

Diagnosekodierung als Interpretation sprachlicher Zeichen

Hans Rudolf Straub, Kreuzlingen

Summary. The ICD (International Classification of Diseases) is used worldwide for the coding of medical diagnoses. A computer system that could automatically relate every possible diagnosis text to the corresponding code would be very useful. The difficulties in achieving this goal are not caused by the absolute number of codes (>15'000), but by the high variability in text phrasing and the complexity of the underlying medical semantics. To attain the goal, we first must distinguish strictly between words (signifiant) and concepts (signifié). On the side of the concepts, we must interweave different hierarchical trees, the weaving design being a critical aspect of the semantic representation. The structure of the ICD itself reflects a typical dialectic tension: on the one hand the full richness and variability of real life should be captured and on the other hand the ICD as a standard should be uniform, simple and unchangeable. The method (concept molecules) that we use allows us to relate the correct ICD codes automatically to diagnoses formulated in natural language.

Zusammenfassung. Die ICD (International Classification of Diseases) wird weltweit angewendet, um Diagnosen zu verschlüsseln. Ideal wäre ein System, das jeder möglichen Textformulierung automatisch den entsprechenden Kode zuordnen würde. Die Problem liegt dabei nicht in der Anzahl der Kodes (>15'000), sondern in der Vielfalt der Textvariationen und der komplexen medizinischen Semantik. Zur Bewältigung der Aufgabe muss klar zwischen Wörtern (signifiant) und Begriffen (signifié) unterschieden werden. Auf der Begriffsseite müssen unterschiedliche Hierarchieebenen miteinander verwoben werden, wobei das Webmuster selber ein wesentlicher Teil der semantischen Darstellung ist. Die ICD-Einteilung der Diagnosen steht als Standard in einem dialektischen Spannungsfeld: Einerseits soll die gesamte Vielfalt und Dynamik des realen Lebens abgebildet werden, andererseits soll der Standard eine Norm sein, die fix, einheitlich und einfach ist. Die von uns eingesetzte Methode (Begriffsmoleküle) ermöglicht es, frei formulierte Diagnosen automatisiert zu kodieren.

1. Die internationale Klassifikation der Krankheiten

Die Diagnose der Krankheit eines Patienten spielt eine zentrale Rolle für seine Behandlung. Doch nicht nur im Einzelfall interessiert die Frage, welche

Diagnose ein Patient hat, auch statistisch möchte man gerne wissen, welche Krankheiten auftreten.

Die Epidemiologen arbeiten deshalb seit dem 19. Jahrhundert an standardisierten Klassifikationen, welche zuerst nur einige wenige, damals lebensbedrohliche Krankheiten wie Cholera und Typhus auflisteten. Die anfänglich nationalen Klassifikationen wurden zusammengefasst und als ICD (International Classification of Diseases) während des 20. Jahrhundert kontinuierlich erweitert. Heute umfasst die 10. Revision dieser Klassifikation (=ICD-10) praktisch das gesamte Krankheitsspektrum. Sie wird von der WHO gepflegt und weltweit angewendet.

Die ICD-10 erlangt aus verschiedenen Gründen eine zunehmende Bedeutung. In der Schweiz sind die Spitäler gesetzlich verpflichtet, die Diagnosen aller Patienten dem Bundesamt für Statistik anonymisiert für epidemiologische Zwecke bekannt zu geben. In Deutschland werden die Krankenhäuser ihre Leistungen in Zukunft über diagnosebasierte Fallpauschalen¹ abrechnen. Damit will man mehr Gerechtigkeit erreichen, indem die Krankenhäuser für die Behandlung von Diagnosen bezahlt werden und nicht unkontrolliert für Leistungen, deren Notwendigkeit niemand überprüfen kann.

In beiden Fällen erfolgt die Mitteilung der Diagnose nicht direkt, sondern über die entsprechenden ICD-10 Codes. Der Grund dafür ist klar: Die Möglichkeiten, eine Diagnose zu formulieren, sind unglaublich vielfältig. Es gibt nicht nur eine Vielzahl von Ausdrücken, die Ärzte teilen ihre Diagnosen auch nicht immer auf die gleiche Weise ein. Die ICD-10 Klassifikation hingegen umfasst eine endliche Anzahl von Codes und ist auf den ersten Blick klar hierarchisch strukturiert. Mit der ICD-Klassifikation erhofft man sich also eine Zunahme an Klarheit.

Jemand muss allerdings die ICD-Codes den als Texten formulierten Diagnosen zuweisen. Dazu muss er sowohl das Kodewerk kennen wie die ärztliche Sprache. Er muss eine Übersetzung durchführen von der Sprache der Ärzte in die Struktur des vereinfachenden Kodewerkes. Für den oberflächlichen Betrachter mag dies als eine Banalität erscheinen, doch die Tücken zeigen sich bei der Durchführung.

Dies gilt für menschliche Kodierer. Soll die Übersetzung automatisiert – durch ein Computerprogramm – durchgeführt werden, stösst die starre Logik der Maschine erst recht an Grenzen. Trotzdem ist eine automatisierte Kodezuordnung durch die Maschine realisierbar. Ich möchte im folgenden

auf einige oft übersehene Aspekte hinweisen, die bei der automatisierten Kodierung eine Rolle spielen.

2. Diagnosen im semiotischen Dreieck

Ein Wort als geschriebenes oder gesprochenes Zeichen ist nicht dasselbe wie die Bedeutung des Wortes. Saussure spricht von *signifiant* und *signifié*; das Zeichen ist das *signifiant* und die Bedeutung das *signifié*.

Mit *signifié* ist bekanntlich nicht das bezeichnete Objekt – zum Beispiel der Gegenstand *Tisch*, auf den das Wort *Tisch* hinweist - gemeint, sondern als ein Drittes die Vorstellung, die das Wort Tisch im Sprecher einnimmt und im Zuhörer erzeugt. Wir haben es also mit drei sehr unterschiedlichen Elementen zu tun:

1. Dem Gegenstand (Objekt, Bezeichnetes, Signifikat)
2. Dem Wort (Zeichenausdruck, *signifiant*, Signifikant, Symbol)
3. Dem Begriff (Zeicheninhalt, *signifié*, Interpretant, Konzept, Vorstellung)

Dass diese drei Elemente drei völlig unterschiedlichen Welten angehören und dass die Regeln, wie wir mit diesen Elementen umgehen, je nach der Welt, der sie angehören, völlig unterschiedlich sind, ist eine Tatsache, auf die Saussure, Peirce, Ogden und Richards und viele andere hingewiesen haben. Jede der drei Welten spielt bei der Diagnosekodierung, das heisst bei der Zuordnung von ICD-Kodes zu Textdiagnosen, ihre Rolle.

