

Biomedizinische Ontologie

Wissen strukturieren für den Informatik-Einsatz

herausgegeben von

Ludger Jansen und Barry Smith

mit Beiträgen von

Thomas Bittner, Berit Brogaard, Boris Hennig, Ludger Jansen,
Ingvar Johansson, Bert Klagges, Ulf Schwarz und Barry Smith

Kapitel 4

Klassifikationen

LUDGER JANSEN

Dinge zu klassifizieren ist seit alters her eine Standardtätigkeit für Wissenschaftler, und so verwundert es nicht, dass sich Philosophen seit zweieinhalb Jahrtausenden Gedanken über das Klassifizieren gemacht haben, angefangen bei Platon und Aristoteles bis hin zur modernen Wissenschaftstheorie. Um einige der Ergebnisse dieses Nachdenkens zusammenzufassen, soll zunächst eine Klassifikationsparodie diskutiert werden: die „chinesische“ Klassifikation der Tiere von Jorge Luis Borges.⁴⁸ Das, was an dieser Klassifikationsparodie komisch erscheint, zählt natürlich zu dem, was in einer idealen Klassifikation vermieden werden sollte (Kap. 4.1). Doch all jene Elemente, die an dieser Parodie komisch erscheinen, kommen auch in dezidiert wissenschaftlichen Klassifikationen vor (Kap. 4.2). Dass wird anhand von Beispielen aus der Terminologie-Datenbank des amerikanischen *National Cancer Institute* (NCI) gezeigt werden, dem *NCI Thesaurus*. Zum Teil sind diese Abweichungen vom klassifikatorischen Ideal abhängig von den technischen Randbedingungen einer solchen Datenbank (Kap. 4.3). Eine Arbeitsteilung zwischen mehreren Informationssystemen – zwischen einer auf Vollständigkeit angelegten Referenzontologie auf der einen und auf bestimmte Anwendungen zugeschnittenen Anwendungsontologien auf der anderen Seite – könnte einige der Probleme lösen (Kap. 4.4).

4.1 Chinesische Tiere: Wie man eine gute Klassifikation erstellt

In einer „gewissen chinesischen Enzyklopädie“, so erzählt Jorge Luis Borges, seien die Tiere in die folgenden Gruppen eingeteilt:

⁴⁸ Vgl. Borges 1966.

1. Tiere, die dem Kaiser gehören,
2. einbalsamierte Tiere,
3. gezähmte Tiere,
4. Milchschweine,
5. Sirenen,
6. Fabeltiere,
7. herrenlose Hunde,
8. in diese Gruppierung gehörige,
9. die sich wie Tolle gebärden,
10. unzählbare,
11. die mit einem ganz feinen Pinsel aus Kamelhaar gezeichnet sind,
12. und so weiter,
13. die den Wasserkrug zerbrochen haben,
14. die von weitem wie Fliegen aussehen.⁴⁹

Wir haben es hier mit einer ausgezeichneten Parodie von Klassifikationen zu tun. Um sich leichter auf sie beziehen zu können, soll diese erfundene Klassifikation im Folgenden als „CTT“, als „Chinesische Tier-Taxonomie“ bezeichnet werden. CTT wirkt dadurch komisch, dass sie einige Sachen grundlegend anders macht, als man es in einer wissenschaftlichen Klassifikation machen sollte. Gute Klassifikationen beruhen also im Gegenschuß darauf, genau diese Fehler zu vermeiden. Was ist es also, das uns in CTT lustig vorkommt? Was können wir von CTT lernen? Hier sind einige der Regeln für das Anfertigen von guten und nützlichen Klassifikationen, gegen die CTT verstößt:

- *Strukturiertheit.* Gute Taxonomien sind strukturiert. Sie berücksichtigen, dass Typen sich in Untertypen aufteilen. In CTT sind alle Gruppen gleichberechtigt. Nun sind Sirenen aber Fabeltiere, (5) mithin eine Unterart von (6). Gefragt ist also eine Hierarchie von Gattungen und Arten.
- *Disjunktivität.* Wenn wir eine solche Hierarchie von Typen und Subtypen haben, dann gilt: Alles, was einen Subtyp instantiiert, instantiiert auch den dazugehörenden Typ. In einer zoologischen Klassifikation (die CTT ja wä-

⁴⁹ Borges 1966, 212. – Als Quelle für die Taxonomie gibt Borges „Dr. Franz Kuhn“ an. Es gibt zwar einen deutschen Sinologen und Übersetzer zahlreicher chinesischer Texte dieses Namens, aber unter seinen Veröffentlichungen ist keine Quelle für CTT zu finden. Vermutlich hat Borges hier in parodistischer Absicht eine falsche Fährte gelegt. Vgl. dazu <http://www.linguistlist.org/issues/7/7-1446.html>. Meinen Zweifel an der Seriosität von Borges' Verweis auf Kuhn und den Hinweis auf diese Seite verdanke ich Jan Westerhoff.

re) gilt beispielsweise, dass alle Tiere, die Säugetiere sind, auch Wirbeltiere sind. Auf derselben Ebene der Klassifikation sollten die verschiedenen Typen jedoch disjunkt sein: Kein Ding sollte zu zwei Typen derselben Ebene gehören, wie ja auch kein Tier sowohl ein Säugetier als auch ein Reptil ist oder sowohl ein Wirbeltier als auch ein Wirbelloses. Wenn Typen derselben Ebene diese Forderung erfüllen, sind sie disjunkt. Die Typen in CTT erfüllen (zumindest nicht notwendigerweise) diese Forderung nicht: Typ (1) der Tiere, die dem Kaiser gehören, kann ja auch gezähmte Tiere umfassen, also Tiere, die zu Typ (3) gehören.

