

Wissensorganisation mit Hilfe des semiotischen Thesaurus – auf der Basis von SGML bzw. XML

von
Walther Umstätter

Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin

Zusammenfassung

Die Dokumentare blicken heute auf eine rund hundertjährige Diskussion über die richtige klassifikatorische Erschließung von Dokumenten und ihren thematischen Inhalten zurück. Daneben hat sich die ISKO speziell der Frage nach der Organisation des in den Dokumenten enthaltenen Wissens durch Klassifikation, Kategorienbildung und Thesaurierung gewidmet. Dieser, durch die Möglichkeiten des Volltextretrievals zeitweilig etwas in den Hintergrund geratene Erfahrungsschatz, gewinnt durch die neuen sogenannten semantischen Thesauri eine ungeahnte Renaissance. Dabei ist der allgemeinen Aufmerksamkeit entgangen, dass die sich aus der Semiotik ableitende Entsprechung des semantischen Thesaurus, die des pragmatischen Thesaurus, völlig unbeachtet blieb. Eingehende Überlegungen zu dieser Thematik zeigen, dass es sich bei den semantischen Thesauri eigentlich um semiotische Thesauri handelt, weil sie beiden Aspekten, dem semantischen und dem pragmatischen, gerecht werden müssen. Thesauri sind in diesem Sinne die semiotische Schnittstelle zwischen den semantischen Objektbezeichnungen des Senders und der pragmatischen Begriffsbestimmung dieser Zeichen durch den Empfänger einer Nachricht. Es soll nun geprüft werden, ob und wieweit über die SGML- (*Standard Generalized Markup Language*) bzw. die XML-Philosophie (*eXtensible Markup Language*), Computern eine gewisse Begrifflichkeit vermittelbar ist.

Einleitung

Das deutsche Wort verstehen hat verschiedene homonyme Bedeutungen. Worin unterscheidet es sich vom Wissen?

Wir können eine Nachricht akustisch verstehen ohne sie zu begreifen. Damit liegt sie auf der Ebene der einfachen Kommunikationstheorie. Dies hat noch nichts mit dem pragmatischen Verständnis im Sinne der Semiotik zu tun, bei dem wir auch die Bedeutung der Nachricht erfassen, indem wir einer Bezeichnung eine Begrifflichkeit zuordnen. Wir können aber auch etwas verstehen im Sinne einer geistigen Durchdringung, die ein Wissen über das Empfangene voraussetzt.

Als Menschen durchschreiten wir diese drei Ebenen des Verstehens mit Hilfe der zunehmenden Komplexität unseres neuronalen Netzes in unserem Gehirn, das rein formal einem Thesaurus entspricht, in dem mehr oder minder komplexen Zeichenvernetzungen von der Brokischen Region unseres Gehirns bestimmte Benennungen zugeordnet werden.

Während Wilhelm von Humboldt, wie viele andere vor und nach ihm, noch die Meinung vertrat, der Mensch könne nur in Sprache denken, gelangte C.S. Peirce (S.186, 1967) zu der etwas allgemeineren Meinung, dass es nur möglich ist, in "Zeichen zu denken". Diese Zeichen können, wie wir heute wissen, recht komplexe neuronal vernetzte Strukturgebilde sein, denen

wir interessanterweise durch eine jeweils lokalisierbare Gehirnregion Namen geben, die wir dann ihrerseits als Zeichen zur zwischenmenschlichen Kommunikation verwenden. Es sei am Rande erwähnt, dass schon unsere „peinliche Verwandtschaft“, die Affen (ein Wort von Goethe), dieses Brokasche Zentrum nicht haben.

Die Translation von Begrifflichkeiten unseres Gehirns, in Benennungen der natürlichen Sprache, ist bei unserem menschlichen Denken zweifellos ein wichtiges Hilfsmittel zur Strukturierung von Wissen. In erster Linie dient es aber der Kommunikation und zum zwischenmenschlichen Austausch unseres Wissens. Dass wir nicht unbedingt in Sprache denken, sondern in Begriffen, lässt sich leicht daran erkennen, dass wir wiederholt in die Situation geraten, dass wir für bestimmte Vorstellungen, die wir haben und beschreiben möchten, keine Worte finden. Außerdem haben wir gerade in der semantischen Konstruktion von Thesauri das Problem, dass wir zu komplexen Begriffen, die unser Gehirn aufgrund von Erfahrung und Logik erzeugt hat, Namen erst neu erzeugen müssen. Selbstverständlich bestehen auch die Begriffe aus Zeichenketten, die im neuronalen Netz unseres Gehirns ausgetauscht werden. Sie haben aber noch nichts mit unserer natürlichen Sprache zu tun.

