterschiedlicher Modellierung der gleichen Weltausschnitte, in der jeweils verfügbaren technischen Infrastruktur oder den gegebenen spezifischen Anforderungen [Busse et al. 1999]. Visser et al. verweisen auf den ambivalenten Aspekt von Heterogenität im Bereich der Informationstechnologie: So müsse Heterogenität bei EDV-Systemen nicht immer eine unwillkommene Eigenschaft sein; heterogene Systeme reflektieren eine starke Anlehnung an konkret zu lösende Aufgaben, die in engem Bezug zu gesteigerter Effizienz zu sehen ist [Visser et al. 1997]. Die Heterogenität von Daten hingegen wird übereinstimmend als wesentliches Hindernis bei der Integration von Datenressourcen aus unterschiedlichen Quellen eingestuft [Hull 1997] [Visser et al. 1997] [Busse et al. 1999].