Der Welt des Bezeichneten, das heisst der realen Objektwelt, entspricht bei der Kodierung der Patient und sein Zustand. Wenn der Arzt diesem Zustand eine Bezeichnung, das heisst eine Diagnose zuordnet, dann geschieht dies aufgrund seiner Vorstellung über die Krankheiten, ihre Ursachen und ihre Symptome. Diese Vorstellungswelt, die sich aus dem, was er während Studium und Beruf gelernt und erfahren hat, zusammensetzt, erlaubt es ihm, das Zustandsbild, das der Patient ihm bietet, zu verstehen, indem er die beobachteten Symptome (Realwelt) mit den Soll-Symptomen seiner Lehrvorstellungen (Konzeptwelt) verbindet.

Natürlich ist dieser Diagnoseprozess ein bidirektionaler: ein Symptom kann für mehrere Diagnosen zutreffen, der Arzt wird deshalb den unentscheidbaren Zustand in seiner Konzeptwelt dahingehend korrigieren, dass er die Direktion der Bestimmung umkehrt. Wenn das beobachtete Symptom (Realwelt) das Konzept nicht vollständig determiniert, wird er aus der Konzeptwelt eine Frage hervorholen, um diese Frage auf die Realwelt anzuwenden. Die Frage bedeutet, dass er in der Realwelt gezielt nach einem weite-

ren Symptom (zum Beispiel einem Laborwert oder einer subjektiven Äußerung des Patienten) sucht, welches ihm die Entscheidung für eine bestimmte Diagnose (Konzeptwelt) erlaubt. Zwischen Realwelt und Konzeptwelt erfolgt auf diese Weise ein unablässiges Hin und Her, bis der Arzt zufrieden ist mit seiner Interpretation der Situation (Abbildung 1).

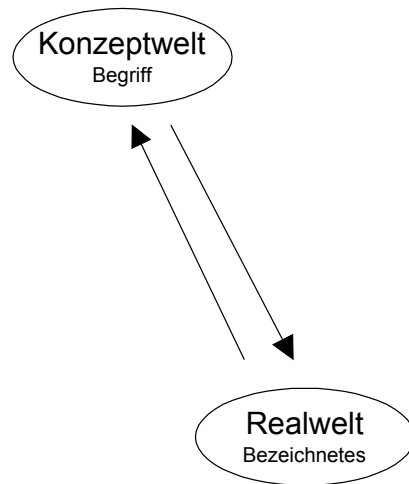


Abbildung 1: Diagnosefindung (bidirektional)

Damit hat er seine Diagnose gefunden. Die Diagnose des individuellen Arztes widerspiegelt in der Regel die Vorstellungen, welche die Ärzteschaft als Gesamtes von der Welt der Krankheiten hat. Der Arzt teilt seine Vorstellungen mit seinen Kollegen. Hätte er nicht während des Studiums und in den Praktika die bereits ausgetesteten Vorstellungen seiner Lehrer in seine eigene Konzeptwelt einbauen können, wäre er nicht in der Lage, die Krankheit des Patienten nach Ärzteart zu verstehen. Er hätte – wie ein Laie – die zur Diagnose führenden Fragen an die Realwelt gar nicht erst stellen können. Die Ärzteschaft ist existentiell auf den Austausch ihrer Ideen und Vorstellungen angewiesen – und damit kommt das dritte Element ins Spiel, das Wort (Abbildung 2).

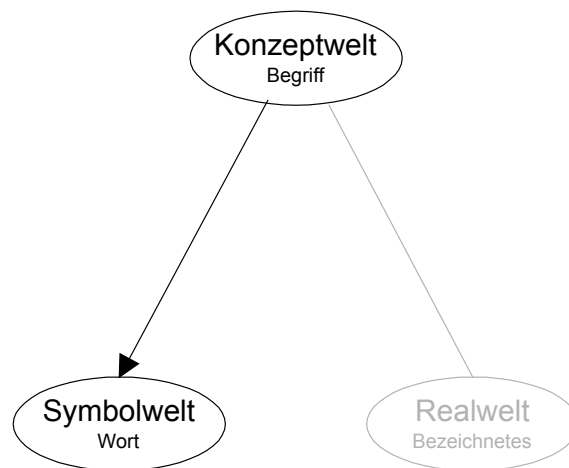


Abbildung 2: Diagnoseäusserung: der Begriff als Wort

Der Arzt spricht eine Diagnose aus, doch das unwichtigste dabei sind die Phoneme. Die Diagnose, die er äussert, entspricht vielmehr einem bestimmten Ort in seinen Vorstellungen, der mit Tausenden von anderen Orten in seiner Konzeptwelt auf eine ganz bestimmte Weise verbunden ist. Erst die Verknüpfung all dieser Punkte definiert die Diagnose (signifié) wirklich. Wenn er die gefundene Diagnose ausspricht (signifiant), ist dies einfach ein Wort. Ein Wort unterliegt – wie ein Begriff – bestimmten Regeln, aber es sind andere als beim Begriff. Wort und Begriff, so eng verbunden sie sein mögen, gehören zwei verschiedenen Welten an. Das Wort folgt den Regeln der Grammatik, der Begriff hingegen denen der Semantik, in unserem Beispiel der Semantik der Medizin, und die Regeln der beiden Welten sind sehr unterschiedlich. Hier geht es vorrangig darum, die Regeln der Semantik zu beschreiben.

Mit Abbildung 2 ist das semiotische Dreieck mit allen drei Ecken gezeichnet. Was geschieht nun bei der Diagnose-Kodierung? Wo im Dreieck findet sich der ICD-Kode?

3. Kode - ein Wort mit unterschiedlichen Bedeutungen

Wir müssen uns klar darüber sein, dass das Wort "Kode" sehr unterschiedlich verwendet wird². Die Bedeutungen des Wortes "Kode" unterscheiden sich in zwei Richtungen³:

3.1. Kode als Regelwerk und als Kodierungsergebnis

Eine Kodierung kann generell angesehen werden als eine Umformung einer Information in einem System in eine Information in einem anderen System.