- *Exhaustivität.* Taxonomien enthalten idealerweise alle Entitäten desjenigen Bereiches, den sie zu klassifizieren vorgeben. CTT ist jedoch weit entfernt davon, alle Tiere zu umfassen, wenn davon abgesehen wird, dass beliebige Tiere unter Typ (8) oder auch unter den Typ (13) „und so weiter“ subsumiert werden können. Exhaustivität und Disjunktivität werden oft mit der Formulierung „gemeinsam exhaustiv und paarweise disjunkt“ in einem Atemzug genannt; im Englischen wird diese Forderung oft mit „JEPD“ abgekürzt (für: *jointly exhaustive, pairwise disjoint*).
- *Verzicht auf Ambiguität.* Gute Taxonomien spielen nicht mit Mehrdeutigkeiten ihres Oberbegriffs. Sind Fabeltiere, gemalte Tiere und tote Tiere im gleichen Sinne Tiere wie herrenlose Hunde? Wohl nicht. Die Typen (2), (5), (6) und (11) passen aus diesem Grund nicht in das Ordnungsschema hinein. Gemalte Tiere sind eben keine Tiere, sondern Bilder, die Tiere darstellen.
- *Einheitlichkeit.* Gute Taxonomien haben einen wohldefinierten Bereich, aus dem sie ihre Unterscheidungsmerkmale wählen. CTT hingegen beweist eine große Kreativität, Ordnungsmerkmale aus möglichst verschiedenen Bereichen zu verwenden: (1) sortiert nach dem Eigentümer, (4) nach der Artzugehörigkeit, (7) nach der Abwesenheit eines Eigentümers plus Artzugehörigkeit, (9) nach dem Verhalten, (13) nach den Auswirkungen des Verhaltens und (14) nach der Wirkung auf einen entfernten Beobachter.
- *Keine Meta-Typen.* Gute Taxonomien vermeiden „Meta-Typen“, die nur durch die Kategorisierung selbst zustande kommen, denn solche Meta-Typen können zu paradoxen Effekten führen. In CTT ist (8) eine solcher Meta-Typ: „Tiere, die zu CTT gehören“. Gehört ein Tier zu CTT? Dann gehört es zu (8). Gehören alle Tiere zu CTT? Dann gehören alle Tiere zu (8). Alle Tiere, die zu (1) bis (7) oder (9) bis (14) gehören, gehören also zu

(8). Gehört ein Tier zu CTT, aber nicht zu diesen Klassen, ist dies auch kein Problem: Es kann schließlich auch zu CTT gehören, wenn es nur zu (8) gehört. (8) ist also ein sehr merkwürdiger Typ von Tier. Klassifikationen, die (8) enthalten, führen zu Problemen, die strukturell dem semantischen Paradox des so genannten Wahr-Sagers entsprechen, das sich an einem Satz wie „Dieser Satz ist wahr“ entzündet. Dieser Satz ist hinsichtlich seines Wahrheitswertes unbestimmt, weil er beliebig bewertet werden kann. Wenn angenommen wird, er sei wahr, dann ist der Fall, was er behauptet, nämlich dass er wahr ist – und das ist genau das, was für die Wahrheit eines Satzes notwendig ist. Wird hingegen angenommen, er sei falsch, dann ist nicht der Fall, was er behauptet, wie dies für einen falschen Satz auch angemessen ist. Jeder der beiden Wahrheitswerte „wahr“ und „falsch“ kann dem Wahr-Sager also konsistent zugeordnet werden. Genauso konsistent können Tiere, die nicht zu den übrigen Typen von CTT gehören, (8) zugeordnet werden oder nicht. Eine solche Möglichkeit der willkürlichen Zuordnung ist freilich sehr merkwürdig.

Noch schlimmer wird es mit einer Russellschen Variante von CTT, nämlich CTT*, die statt des Typs (8) den Typ (8*) enthält:

8*. Tiere, die nicht zu CTT* gehören.

Ein solcher Typ führt zu Problemen, die der Russellschen Antinomie oder dem Lügner-Paradox strukturell entsprechen: Gehört ein Tier zu den Typen (1) bis (7) oder (9) bis (13), dann gehört es zu CTT* und damit nicht zu (8*). So weit ist die Sache klar. Gehört ein Tier aber nicht zu diesen Typen, geraten wir in eine paradoxe Situation. Denn wenn das Tier nun auch nicht zu (8*) gehören würde, würde es zu gar keinem Typ von CTT* gehören, und damit nicht zu CTT*. Tiere, die nicht zu CTT* gehören, gehören aber zu (8*). Wenn man davon ausgeht, dass das Tier nicht zu den übrigen Typen gehört, folgt also: Wenn etwas nicht zu (8*) gehört, gehört es zu (8*). Aber was auch immer zu (8*) gehört, gehört zu CTT*. Dann jedoch gehört das Tier nicht zu (8*). Wieder vorausgesetzt, dass das Tier nicht zu den übrigen Typen gehört, gilt also auch: Was zu (8*) gehört, gehört nicht zu (8*). Das Auftreten solcher Situationen sollte man in wissenschaftlichen Klassifikationen tunlichst vermeiden.

- *Explizitheit und Präzision.* Gute Taxonomien sind explizit und präzise. „Und so weiter“-Typen wie (12) genügen diesen Ansprüchen selbstverständlich nicht.

- *Ontologische Fundierung.* Gute Taxonomien klassifizieren Dinge aufgrund von entsprechenden Merkmalen. Das schließt Meta-Typen genauso aus wie Typ (12), „und so weiter“, denn zur „und so weiter“-Gruppe gehören Dinge nicht, weil sie ein sie einigendes Merkmal besitzen. Und auch Typ (14), „Die von weitem aussehen wie Fliegen“, klassifiziert nicht direkt aufgrund von Merkmalen der Dinge selbst, sondern aufgrund ihrer Wirkung auf einen entfernten Beobachter.

4.2 Medizindatenbanken:

Wie man eine schlechte Klassifikation erstellt

Wenden wir uns nun einer ernstgemeinten Klassifikation zu, dem *National Cancer Institute Thesaurus* (NCIT). Dieser wurde vom amerikanischen *National Cancer Institute* erstellt, um dessen Kampf gegen den Krebs in vielfacher Weise zu unterstützen: Zum einen stellt der Thesaurus ein online zugängliches Nachschlagewerk dar, zum anderen kann er als ein kontrolliertes Vokabular zur Annotation und Indexierung der für die Erforschung und die Bekämpfung von Krebs einschlägigen Literatur dienen.⁵⁰ Dazu enthält er mehr als 110.000 Ausdrücke und 36.000 Begriffe, die sich auf den Krebs beziehen und für die Krebsforschung wichtig sind. Darunter befinden sich 10.000 Arten medizinischer Diagnosen oder Störungen, mehr als 5.000 anatomische Arten, mehr als 3.500 Chemikalien und Medikamente, etwa 2.000 Arten von Genen etc.⁵¹ Der Thesaurus hat also einen durchaus beeindruckenden Umfangs. Doch seine Anlage erinnert allzu oft an CTT.

4.2.1 *Strukturiertheit: Gruppen und Tiere*

Anders als CTT besitzt der NCIT eine Hierarchie von Ober- und Untertypen. Doch ist diese Hierarchie oft nicht oder nicht richtig ausgearbeitet. Ein Beispiel für eine fehlende Hierarchisierung bietet der Eintrag *Subgroup*, der vom NCIT wie als eine „subdivision of a larger group with members often exhibiting similar characteristics“ definiert. Nun sollte man annehmen, dass solche Teile von Gruppen auch Gruppen sind, und in der Tat wird dies auch von der

⁵⁰ Vgl. Ceusters, Smith und Goldberg 2005.

⁵¹ Vgl. Fragoso 2004.

Definition von *Group* im NCIT impliziert, denn *Group* wird definiert als „any number of entities (members) considered as a unit“. Die Verbindung zwischen *Subgroup* und *Group* fehlt jedoch im NCIT, und damit fehlt ein Teil der hierarchischen Struktur.

Zudem ist der NCIT in vielen Fällen auch falsch strukturiert. Als Obertyp von *Group* wird beispielsweise *Grouping* angegeben, das definiert wird als ein „system for classifying things into groups or the activity of putting things together in groups“. Aber, wie die philosophische Tradition weiß,⁵² muss die Definition des Supertyps auch von den Subtypen ausgesagt werden können. Aus der Definition von *Grouping* und der Tatsache, dass *Group* als Untertyp von *Grouping* angesehen wird, folgt jedoch, dass eine *Group* ein „system for classifying things into groups or the activity of putting things together in groups“ ist – eine absurde Konklusion.

Die Klassifikation der Tiere im NCIT ist nahezu von ähnlicher Qualität wie die in Borges' CTT. Der Typ *Tier* hat im NCIT die folgenden Untertypen: *Wirbelloses*, *Labortier*, *Wechselwarmes*, *Wirbeltier*. Eine höchst merkwürdige Reihung fürwahr, ist doch erstens schon das Paar *Wirbelloses*/*Wirbeltier* eine vollständige Aufteilung aller Tiere. Zweitens nimmt sich der Typ *Labortier* neben drei natürlichen Klassen höchst seltsam aus, da Labortiere eben keine natürliche Art bilden; die Einteilung verwendet hier folglich Merkmale ganz verschiedener Art. Drittens schließlich ist *Wechselwarmes* ein Untertyp von *Wirbeltier*, gehört damit also gar nicht auf die gleiche Ebene wie sein Obertyp.

4.2.2 Disjunkтивität und Exhaustivität: Patienten

Häufig gibt es im NCIT Untertypen desselben Obertyps, die nicht disjunkt sind. Ein Beispiel: Als medizinische Terminologie hat der NCIT natürlich einen Eintrag *Patient*. Dieser Eintrag hat zwei Untertypen: *Krebspatient* und *Ambulanter Patient* (*Cancerpatient*, *Outpatient*). Diese beiden Einträge sind augenscheinlich nicht disjunkt, denn viele Krebspatienten werden ambulant behandelt. Und natürlich sind diese beiden Untertypen auch keine exhaustive Klassifikation der Patienten, da es zahlreiche Patienten gibt, die weder Krebspatienten noch ambulante Patienten sind.

⁵² Vgl. z.B. Aristoteles, *Kategorien* 3.

Normalerweise würde man dieses Beispiel als einen typischen Anwendungsfall einer Kreuzklassifikation ansehen: Wir haben es mit zwei voneinander unabhängigen Merkmalen zu tun, die jeweils vorliegen oder nicht vorliegen können. Miteinander kombiniert bilden diese beiden Merkmale insgesamt vier verschiedene Klassen, in unserem Fall vier Klassen von Patienten, die in Abbildung 4.1 zu sehen sind.

Abb. 4.1: Vier Klassen von Patienten – eine Kreuzklassifikation

PATIENTEN	Ambulant? Ja.	Ambulant? Nein.
Krebs? Ja.	<i>Ambulante Krebspatienten</i>	<i>Stationäre Krebspatienten</i>
Krebs? Nein.	<i>Ambulante Nichtkrebspatienten</i>	<i>Stationäre Nichtkrebspatienten</i>

In den Informationswissenschaften stellt man sich ideale Klassifikationen als Baumstrukturen vor. Das steht in der langen Tradition der *Arbor Porphyriana*, des Porphyrischen Baumes, benannt nach dem Neuplatoniker Porphyrios (ca. 234–304), dessen Einleitungsschrift zum Organon des Aristoteles, die *Isagogé*, die zentralen Begriffe des klassischen aristotelischen Begriffsbaumes erläutert, wie sie in Abbildung 4.2 dargestellt werden. Mit Hilfe solcher Bäume können gemäß der Formel *Definitio fit per genus proximum et differentiam specificam* leicht Definitionen nach Aristotelischem Muster erstellt werden: Eine Spezies wird danach so definiert, dass die ihr als nächstes übergeordnete Gattung (das *Genus proximum*) gemeinsam mit dem sie konstituierenden spezifischen Merkmal (der *Differentia specifica*) angegeben wird. Als Definition des Menschen könnte man gemäß dieser Struktur „vernunftbegabtes Lebewesen“ angeben, wobei „Lebewesen“ die übergeordnete Gattung und „vernunftbegabt“ das spezifische Merkmal des Menschen bezeichnet.