Das Wort "Kode" wird bei dieser Umformung für zwei verschiedene Dinge benützt, einerseits für das gesamte Regelwerk, den Hintergrund also, auf dem die Kodierung stattfindet, und andererseits für den konkreten Output einer Umformung, das heisst für das einzelne Kodierungsergebnis. Am Beispiel des Morsecodes bedeutet Kode somit einerseits das gesamte System mit allen Zuordnungsregeln⁴ und andererseits das Resultat einer einzelnen Kodierung (zum Beispiel die Zeichenfolge ". -" als Morsecode für den Buchstaben "A").

Im Zusammenhang mit der Diagnosekodierung kommen beide Vorstellungen zum Zug. Wenn wir von einem ICD-Kode sprechen, dann meinen wir das Resultat der Kodierung, also zum Beispiel den Code "A00.0" für die klassische Cholera. Man könnte mit Kode zwar auch die gesamte ICD-Regelammlung meinen, in der Regel ist mit ICD-10-Kode aber der konkrete Einzelcode gemeint. Anders ist es beim Verhältnis von Wort und Begriff. Viele Autoren sprechen bei der Relation von signifiant und signifié von einem Kode. Damit ist natürlich das Gesamtwerk der Regeln gemeint, die signifiant und signifié verbinden. Wie genau dieses Regelwerk aussieht, ist damit aber noch nicht gesagt. Genau das interessiert uns aber. Wenn wir eine Kodierung von medizinischen Diagnosetexten durch den Computer durchführen lassen, müssen wir explizit wissen, wie der Bezug zwischen Worten und Begriffen ist, das heisst wir müssen diesen "Kode" im Detail aufschlüsseln.

3.2. Informationserhaltende und -verlierende (gruppierende) Kodierung

Beim Morsecode enthält die kodierte Form genau die gleiche Information wie die unkodierte, denn es handelt sich um eine 1:1-Abbildung, bei der jedem Buchstaben genau ein Morsezeichen entspricht. Deshalb ist es möglich, den Morsecode auch wieder zu entziffern und den Vorzustand wiederherzustellen. Das gleiche – wenn auch nicht in derselben offensichtlichen 1:1-Abbildung – geschieht bei der Transformierung eines akustischen Signals in den Strom eines Telefondrahtes und zurück und bei der Kryptographie zur Verschlüsselung von Nachrichten, die nicht jedermann lesen soll. Immer ist es die Absicht, das ursprüngliche Signal am Ende möglichst gleich wieder herstellen zu können. Die kodierte Form ist nur ein Zwischenzustand. Wenn das Ziel der 1:1-Abbildung nicht perfekt erreicht wird, zum Beispiel bei der akustischen Wiederherstellung aus dem elektrischen Telefonsignal, dann liegt das nicht in der Absicht der Kodierung, sondern daran, dass technisch der Prozess nur annähernd perfekt sein kann.

Etwas ganz anderes ist es bei der Diagnosekodierung. Hier ist es das *Ziel* der Kodierung, Information zu verlieren. Eigentlich handelt es sich bei der ICD-Kodierung nicht um eine Kodierung, sondern um eine *Klassifizierung*, das heisst um eine *Gruppierung*. Ein und derselbe "Kode" fasst verschiedene Diagnosen zusammen, was durchaus in der Absicht der Kodierung liegt. Es geht darum, viele ähnliche Fälle zusammenzufassen, um so eine Übersicht über die Fälle zu bekommen und sich nicht im Detail zu verlieren. Das "C" im Kürzel ICD bedeutet dementsprechend auch "Classification" – und nicht etwa "Coding". Eine Kodierung ist die ICD aber insofern, als die Bezeichnungen der ICD-Klassen "Kodes" sind, das heisst kryptisch wirkende Buchstaben-Zahlen-Kombination, deren Sinn erst mit einer nicht allen bekannten Anleitung verständlich wird.

Nicht nur die ICD-Kodes, schon die Diagnosen aus denen sie entstehen, sind gruppierende Klassen (siehe Abbildung 1), die verschiedene ähnliche, aber durchaus unterscheidbare Krankheitszustände der Realität - zum Beispiel die Blinddarmentzündung von Frau Müller und die von Herrn Meier – zusammenfassen. Ganz allgemein ist jedes Konzept eine Vereinfachung, die auf mehrere Zustände der Realwelt angewendet werden kann. Bei dieser Vereinfachung gehen die Informationen über den Einzelfall verloren.

Gerade dies macht ja die Idee eines Konzeptes oder einer Klasse aus, dass es sich um eine verallgemeinernde Vereinfachung handelt. Die Verallgemeinerung, die das Konzept darstellt, ist die Bedingung, dass das gleiche Konzept in verschiedenen Situationen immer wieder neu anwendbar ist. Dies erlaubt es uns, die Welt nicht jedes Mal von Grund auf neu verstehen zu müssen, sondern im Neuen das bekannte Muster zu erkennen und uns so in der neuen Situation an der vorhandenen Erfahrung orientieren zu können. Der Preis aber ist, dass durch die Vereinfachung viele Detailinformationen verloren gehen.

Die ICD-Kodierung ist in diesem Sinn eine typische Vereinfachung und als solche eine Fortführung der bereits stattgehabten Vereinfachung von der Realwelt zur Text-Diagnose. Beides sind gruppierende Prozesse.

Eine gruppierende Kodierung unterscheidet sich von der informationserhaltenden Kodierung des Morsecodes und den kryptographischen Verfahren darin, dass mit Absicht und unwiederbringlich Information verloren geht⁵. Diese Beobachtung mag auf den ersten Blick abstrakt oder theoretisch erscheinen, ist aber von eminent praktischer Bedeutung. Denn der Informationsverlust, der bei jeder Gruppierung notwendigerweise stattfindet, führt

bei der Diagnosekodierung zu einigen der typischen Probleme, auf die ich weiter unten eingehen möchte.

4. Diagnosekodes im semiotischen Dreieck

Wenn ich hier vom Diagnosekode spreche, meine ich – wie dargestellt – das Resultat der ICD-Kodierung. Im semiotischen Dreieck ist ein solcher Diagnosekode ein Symbol (signifiant) für eine Vorstellung (signifié), genauso wie ein Wort ein Symbol für ein signifié ist. Abbildung 3 zeigt den Prozess der Diagnosekodierung in drei Schritten.

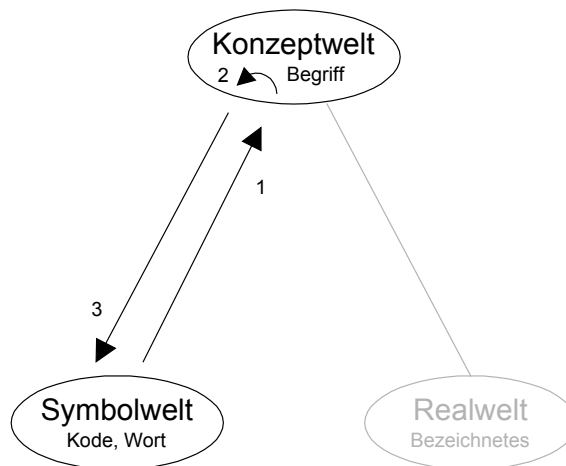


Abbildung 3: Diagnosekodierung: vom Wort zum Kode

Wieder geht der Prozess in zwei Richtungen. Einmal muss zur textlichen Formulierung der Diagnose (Wort) die entsprechende Vorstellung in der Konzeptwelt gefunden werden (1). Dann muss – unter Kenntnis des Regelwerkes der ICD-10 – diese Vorstellung einem Konzept der ICD-10 zugeordnet werden (2). Sobald das ICD-Konzept klar ist, kann der entsprechende Kode ausgegeben werden (3).

Der Hauptteil der Arbeit findet dabei in der Konzeptwelt statt. Weil die textliche Formulierung der Diagnose sich nicht an denselben Konzepten ausrichtet wie die Einteilung der ICD (auf Details gehe ich etwas später ein), ist insbesondere Schritt 2 oft sehr problematisch. Doch auch Schritt 1 ist nicht einfach. Die Konzeptualisierung der Wörter beinhaltet die Rekonstruktion ihrer impliziten Inhalte. Dies ist keinesfalls banal. Die Begriffe, die hinter den Diagnoseformulierungen stecken, müssen für eine Kodierung durch den Computer klar und explizit dargestellt werden können. Dabei entstehen viele Probleme; auf einige werde ich weiter unten eingehen. Die Frage der

formalen Notation der Konzepte und ihrer Relationen ist ganz entscheidend. Die Welt der medizinischen Semantik ist weitläufig und die Regeln in dieser Welt sind komplex. Ohne eine gleichzeitig einfache, präzise und mächtige Notation für die Semantik lässt sich ein Regelwerk für die automatisierte Diagnosekodierung nicht erstellen und schon gar nicht warten.

Die Welt der Textdiagnosen ist reicher als die Welt der Codes, weshalb in der Regel sehr viele Diagnosen zum selben Code führen. Weil die Konzeptgrenzen von Code und Text nicht übereinstimmen, kann es umgekehrt aber auch sein, dass der *gleiche* Text zu verschiedenen Codes führen könnte. In diesem Fall muss der Ast nach rechts in Abbildung 3 verfolgt werden, das heisst die Diagnose muss durch Nachfragen um diejenigen Aspekte aus der Realwelt ergänzt werden, die in der textlichen Formulierung bisher fehlten.

Insgesamt ist der Prozess der Diagnosekodierung ein sehr komplexes Geschehen, das sich aber trotzdem weitgehend automatisieren, das heisst durch ein wissensbasiertes Computerprogramm durchführen lässt. Auf einige typische Tücken der Diagnosekodierung werde ich jetzt im Detail eingehen.

5. Eigenschaften der Konzeptwelt

5.1. Unsichtbarkeit

Ein Gedanke ist nicht sichtbar. Wenn er sich auf die Realwelt bezieht, ist sichtbar, worauf er sich bezieht, der Gedanke selber ist aber nicht sichtbar. Er ist formulierbar und die Worte sind dann wieder sichtbar. Doch die Worte sind nicht der Gedanke, sondern nur Symbole des Gedankens. Die Worte gehören zur Symbolwelt, die Konzeptwelt der Gedanken ist durch Worte nur indirekt, nicht direkt sichtbar.

Da – wie in Abbildung 3 dargestellt – die Hauptarbeit der Diagnosekodierung in der Konzeptwelt stattfindet, ist es sehr erstrebenswert, eine Notation für die Elemente der Konzeptwelt zu haben. Weil die Elemente aber unsichtbar sind, kann man leicht Gefahr laufen, sie falsch darzustellen.

Wenn wir die Elemente der Konzeptwelt so darstellen wie die Elemente der Symbolwelt, das heisst die Wörter, dann gehen wir von einer zwar naheliegenden, aber unbewiesenen Annahme aus. Zwischen den Wörtern (Symbolwelt) und den Begriffen (Konzeptwelt) bestehen wesentliche Unterschiede, die meines Erachtens bei jeder Wortinterpretation berücksichtigt werden müssen.

5.2. *Mehrdimensionalität kontra Linearität*

In der Sprache unterliegen wir einer sehr grossen Einschränkung: der Linearität. Ob gesprochen oder geschrieben, immer folgt ein Wort auf ein anderes, ein Buchstabe (Phonem) auf einen andern. Jedes Wort hat maximal zwei Nachbarn. Komplexere Strukturen sind nicht darstellbar, jedenfalls nicht explizit. Bei den Gedanken ist das anders: Ich kann sehr viele Begriffe gleichzeitig miteinander verbinden und die gleichzeitige Verbindung der Begriffe, die genaue Anordnung der einzelnen Elemente mit eingeschlossen, macht den Gedanken aus. Wenn ich ihn formuliere, bin ich jedoch an die Zeitachse gebunden und kann jeweils nur genau ein Wort hinter das andere setzen. Es macht die Kunst der Formulierung aus, dass im Kopf des Zuhörers trotzdem ein komplexes Gebilde entstehen kann.

Für eine Darstellung der Begriffe in der Konzeptwelt hingegen muss ich die komplexe Verknüpfung der Begriffe anstreben, das heisst ich sollte mehrere Begriffe in ihrer gegenseitigen Verknüpfung explizit darstellen können.

5.3. *Überschneidungen*

Ein Wort in einem Satz ist von einem anderen klar abgegrenzt. Es können nicht zwei Wörter an der gleichen Stelle stehen. Bei den Begriffen ist das nicht so. Wenn ich "Tisch" sage, kann ich nicht gleichzeitig "Möbel" sagen, meinen kann ich das aber schon. Es ist typisch für die Konzeptwelt, dass die Begriffe auf diese Weise einander durchdringen. Der Begriff "Tisch" beinhaltet den Begriff "Möbel" und dieser den Begriff "Gegenstand". Wir denken uns die kleinsten Teile eines Gebietes, die Atome, gerne als kleine Kügelchen, mit einem klaren Ort und einer klaren Oberfläche und Begrenzung. Bei den Begriffen ist das – im Gegensatz zu den Wörtern – nicht so.