Solche Baumstrukturen nennt man in der Informatik auch „gerichtete Graphen“. In ihnen ist eine Richtung ausgezeichnet: Bäume haben einen eindeutigen „Stamm“, das *Genus ultimum*, von dem ausgehend immer feiner werdende Äste abzweigen, die schließlich in den „Blättern“, den Spezies, enden. Jedes Element in einem solchen Baum (jeder „Knoten“ des Graphen) hat einen eindeutigen Obertyp. Versucht man nun aber, eine Kreuzklassifikation, wie die in Abbildung 4.1 dargestellte, in einen solchen Graphen umzuwandeln, erlebt man zwei Überraschungen: Erstens geht die Eindeutigkeit des Ober-

typs verloren – ambulante Krebspatienten gehören sowohl zum Obertyp „Krebspatienten“ als auch zum Obertyp „Ambulante Patienten“. Die Verzweigungen gehen in einem solchen Diagramm nicht mehr nur in eine Richtung, und ein Element des Diagramms kann sowohl mehrere Untertypen als auch mehrere Obertypen haben. In diesen Fällen wird gesagt, dass diese Diagramme eine „multiple Vererbung“ erlauben (*multiple inheritance*) bzw. dass in ihnen *diamonds* vorkommen dürfen, also rautenförmige Strukturen wie in Abbildung 4.3.

Abb. 4.2: Schema eines Porphyrischen Baumes

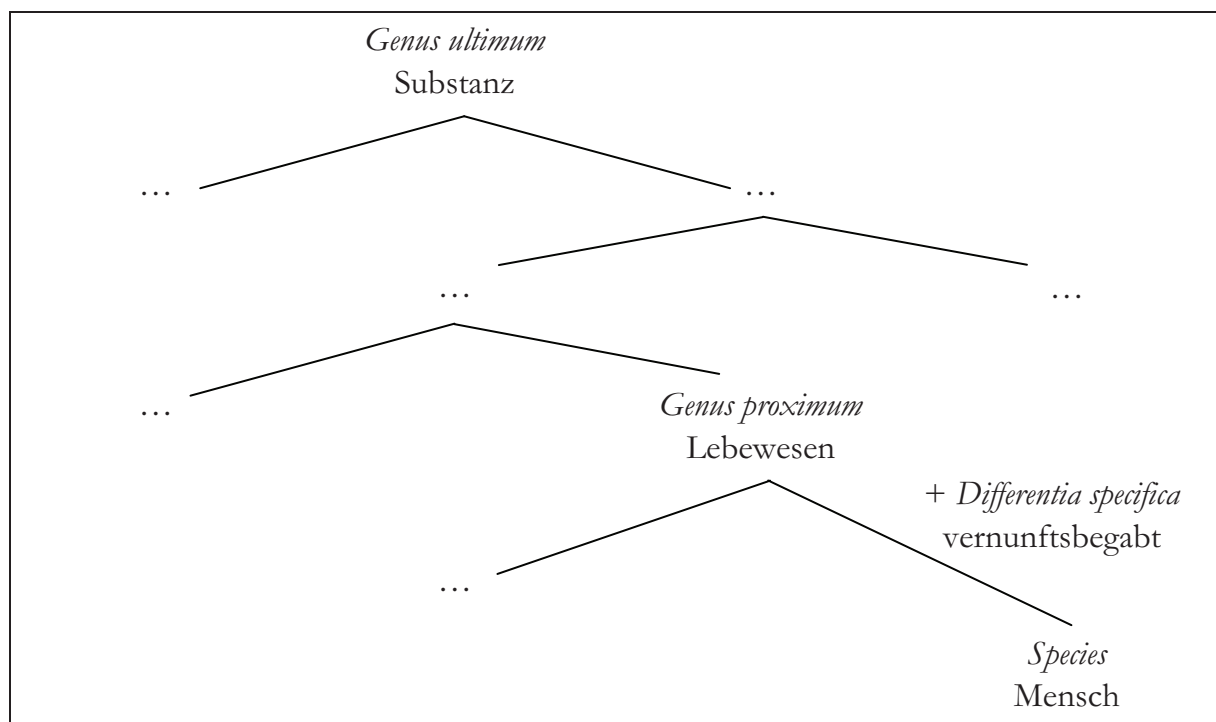
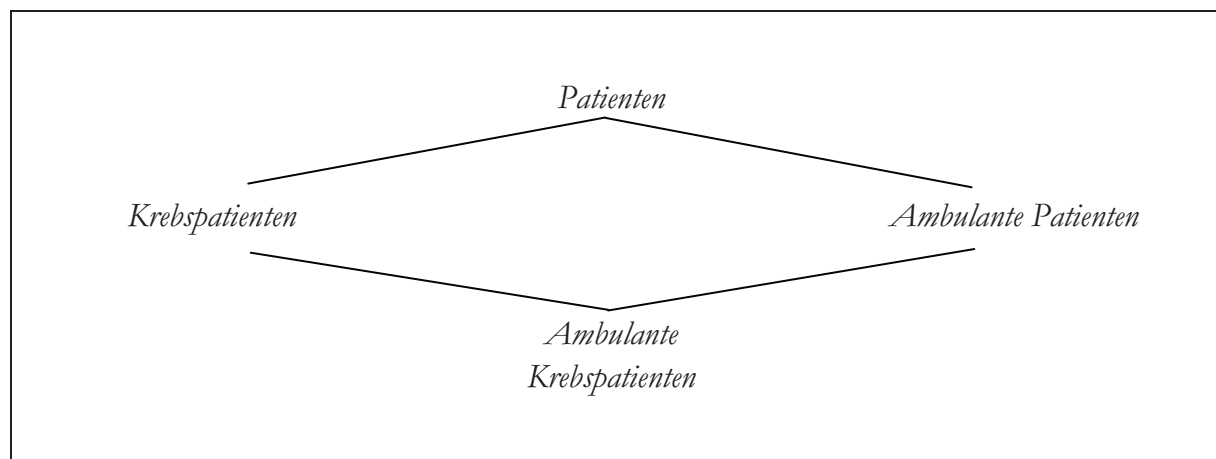
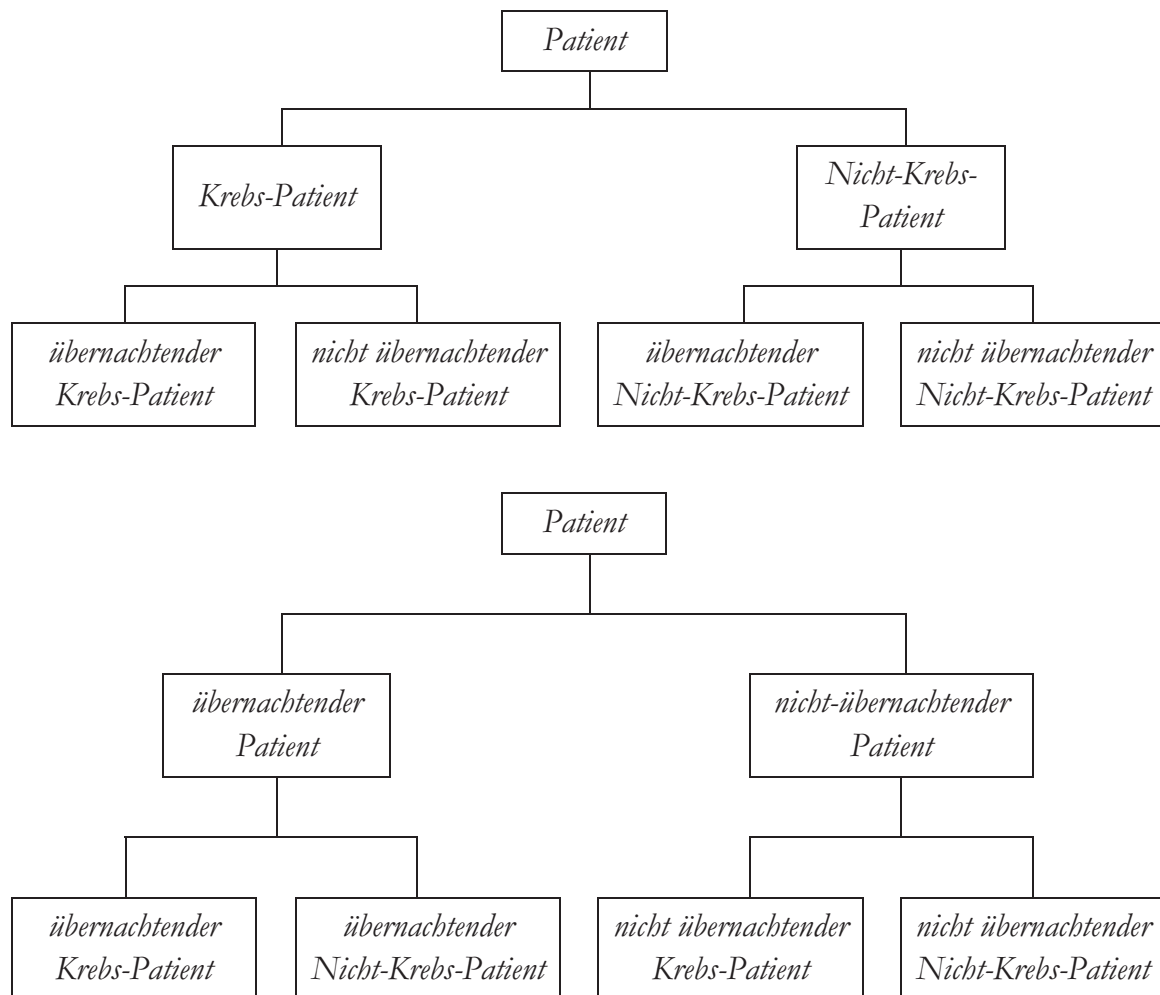