Bei einer Darstellung der Konzeptwelt müssen wir auf diesen Sachverhalt natürlich Rücksicht nehmen. Die Begriffe überdecken sich. Dies hat nichts mit Ungenauigkeit zu tun. Ungenau und mehrdeutig definierte Ausdrücke sind natürlich ebenfalls ein Problem bei der Textanalyse, die hier thematisierten Überschneidungen treten aber immer, auch bei ganz klar definierten Verhältnissen auf. Dass zwei Begriffe einander durchdringen, gehört zu ihrer Natur.

5.4. *Hierarchien – unter anderem*

Die oben genannte Begriffsfolge "Gegenstand – Möbel – Tisch" ist eine typische hierarchische Kette, ein Ausschnitt aus einem grösseren hierarchischen

Baum. Man könnte nun versucht sein, einen ganz grossen Baum herzustellen, der alle Begriffe des Fachgebiets beinhaltet. Wenn man diesen Baum erstellen würde, hätte man nicht nur alle Begriffe aufgezählt, sondern ihnen darüber hinaus eine Ordnung gegeben. Man könnte mittels der Baumstruktur vom Unterbegriff (dem Begriff auf der tieferen Hierarchiestufe) automatisiert auf den Oberbegriff (den Begriff der jeweils höheren Hierarchiestufe) schliessen und automatisiert alle Eigenschaften des Oberbegriffes auf den Unterbegriff vererben.

Berühmt und durch alle Jahrhunderte seither überliefert sind die Versuche der Athener Akademie von Sokrates, Platon und Aristoteles, solche Begriffsbäume zu erstellen. Doch nicht nur die Philosophen, auch die Informatiker mögen Bäume, denn die klare Baumstruktur kommt der Maschinennatur der Computer entgegen. Die Klarheit und Übersichtlichkeit eines Baumes macht ihn zudem für die Epidemiologie attraktiv. So wundert es nicht, dass auch die ICD-10 im Prinzip ein einziger grosser Baum ist.

Allerdings müssen wir hier den Traum der einen Begriffshierarchie mit aller Deutlichkeit relativieren. Bäume führen in der Praxis zu vielen Problemen. Und Bäume sind nicht die einzige Möglichkeit, Konzepte anzuordnen.

6. Probleme bei der Erstellung von Hierarchien

6.1. Vererbung mit Ausnahmen

Die Eigenschaften einer Kategorie werden auf die Unterkategorie vererbt. Wenn die Vögel fliegen, fliegen auch die Amseln und Spatzen. Die Strausse fliegen aber nicht und dürfen diese Eigenschaft nicht vererbt bekommen. Die hierarchische Struktur scheint hier verletzt, doch dies ist nur scheinbar. Der Eindruck entstand dadurch, dass die Eigenschaft "kann fliegen" als definierend für die Klasse "Vogel" angenommen wurde, was natürlich falsch ist. "Kann fliegen" ist eine typische, aber keine definierende Eigenschaft. Wenn wir beides auseinander halten, vererben wir zwar die Eigenschaft "kann fliegen" als typische Eigenschaft auf die Unterkategorien, setzen sie aber bei der Kategorie "Strauss" auf "fliegt nicht" um – ohne dass die Klasseneinteilung geändert werden muss.

Damit ist das Problem gelöst. Die wirklichen Probleme der hierarchischen Einteilung der Phänomene der Welt sind anderer Art, und ich werde gleich auf die wichtigsten eingehen:

6.2. Multihierarchie

Man beachte die folgenden zwei hierarchischen Ketten:

Diagnose – Leberkrankheit – Hepatitis – infektiöse Hepatitis – Virushepatitis
Diagnose – Infektionskrankheit – Viruskrankheit – Virushepatitis

Das gleiche Endblatt in der Hierarchie (Virushepatitis) kann über verschiedene Äste mit der Wurzel (Diagnose) verknüpft sein. Eine solche Struktur ist mathematisch⁶ gesehen kein Baum mehr, sondern eine Kreise enthaltende Netzstruktur. Trotzdem sind die beiden Hierarchiestränge die Art, wie Ärzte die Virushepatitis sehen; sie ist gleichzeitig eine Leberkrankheit und eine Infektionskrankheit. An dieser Stelle lässt sich die Baumstruktur nicht mehr aufrechterhalten. Die konzeptuelle Darstellung der medizinischen Diagnosen ist voll von solchen multihierarchischen Verknüpfungen. Wir sind deshalb gezwungen, eine bessere Darstellung für Begriffe zu finden als eine einfache Hierarchie es ist.

6.3. Hierarchiewechsel

Die Einteilung der Diagnosen spiegelt das medizinische Denken wider. Dieses ist aber einem steten Wechsel unterworfen. Ein typisches Beispiel sind die Geschwüre von Magen und Duodenum (Ulcus ventriculi und duodeni). Sie galten früher als ein diätetisches Problem. Durch zuviel Kaffee und andere Substanzen wird die Magenschleimhaut angegriffen. Die Therapie erfolgt dann durch einen Diätwechsel, zum Beispiel mit mehr Milch. Später galten diese Geschwüre als typische psychosomatische Krankheit. Ein bestimmter Persönlichkeitstyp reagiert auf Stress mit der Bildung von Magengeschwüren. Die Therapie erfolgt entsprechend psychosomatisch. Eine weitere Einteilungsmöglichkeit ist gegeben durch die Beobachtung, dass die Magensäure bei dieser Krankheit vermehrt ist. Durch Medikamente kann die Magensäure vermindert werden und das Symptom verschwindet. In jüngerer Zeit konnte überraschend nachgewiesen werden, dass ein Bakterium (*Helicobacter pylori*) bei diesen Geschwüren die Schleimhaut befällt. Heute werden deshalb oft Antibiotika zur Therapie der Ulzera verabreicht.

Die Frage für die Erstellung des Baumes ist jetzt folgende: In welches Kapitel fällt das Magengeschwür? In das Kapitel der Stoffwechselkrankheiten, der psychosomatischen Krankheiten oder der Infektionskrankheiten? Was ist, wenn wir den Baum strukturiert und das Magengeschwür entsprechend eingeteilt haben und sich die Lehrmeinung ändert? Das Beispiel zeigt, wie problematisch hierarchische Einteilungen sind.