Abb. 4.3: Ein Beispiel für multiple Vererbung



Zweitens stellt sich die Frage, welches der beiden Merkmale in der Hierarchie der Klassifikation zuerst abgearbeitet werden soll: Sollen wir die Patienten zuerst in solche mit bzw. ohne Krebs unterteilen und diese beiden Patientenklassen jeweils in solche, die ambulant und solche, die stationär behandelt werden? Oder sollen wir vielmehr die Patienten zuerst in einerseits ambulant und andererseits stationär zu behandelnde unterteilen, und dann diese beiden Klassen jeweils in Unterklassen von Patienten, die an Krebs leiden oder die nicht an Krebs leiden? Wir haben also die Wahl zwischen den beiden in Abbildung 4.4 dargestellten Baumstrukturen.

Abb. 4.4: Zwei alternative Baumdiagramme



Auf der untersten Ebene der Bäume, der Ebene der „Blätter“, ergeben sich in beiden Fällen dieselben vier Typen. Vermutlich wäre es für die medizinische Praxis völlig irrelevant, welche der beiden Möglichkeiten wir wählen würden. Aus philosophischer Sicht ist eine solche Arbitrarität, also die Mög-

lichkeit zu einer willkürlichen Entscheidung dieser Frage, allerdings ein bemerkenswertes Phänomen. In Kapitel 9 wird eine Diagnose der Ursache dieser Arbitrarität vorgeschlagen werden, aus der sich dann auch eine Empfehlung zum Umgang mit diesem Phänomen ergibt. An dieser Stelle sei nur soviel angedeutet: Eine Kreuzklassifikation beruht gerade darauf, dass sie nach dem Vorliegen bzw. Nichtvorliegen zweier voneinander unabhängiger Merkmale fragt. Im Falle der Patienten im NCIT sind dies die Fragen:

- (Q1) Aus welchem Grund wird der Patient behandelt?
 (Q2) Bleibt der Patient über Nacht?