6.4. Kombinatorische Explosion

Wenn wir in unserer Konzeptrepräsentation 1000 Krankheitserreger (Bakterien, Viren, Pilze und andere) aufnehmen und im Körper 20 Organe unterscheiden, dann erhalten wir 20'000 verschiedene Kombinationen von Erreger und Organ. Wenn wir nun die medizinischen Diagnosen in einem Baum darstellen wollen, müsste dieser Baum eigentlich alle 20'000 Kombinationen kennen. Auch wenn wir uns klar darüber sind, dass vielleicht die meisten dieser Kombination nicht möglich oder nicht wichtig sind, das Prinzip der Kombination von frei kombinierbaren Eigenschaften bleibt, und wenn wir die Diagnosen in einem grossen Baum darstellen wollen, dann bekommen wir zu viele Verästelungen. Wir müssen für die Lunge die wichtigen Erreger darstellen, genauso wie für die Hirnhäute oder die Blase; wir müssen für die Leistenhernie die relevanten Komplikationen aufführen und dieselben Komplikation für die Schenkel-, die Nabel- und die Narbenhernie wiederholen. Dies führt dazu, dass sich die Konzeptrepräsentation unglaublich aufbläht. Um dem abzuhelpfen, müssen wir der kombinatorischen Explosion der Realwelt in der Konzeptwelt etwas entgegensetzen, was selbst kombinatorische Potenz besitzt. *Kombinante Begriffsarchitekturen*, die einen Begriff aus mehreren Teilbegriffen zusammenstellen, besitzen diese Kompetenz, sind aber keine Bäume mehr.

6.5. Multipunktualität

Viele medizinische Diagnosen sind Syndrome mit mehreren einzelnen Unterdiagnosen, deren Kombination erst die Diagnose ausmacht. Die einzelne Unterdiagnose kann aber auch allein vorkommen und ist wieder eine Diagnose (Part-Whole-Problematik). In einem Baum bekommen wir bei der Darstellung eines solchen Sachverhalts ein Problem.

6.6. Variabel kombinierte Ursache – Folge – Komplexe

Ursache-Folge-Komplexe sind ein Spezialfall der eben genannten Multipunktualität. Medizinische Diagnosen sind oft Folgen von anderen Diagnosen, wobei der Bezug von Ursache und Folge zwingend ist – oder auch nicht. Eine Struma (= Kropf) zum Beispiel kann eine Folge eines Jodmangels, eines Schilddrüsentumors oder auch einer ganz anderen Krankheit sein. Ein Jodmangel kann, muss aber nicht zwingend zu einer Struma führen. Der Jodmangel kann zudem mit einer Hypothyreose (Mangel an Schilddrüsenhormon) verbunden sein, oder auch nicht. Die drei Diagnosen (Jodmangel, Hypothyreose und Struma) sind kausal miteinander verknüpft,

aber jeweils auch eigenständige Diagnosen und können in beliebiger Kombination miteinander oder mit anderen Krankheiten kausal verknüpft sein. In einem hierarchischen Baum, in dem jede Krankheit mit einem einzigen Endpunkt (=Blatt) repräsentiert sein soll, lässt sich so etwas kaum darstellen.

7. Die Abbildung der Diagnosen auf die ICD-10

7.1. Die ICD-10 als hierarchische Klassifikation

Das grundlegende Problem der ICD-10 ist, dass sie sich aus einer rein hierarchischen Einteilung der Diagnosen entwickelt hat. Im 19. Jahrhundert, mit nur wenigen in der Klassifikation aufgenommenen Diagnosen war das noch möglich, heute, wo die ICD-10 das gesamte Krankheitsspektrum abdecken will, kaum mehr. Die oben genannten Probleme der Multihierarchie, der kombinatorischen Explosion, der Multipunktualität und der Ursache-Folgen-Komplexe treten uns praktisch auf jeder der über 1000 Seiten des systematischen Verzeichnisses entgegen. Die Erbauer der ICD-10 sind sich der Probleme durchaus bewusst. Sie haben versucht, die inhärente Multihierarchie und die kombinatorische Explosion mit einem kombinanten Ansatz zu überwinden, und für gewisse Diagnosen wird nun die Kombination von zwei Codes, einem Primär- und einem Sekundärkode verlangt, zum Beispiel bei der Masernpneumonie:

Primärkode: *B05.2+ Masern, kompliziert durch Pneumonie*

Sekundärkode: *J17.1* Pneumonie bei anderen Viruskrankheiten*

Allerdings ist diese kombinante Kodierung keinesfalls konsequent durchgeführt.

Weil zudem viele Krankheiten in mehreren Kapiteln der ICD-10 eingeteilt sein können – je nachdem, ob die Pathogenese (Infekt, Stoffwechsel), das betroffene Organ (Atmungsorgan, Kreislauforgan), die Lebensumstände (Neonatalogie, Schwangerschaft), die Vererbung (Erbkrankheit), oder spezielle Ursachen (Verletzungen, Operationskomplikationen) im Vordergrund stehen – , überschneiden sich alle Kapitel. Die ursprünglich geplante Hierarchie lässt sich dadurch nicht konsequent einhalten, und deshalb finden sich im systematischen Verzeichnis viele Anweisungen, wann man welche Diagnose wo einzuordnen und zu kodieren hat. Diese oft sehr komplexen Anweisungen müssen bei der Kodierung berücksichtigt werden⁷.

Natürlich ist es sehr spannend, in dieser Situation eine automatisierte Kodeweisung durch den Computer durchzuführen zu lassen. Wir müssen dazu

vorgängig eine ausführliche semantische Analyse der Diagnosebegriffe vornehmen und sie in einem möglichst präzisen und übersichtlichen nichthierarchischen System darstellen. Gleichzeitig müssen wir die Systematik der ICD-10 analysieren und die einzelnen Codes im selben nichthierarchischen semantischen System darstellen. Die Zuordnung erfolgt dann auf der Basis des semantischen Systems (Schritt 2 in Abbildung 3).

Bei aller Kritik, die in meinen Ausführungen gegenüber der ICD-10 enthalten sind, möchte ich klarstellen, dass es viel einfacher ist, die ICD-10 zu kritisieren als eine bessere Klassifikation herzustellen. Die ICD-10 muss viele Zwecke gleichzeitig erfüllen. Dass sie dabei nicht jeden Zweck gleich optimal erfüllt, ist verständlich.

7.2. Der Einsatz komplexer Architekturen

Hierarchien sind, wie oben dargestellt, eine Möglichkeit für die Begriffsanordnung in der Konzeptwelt, wobei die begrenzte Anwendbarkeit der Hierarchien zu berücksichtigen ist. Um komplexe Verhältnisse darzustellen, müssen mehrere Hierarchien kombiniert werden. Am einfachsten geht das, indem man mehrere Achsen definiert und auf diesen Achsen Hierarchien erstellt. Ein solches mehrachsiges oder multidimensionales System ist zum Beispiel SNOMED⁸ mit – je nach Version – sieben oder elf Achsen.

Das Problem solcher Systeme ist, dass die Achsenzahl immer zu klein ist und deshalb auf den einzelnen – nun zu gross geratenen – Achsen wieder das Problem der kombinatorischen Explosion auftritt. Eine Vermehrung der Achsenzahl führt jedoch schnell zu Unübersichtlichkeit. Einfache mehrdimensionale Systeme im Stil von SNOMED sind deshalb für die Repräsentation von medizinischen Sachverhalten im allgemeinen und für die automatisierte Kodezuweisung im speziellen ungenügend.

Die Lösung besteht darin, die Achsenstruktur selber zum Thema der Begriffsrepräsentation zu machen. Die Achsen treffen sich jetzt nicht mehr alle im gleichen Punkt, sondern die einen Achsen können den anderen Achsen untergeordnet sein und von genau bestimmten Punkten auf den übergeordneten Achsen ausgehen. So wird die Achsenstruktur gleichzeitig übersichtlich und präzise (=multifokale Architektur⁹).

Allerdings garantiert auch eine multifokale Architektur nicht automatisch den Erfolg bei der Erstellung einer grossen Wissensbasis. Entscheidend ist es, die Darstellung von semantischem Wissen möglichst effizient zu gestalten, das heisst möglichst wenig Zeichen für das Dargestellte zu verwenden.

Man könnte hier von *Informationsdichte* sprechen: Bedeutung pro Zeichen. In der Praxis geht es darum, dass ein Mensch (der Wissensingenieur) möglichst schnell verstehen muss, was eine Wissensregel bewirkt. Er muss eine grosse Sammlung von Regeln überblicken können. Deshalb ist es wichtig, dass die Regeln komprimiert die relevante Information darstellen und nichts anderes.

Mit der Notation der Prädikatenlogik zum Beispiel können recht komplexe Sachverhalte ausgedrückt werden. Dabei wird die Darstellung aber schnell sehr umständlich und ein ausgedehntes, wissensbasiertes System ist in einer solchen Darstellung nicht wartbar. Die *Conceptual Graphs* nach J.F. Sowa (Sowa 1984) ermöglichen demgegenüber eine mehr intuitive Darstellung von komplexen Sachverhalten. Trotzdem lässt sich die Notation noch weiter verbessern. Um die ICD-Kodierung über das gesamte Diagnosespektrum der Medizin durchführen zu können, haben wir intensiv nach der dafür geeigneten semantischen Notation gesucht und notieren heute die Regeln in einer Form, die wir als *Begriffsmoleküle* bezeichnen (Straub 2001). Diese Notation erlaubt das Erstellen und Warten von sehr grossen Wissenssammlungen.

Begriffsmoleküle sind insbesondere dafür gebaut, komplexe Achsenhierarchien darzustellen; sie sind von Menschen und Maschinen gleich einfach und sicher zu lesen und erlauben eine übersichtliche und dabei sehr detaillierte semantische Modellierung.

Begriffsmoleküle nutzen die implizite Information, die in der "Topographie" des Papiers oder der Bildschirmoberfläche steckt: Ein Begriff der links von einem anderen steht wird automatisch als sein Oberbegriff interpretiert, ein Begriff der unter einem anderen steht, als sein Attribut. So lässt sich die Struktur der Konzepte konzentriert darstellen und schnell lesen. Die Erfahrung zeigt, dass auch Nicht-Mathematiker und Nicht-Programmierer (das heisst ganz normale Ärzte) eine ausgedehnte Wissensbasis ohne lange Einarbeitungsphase perfekt warten können.

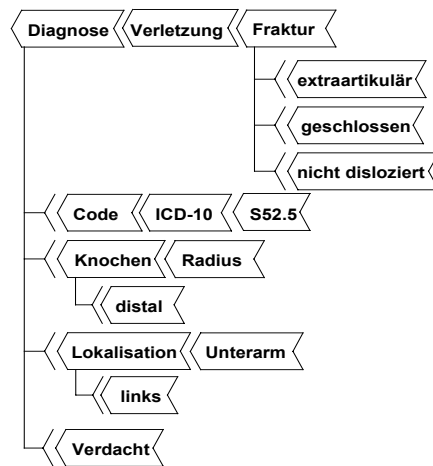


Abbildung 4: Beispiel eines Begriffsmoleküls für eine Radiusfraktur

8. Zur Dialektik von Vielfalt und Ordnung

Das Leben, das heisst die Realwelt (siehe Abbildung 3), hat viele Facetten und enthält viel mehr Information als jede noch so ausgeklügelte Interpretation je aufnehmen kann. Deshalb ist eine Konzeptwelt immer eine Vereinfachung der realen Verhältnisse. Der Verlust an Information wird dabei durch den Gewinn an Übersicht und Klarheit wettgemacht. Pointiert gesagt: Durch den Verlust an Information (Details der Realwelt) wird die Information (Aussage in der Konzeptwelt) überhaupt erst möglich (Straub 2001: 63-65 und 152).

Im Bereich der Lebewesen, der Tiere und Pflanzen, scheint eine eindeutige hierarchische Ordnung zu bestehen, derart, dass die Klassen, in denen sich die Arten einordnen lassen, wirklich einen typischen hierarchischen Baum ergeben. Diese Ordnung darf aber nicht auf alle Dinge, das heisst auf die gesamte Ontologie ausgedehnt werden. Die hierarchische Struktur der Pflanzen- und Tierarten ist eine Ausnahme und hat mit ihrer Entstehung, das heisst mit der Art der Fortpflanzung und der Evolution zu tun. Weil eine Art sich nicht mit einer anderen paaren kann, sind die Klassengrenzen fest. Und weil die eine Art sich aus der anderen entwickelt hat, widerspiegelt der entstandene Artenbaum die Geschichte der Artenentstehung. Die Artenentstehung erfolgt entlang des unidirektionalen Zeitpfeils. Dadurch erfolgen Verzweigungen immer in die gleiche Richtung und Kreisstrukturen sind ausgeschlossen - mit anderen Worten: ein idealer Baum entsteht. Diese Art Einteilung oder Ontologie ist jedoch eine Ausnahme.