Abb. 4.5: Mehrdimensionale Klassifikation in SNOMED II⁵³

Welche morphologische Struktur?	Morphologie	M
Findet sich wo?	Topographie	T
Was ist die Ursache?	Ätiologie (<i>etiology</i>)	E
Mit welchem Effekt?	Funktion	F
Bei welcher Krankheit?	Krankheit (<i>disease</i>)	D
Welche Prozeduren sind erfolgt?	Prozedur	P
Zusammenhang mit dem Beruf?	Beruf (<i>job</i>)	J

Eine Möglichkeit, solche Aspekte voneinander zu trennen, ist die „mehrachsig“ oder „mehrdimensionale“ Klassifikation. Dieses Verfahren wird zum Beispiel von SNOMED, der *Systematized Nomenclature of Human and Veterinary Medicine*, angewandt, die vom amerikanischen *College of American Pathologists* entwickelt worden ist.⁵⁴ In der Version 3 unterscheidet SNOMED elf verschiedene Klassifikationsachsen, die mit 17 Qualifikatoren verknüpft werden können. Abbildung 4.5 listet einige dieser „semantischen Achsen“ von SNOMED auf. Nicht für jedes Krankheitsbild wird jede dieser Achsen benötigt, aber indem mehrere dieser Achsen herangezogen werden, kann zum Beispiel eine Virusenzephalitis bei einem Waldarbeiter kodiert werden als

TX2000 M40000 E30000 J63230,

⁵³ Dugas und Schmidt 2003, 80. Zur Kritik an SNOMED vgl. Bodenreider et al. 2007.

⁵⁴ Vgl. <http://www.snomed.org> und Winger 1984.

wobei der mit „T“ beginnende Teil der Kodierung den Ort der Krankheit, der mit „M“ beginnende Teil den betroffenen Körperteil, der mit „E“ beginnende Teil die Ursache der Krankheit (den Virus) und der mit „J“ beginnende Teil den Beruf des Betroffenen angibt.

4.2.3 Uniformität: Labortiere

Patienten zugleich nach ihrer Krebsdiagnose und nach ihrem Übernachtungsstatus zu klassifizieren, führt nicht nur zu Problemen mit der Disjunktivität und Exhaustivität, sondern verstößt zugleich auch gegen das Uniformitätsgebot: Merkmale aus zwei ganz verschiedenen Bereichen werden hier zusammengebracht. Noch drastischer als bei der Klassifikation der Patienten fällt dies bei der Klassifikation der Labortiere im NCIT auf, die, soviel steht fest, für die Krebsforschung äußerst wichtig sind. Dies lässt sich auch daran erkennen, dass der Obertyp *Laboratory_animal* im NCIT in zwölf vielfältige Untertypen unterteilt wird. Die Merkmale, mit Hilfe derer diese Untertypen gebildet werden, stammen allerdings aus ganz unterschiedlichen Bereichen. Der NCIT verstößt hier also gegen das Uniformitätsgebot.

Einige der Untertypen von *Labortier* basieren auf Dingen, die den fraglichen Tieren widerfahren sind. Eine *Genetically_Engineered_Mouse* wird zum Beispiel vom NCIT definiert als eine „Maus, die durch das Hinzufügen neuer genetischer Charakteristika genetisch verändert wurde“. In diesem Beispiel wird das Erleiden der DNA-Manipulation als Wesen der *Genetically_Engineered_Mouse* angegeben. Andere Untertypen wie *Control_Animal* basieren auf der jeweiligen Rolle, die das Tier innerhalb eines Versuchaufbaus einnimmt. Denn Kontrolltiere, so vermerkt der NCIT, sind „die Tiere in einer Studie, die nicht die zu testende Behandlung erfahren“.

Diese Definitionen berufen sich folglich auf Unterscheidungsmerkmale aus ganz unterschiedlichen Bereichen. Sie stammen aus ganz unterschiedlichen Kategorien: natürliche Arten, Leiden, funktionale Rollen. Diese kategorialen Unterscheidungen (mit denen sich das nächste Kapitel näher befassen wird) sollten in einer guten Klassifikation berücksichtigt werden.

4.2.4 Meta-Typen und „Anderes“

Auch hinsichtlich der Explizitheit und Präzision lässt der NCIT zu wünschen übrig. So, wie es in CTT einen Eintrag „und so weiter“ gibt, gibt es im NCIT einen Eintrag *Other*, und zwar als Untertyp des Typs *General_Modifier* (ein Un-

tertyp von *Qualifier*, das wiederum ein Untertyp von *Properties_and_Attributes* ist), der vom NCIT als „different than the one(s) previously specified or mentioned“ definiert wird. Außer der Entsprechung zum bloßen „Andere(r/s)“ gibt es im NCIT noch ca. 80 andere Einträge der Form „Andere F“, zum Beispiel *Carcinoma,_Other,_of_the_Mouse_Pulmonary_System*.⁵⁵

Ganz wie CTT besitzt der NCIT auch Meta-Typen, also Typen, die ihrerseits von der Klassifikation abhängig sind, von der sie ein Teil sind. So enthält der NCIT zum Beispiel die Kategorie *NCI-Thesaurus_Property*, die ein Untertyp von *Property* ist und definiert wird als eine „specific terminology property present in the NCI Thesaurus“. Meta-Typen kommen im NCIT sogar in der obersten Ebene der Klassifikation vor. In dieser Ebene listet der NCIT nämlich den Typ *Retired_Concept* auf, den er definiert als ein „concept [which] has been retired, and should not be used except to deal with old data“. Hier geraten, um Freges Terminologie zu verwenden, eindeutig *Begriffseigenschaften* und *Begriffsmerkmale* durcheinander.⁵⁶ Während es ohne Zweifel nützlich ist, in einer Klassifikation auch Eigenschaften der vorkommenden Ausdrücke oder Begriffe festzuhalten, so sollten diese doch nicht so behandelt werden, als wären sie unabdingbare Merkmale, die Einzeldinge haben müssen, um die fragliche Universalie zu instantiieren.

4.3 Einschränkende Randbedingungen für Klassifikationen

Die obige Kritik an Borges' CTT und am NCIT hat diese an einem Idealbild von Klassifikationen gemessen. Nach diesem Idealbild besteht eine Klassifikation aus paarweise disjunkten Klassen, die gemeinsam den Gegenstandsbereich ausschöpfen und durch ontologisch fundierte Merkmale konstituiert werden. Es gibt eine Reihe von Gründen, warum real existierende Klassifikationen von diesem Idealbild abweichen.

Eine erste Gruppe von einschränkenden Randbedingungen hängt von dem Gegenstandsbereich ab, der zu klassifizieren ist. Gerade wenn es um die Klas-

⁵⁵ Für eine eingehende Kritik an solchen Formulierungen wie „other“ bzw. „not otherwise specified“ („NOS“) vgl. Bodenreider, Smith und Burgun 2004. – In deutschsprachigen Informationssystemen wird oft der Ausdruck „o.n.A.“ („ohne nähere Angabe“) verwendet.

⁵⁶ Vgl. Frege 1884, § 53 und ders. 1892.

sifikation biologischer Arten geht, besteht das Problem, dass eine große Zahl von Tier- und Pflanzenarten wissenschaftlich noch nicht beschrieben oder noch gar nicht entdeckt worden ist. Zudem werden mit neuen genetischen Methoden Artunterschiede entdeckt, die durch traditionelle phänotypische Methoden nicht unterschieden werden konnten. Schon die schiere Anzahl der Arten bietet eine Garantie dafür, dass den Biologen auf absehbare Zeit die Arbeit nicht ausgehen wird: Allein die Anzahl der Tierarten wird auf etwa 30 Millionen geschätzt. Es mag Bereiche geben, die, wie zum Beispiel die Anatomie des Menschen, so gut wie vollständig und abgeschlossen wissenschaftlich erfasst sind. Andere Bereiche hingegen haben ein stetes Wissenswachstum zu verzeichnen, wie die Zoologie, die Botanik und ganz besonders die Genetik, welche aufgrund der Datenintensität der genetischen Analysen ohne computergestützte Datenverarbeitung heute kaum mehr möglich wäre. Prinzipiell muss jedoch immer mit der Entdeckung weiterer Arten gerechnet werden, nicht zuletzt auch deswegen, weil neue Arten entstehen können. Dies würde eine Grenze für die Exhaustivität einer Klassifikation setzen, die durch den zu klassifizierenden Gegenstandsbereich bedingt ist. Manche Gegenstandsbereiche geben dem Klassifikator auch grundsätzliche Probleme auf. Da die Gene von Bakterien auf vielfache Weise untereinander ausgetauscht werden können und sie sich durch die hohe Teilungsrate schnell ändern können, ist es besonders schwer, bei Bakterien stabile und trennscharfe Gattungen und Arten zu unterscheiden. Auch dies sind Grenzen der Klassifikation, die sich aufgrund des Gegenstandsbereiches ergeben.

Eine zweite Gruppe von einschränkenden Randbedingungen hängt ab von der technischen Seite der Realisierung und Anwendung der Klassifikation. Ganz gleich, ob diese in traditioneller, gedruckter Form oder in Form einer computergestützten Datenbank vorliegt: Der zur Verfügung stehende Speicherplatz ist endlich, ob er nun in Druckseiten oder in Festplattenkapazität besteht. Sollen Computerprogramme automatische Schlussfolgerungen aufgrund der eingegebenen Daten durchführen (*automated reasoning*), gibt es zusätzlich zum Speicherproblem das Problem der Berechenbarkeit. Die nötige Rechenzeit wächst mit der Anzahl der Klassen, die zu verwalten sind, und den Relationen zwischen den Klassen, die repräsentiert werden sollen (vgl. Kap. 8), an. Und je nach gewählter Programmiersprache und der ihr zugrunde

liegenden Beschreibungslogik und ihrer Ausdruckskraft⁵⁷ besteht stets die Gefahr, dass eine gestellte Aufgabe überhaupt nicht berechenbar ist.

Unabhängig von der Leistungsfähigkeit der verwendeten Hard- und Software ergeben sich auch durch den menschlichen Benutzer einschränkende Randbedingungen für Klassifikationen. Denn während es immer leichter und billiger wird, die Speicher der Computer zu erweitern, haben die kognitiven Fähigkeiten ihrer menschlichen Benutzer enge Grenzen. Menschlichen Dokumentatoren in Archiven und Bibliotheken wird empfohlen, nicht mehr als etwa eintausend systematisch angeordnete Schlagworte zum Indexieren von Büchern oder Dokumenten zu verwenden.⁵⁸ Computer können selbstverständlich deutlich mehr Termini verwalten; der *NCI Thesaurus* mit seinen 36.000 Termini ist dabei keineswegs die größte Terminologie-Datenbank. Das *Unified Medical Language System* (UMLS) zum Beispiel umfasste im Jahr 2001 schon 1,9 Millionen Ausdrücke mit über 800.000 unterschiedenen Bedeutungen.⁵⁹ Jedoch sind es menschliche Kuratoren, die diese Datenbanken erstellen, und menschlichen Benutzer, die sie später verwenden. Die Kuratoren sind Datenbank-Experten, die sich oft ausschließlich der Entwicklung dieser Art von Wissensspeichern widmen. Aber spätestens dann, wenn ein Hausarzt seine Abrechnungen aufgrund der Diagnosen-Codierung einer solchen umfangreichen Datenbank erstellen soll, stellt sich erneut die Frage, wie viele Diagnosen-Codes er in seinem Praxisalltag zu beherrschen vermag.

Es sieht so aus, als ob das anvisierte Ideal nicht vollständig realisiert werden kann, weil unsere Wünsche sich wechselseitig beschränken: Wenn einerseits eine vollständige Repräsentation eines Wissenschaftsbereiches angestrebt wird, ist dies vermutlich weit von einem einfachen, verständlichen und benutzerfreundlichen System entfernt. Zudem können Probleme mit der Berechenbarkeit auftreten, wenn Vollständigkeit darin besteht, eine große Menge an Daten anzuhäufen und viele Relationen zwischen den Einträgen zu definieren. Wenn andererseits vereinfachende Formulierungen wie „Andere“ oder „Nicht anderweitig bestimmt“ verwendet werden, entstehen Probleme beim Aktualisieren der Klassifikation, denn in unterschiedlichen Versionen kann „Andere“ ganz unterschiedliche Extensionen haben. Wird jedoch davon

⁵⁷ Für einen Überblick vgl. Nardi und Brachman 2002.