Gegenstände können nicht auf die gleiche zwingende Weise in eine monohierarchische Klassierung eingeordnet werden, weil sie nicht wie die biologischen Arten in einer gemeinsamen Geschichte entstehen und weil ihre Klassengrenzen nicht eindeutig sind. Dasselbe gilt für die Einteilungen innerhalb einer biologischen Art. Die Menschen kann man zum Beispiel nach Sprachen einteilen, nach Ländern, nach Rasse, nach Geschlecht, nach Beruf, nach Religion, nach Haarfarbe, nach intellektuellen oder musischen Befähigungen und so weiter. Keine dieser Einteilungen ist eindeutig und keine ist der anderen wirklich überlegen. Monohierarchische Klassierungen, die die gesamte Welt einteilen, sind eine Illusion.

Trotzdem haben hierarchische Einteilungen ihren Zweck. Die Klarheit und Einfachheit, die in einem hierarchischen Baum steckt, hat unbestreitbare Vorteile. Gerade bei medizinischen Diagnosen ist es sehr zweckmässig, einfach und übersichtlich zu strukturieren. Primär um die Konzepte der Diagnosefindung und der Therapiestellung zu ordnen. Denn ohne Ordnung in seinem Kopf könnte kein Arzt Patienten behandeln. Das medizinische Wissen kann nur mit einer bestimmten Ordnung gelernt und gelehrt werden.

Auch wenn wir wissen wollen, welche Krankheiten wie häufig vorkommen und wie sich diese Häufigkeiten entwickeln (Epidemiologie), brauchen wir einfache und übersichtliche Einteilungen. Vieles spricht für die Klarheit der Hierarchie. Trotzdem müssen wir uns bewusst sein, dass die Hierarchie in den meisten Fällen ein künstliches Konstrukt ist, entstanden aus der Notwendigkeit, die Realität für die Interpretation zu vereinfachen.

Weil die Strukturierung in der Konzeptwelt eine Vereinfachung gegenüber der Realwelt ist, enthält sie immer eine gewisse Willkür. Nur ein Teil der ursprünglich in der Realwelt vorhandenen Informationen (Zeichen) kann beachtet werden, der grösste Teil muss ignoriert werden. Was wird behalten und was verworfen? Die Entscheidung darüber liegt nicht allein in der betrachteten Sache (Objekt), sondern auch im Kenntnisstand und in der Interessenlage des Betrachters (Subjekt) und ist nicht endgültig zu treffen, wie nicht zuletzt die Medizingeschichte nahelegt.

Anmerkungen

¹ siehe (Fischer 1997).

-
- ² Das Wort "Kode" kodiert für die unterschiedlichsten *Begriffe*. "Kode" ist somit ein typischer Fall dafür, dass es sehr nützlich ist, Wort (signifiant) und Begriff (signifié) voneinander zu unterscheiden.
- ³ Mit den beiden "Richtungen" sind hier semantische Achsen, Dimensionen oder Freiheitsgrade gemeint: voneinander unabhängige offene Wahlmöglichkeiten der Interpretation eines Zeichens oder Wortes. Die Vorstellung der *semantischen Achse* spielt bei der Konzeptmodellierung eine grosse Rolle (siehe Kapitel 7.2.).
- ⁴ Beim Morsekode eine einfache zweispaltige Tabelle. Andere Codes sind aber etwas komplexer.
- ⁵ Zu den beiden Bedeutungen von Kode siehe (Straub 2001: 151).
- ⁶ Graphentheorie.
- ⁷ Die Anweisung in der ICD-10 Systematik sind nicht immer eindeutig genug, weshalb in Deutschland ausführliche zusätzliche Kodierrichtlinien (DKG 2001) erlassen wurden. Eindrückliche Beispiele für die Komplexität der Richtlinien siehe bei (Fiori 2002).
- ⁸ Standardized Nomenclature of Medicine.
- ⁹ Eine detaillierte Beschreibung der verschiedenen semantischen Architekturen erfolgt in (Straub 2002).

Literatur

- Cimino J.J (1998), "Desiderata for Controlled Medical Vocabularies in the Twenty-First Century". *Methods of Information in Medicine* 37: 394-403.
- DIMDI, Deutsches Institut für Medizinische Dokumentation und Information (1994/95): ICD-10, *Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision* (Bände I/II/III). München: Urban & Schwarzenberg.
- DIMDI, Deutsches Institut für Medizinische Dokumentation und Information (2003): *ICD-10-GM Systematisches Verzeichnis. Version 2004*, Köln: MediComBooks.
- DKG, Deutsche Krankenhausgesellschaft (2001), *Deutsche Kodierrichtlinien Version 2002*. Düsseldorf: Deutsche Krankenhausgesellschaft.
- Fiori W. et al. (2002), "Kodierung in der Geburtshilfe – ein Buch mit 7 Siegeln?". *Das Krankenhaus* 8: 620-629.
- Fischer W. (1997), *Patientenklassifikationssysteme zur Bildung von Behandlungsfallgruppen im stationären Bereich*. Wolfertswil: Z/I/M-Verlag.
- Hausser R. (2001), "The Four Basic Ontologies of Semantic Interpretation". In: H. Kangassalo et al. (eds.), *Information Modeling and Knowledge Bases XII*. Amsterdam: IOS Press Ohmsha.
- MacKay D.M. (1969), *Information, Mechanism and Meaning*. London and Beccles: Clowes and Sons.

-
- Rector A.L. (1999), "Clinical Terminology: Why Is it so Hard?". *Methods of Information in Medicine* 38: 239-52.
- Sowa J.F. (1984), *Conceptual Structures: Information Processing in Mind and Machine*. Reading: Addison-Wesley.
- Sowa J.F. (2000), *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Pacific Grove: Brooks/Cole.
- Straub H.R. (1994), "Wissensbasierte Interpretation, Kontrolle und Auswertung elektronischer Patientendossiers". In: *Kongressband der IX. Jahrestagung der Schweizerischen Gesellschaft für Medizininformatik*. Nottwil: Schweizerische Gesellschaft für Medizininformatik.
- Straub H.R. (2001), *Das interpretierende System*. Wolfertswil: Z/I/M-Verlag.
- Straub H.R. (2002), "Four Different Types of Classification Models". In: R. Grütter (ed.), *Knowledge Media in Healthcare: Opportunities & Challenges*. Hershey / London: Idea Group Publishing.
- Wingert F. (1984), *SNOMED Manual*. Berlin: Springer.

Hans Rudolf Straub
straub@semfinder.com