⁵⁸ Vgl. Gaus 2003, 93–94.

⁵⁹ Vgl. Dugas und Schmidt 2003, 81.

abgesehen, solche „Andere“-Typen zu verwenden und stattdessen kurzerhand das JEPD-Kriterium aufzugeben, gehen beträchtliche Möglichkeiten verloren, Schlüsse aus unseren Daten zu ziehen. Denn dann wissen wir zum Beispiel nicht mehr, ob eine Entität, die zu einem Obertyp gehört, auch zu genau einem seiner Untertypen gehört.

4.4 Referenzontologien: Ein Lösungsversuch

Ein jüngerer Vorschlag zur Lösung dieses Dilemmas besteht in einer klaren Aufgabenteilung. Es werden schlicht zwei Arten von Systemen benötigt: Referenzontologien und Anwendungsontologien. Die Referenzontologien sollen ohne Rücksicht auf Speicherprobleme und Rechenzeiten entwickelt werden. Sie sollen den Stand der jeweiligen Wissenschaftsdisziplin, aus der sie stammen, möglichst vollständig widerspiegeln:

„Eine Referenzontologie ist analog zu einer wissenschaftlichen Theorie; sie hat einen einheitlichen Gegenstandsbereich, der aus Dingen besteht, die unabhängig von der Ontologie existieren, und sie versucht, die deskriptive oder repräsentationale Angemessenheit ihrem Gegenstandsbereich gegenüber zu optimieren, in dem höchsten Grad, der mit formaler Strenge und informationstechnischer Nützlichkeit kompatibel ist. Da eine Referenzontologie analog zu einer wissenschaftlichen Theorie ist, besteht sie aus Repräsentationen der biologischen Realität, die korrekt sind, wenn sie im Licht unserer gegenwärtigen Verständnisses betrachtet werden (und sie sollte daher einem regelmäßigen Update im Lichte des wissenschaftlichen Fortschritts unterworfen werden).“⁶⁰

Eine Anwendungsontologie hingegen ist analog einem technischen Produkt wie einem Computerprogramm. Bisher ist es üblich gewesen, für jeden Anwendungsfall von Grund auf neue Ontologien zu entwickeln. Das erschwert folglich den Austausch und das Vergleichen von Daten. Anders ist es, wenn auf eine bereits bestehende Referenzontologie zurückgegriffen werden kann, von der ausgehend man die Anwendungsontologie erhält, indem man eine auf einen bestimmten Zweck hin zugeschnittene Auswahl oder Kombination von Typen der Referenzontologie vornimmt. Verschiedene solcher Anwen-

⁶⁰ OBO Evaluation Criteria, Version 7 (January 14, 2006), http://sourceforge.net/mailarchive/forum.php?forum_id=12140&max_rows=25&style=nested&viewmonth=200602; meine Übersetzung. Dort auch das folgende Zitat.

dungsontologien sind dann durch ihre jeweilige Rückbindung an eine gemeinsame Referenzontologie leicht aufeinander abbildbar.

Während also die Aufgabe der maximal-adäquaten Realitäts-Repräsentation der Referenzontologie zugewiesen wird, übernehmen es die Anwendungsontologien, das Speicher- und Berechenbarkeitsproblem in den Griff zu bekommen und Rücksicht auf die menschlichen Benutzer zu nehmen. Die zur *OBO-Foundry* zusammengeschlossenen Wissenschaftler sehen darin einen entscheidenden Fortschritt:⁶¹ „Die Methode, Anwendungsontologien ausschließlich vor dem Hintergrund von formal robusten Referenzontologien zu entwickeln und ihre Aktualisierung im Lichte der Aktualisierung der zugrunde liegenden Referenzontologie sicherzustellen, kann sowohl diesen Tendenzen der Vermehrung von Ontologien entgegenwirken, als auch die Interoperabilität der Anwendungsontologien sicherstellen, die nach ihren Vorgaben konstruiert worden sind.“

4.5 Exotisches Denken? Oder ein ungeeignetes Werkzeug?

Manche Philosophen haben gemeinsam mit Foucault darin übereingestimmt, dass Borges' CTT den „exotischen Zauber eines anderen Denkens“ offenbart.⁶² Es konnte gezeigt werden, dass CTT in der Tat darin bezaubernd ist, eine große Bandbreite von Fehlern und Problemen illustrieren zu können, die bei der Konstruktion von Taxonomien vorkommen können. Davon abgesehen ist CTT ein großartiges Stück Literatur, aber kein Zeugnis der Wissenschaft. Selbst wenn es ein solches wäre, wäre es kein Beweis für ein exotisches Denken, sondern eher ein Beispiel für eine sehr unpraktische Taxonomie. Es wäre jedenfalls ebensoviel oder ebensowenig ein Beispiel für exotisches Denken wie der NCIT, der seinem Anspruch nach ein wissenschaftliches Produkt ist, aber dieselben Eigentümlichkeiten aufweist, die CTT so seltsam erscheinen lassen. Die diskutierten Teile des NCIT sind allerdings keineswegs Beispiele exotischen Denkens, sondern tatsächliche Beispiele für höchst unpraktische Taxonomien. Das *National Cancer Institute* jedenfalls, das den NCIT geschaffen hat und unterhält, ist nicht mit seinem gegenwärtigen Zustand und seinem mutmaßlichen „exotischen Charme“ zufrieden und will den NCIT

⁶¹ Zum Projekt der *OBO Foundry* vgl. auch Smith et al. 2007.

⁶² Foucault 1974, 17. Vgl. auch Jullien 2004.

dahingehend verbessern, solche Klassifikationsfehler zu vermeiden.⁶³ Dies wäre ein hervorragendes Beispiel dafür, dass technische Anwendungen auf den von der Philosophie gelegten Fundamenten aufbauen können.

⁶³ Review of NCI Thesaurus and Development of Plan to Achieve OBO Compliance, <http://www.cbd-net.com/index.php/search/show/938464> (eingesehen am 24.01.2006). Für weitere Kritik am NCIT vgl. Ceusters, Smith und Goldberg 2005.