Wir schicken uns heute an, durch den Einsatz von entsprechenden Thesauri in Computern, diesen Maschinen ein Verständnis dafür zu vermitteln, was sie verarbeiten. Diese Thesauri können wir den Computer in vorbereiteter Form eingeben. Wir können die Maschinen aber auch über Programme in den Stand setzen, sich ihre eigenen Thesauri über sogenannte Ontologien selbst zu erwerben. Ein Beispiel dafür ist *CYC*.

Stellen wir uns dazu die Frage, wie der Mensch beim Erwerb von Bedeutungswissen selbst vorgeht. Wenn wir eine Person fragen: „Was verstehen Sie unter einer Bibliothek?“ So wird sie als Laie, also als Nichtbibliothekar, bisherige Erfahrungen assoziieren, die aus Erwähnungen ihrer Eltern, Lehrer oder Bekannten stammen. Sie erinnert sich an Gebäude, an denen sie schon vorbei ging oder die sie abgebildet sah, und an denen das Wort Bibliothek angeschrieben stand. Theaterstücke, Kriminalromane oder Fernsehfilme kommen ihr in den Sinn. Und aller Wahrscheinlichkeit nach wird sie sich auch an die eine oder andere Bibliothek erinnern, die sie schon selbst besucht hat. Der Mensch ordnet somit einer Vielzahl von Objekten, die alle eines gemein haben, eine Bibliothek zu sein, in eine Klasse. Er kategorisiert, in dem er benachbarte Begriffe mehr oder minder ausgrenzt. Bei genauer Betrachtung und möglichst exakter Ein- bzw. Ausgrenzung definiert er, was alle Bibliotheken gemein haben.

Kurzum, der Mensch treibt immer und überall Ontologie, in dem er versucht, die wahre Bedeutung von Worten und Zeichen aus ihren Zusammenhängen, in denen die Zeichen für ihn auftauchen, zu rekonstruieren. Er bemüht sich um Seinslehre im Sinne der modernen Linguistik, in dem durch Induktion von möglichst vielen Einzelfällen auf das Allgemeine geschlossen wird. Je einheitlicher seine Erfahrung ist, desto einfacher fällt ihm dabei die Definition und je umfangreicher sie ist, desto sicherer erscheint ihm seine Aussage. Andererseits wird sie aller Wahrscheinlichkeit nach um so heterogener sein, je umfangreicher seine Erinnerung ist. Als Laie wird man also, aufgrund der Frage „Was verstehen Sie unter einer Bibliothek?“, versuchen, eine möglichst typische Bibliothek die man kennt, zu beschreiben.

Der Bibliothekar dagegen hat nicht nur eine weitaus größere Erfahrung als der Laie im Hinblick auf die Vielfalt des Begriffs Bibliothek, er kennt meist sehr viel mehr Typen aus eigener Anschauung, er hat Fachliteratur gelesen und, was das Wichtigste ist, er hat gelernt die Fülle seiner Kenntnisse zu klassifizieren, und zu unterscheiden zwischen Öffentlichen, Wissenschaftlichen und Spezialbibliotheken, zwischen verschiedenen Systemen, Aufgaben und Bautypen etc.

Er hat damit einen Sprachgebrauch erlernt, den er mit seinen Fachkollegen in möglichst gemeinsamer Eindeutigkeit benutzt. Schon allein sein umfangreiches Wissen und die Vielzahl an Gesprächen und Diskussionen mit den Fachkollegen geben ihm ein recht gutes Bild davon, welche Worte, in welchen Zusammenhängen, wie gebraucht werden. Damit kennt er beispielsweise nicht nur die Übersetzung des deutschen Wortes Öffentliche Bibliothek in das englische *public library*, er weiß meist auch um den Unterschied zwischen diesen beiden verwandten Typen.

Es ist oft gesagt worden und auch zweifellos richtig, dass sich Fachsprache und Allgemeinsprache dadurch unterscheiden, dass die wissenschaftlich begründete Fachterminologie Begriffe klarer definiert, und es ist auch eines der wichtigsten Ziele der Wissenschaft zu klaren Begriffen zu gelangen. Auch die Lehre erfüllt folgerichtig ihre Funktion darin, die jeweils existente Fachterminologie als Voraussetzung dafür zu nehmen, ob jemand die Qualifikation als Fachfrau bzw. Fachmann erhält. In gewissem Sinne ist damit der Erwerb einer definierten Fachsprache die Basis für die Fachzugehörigkeit. Trotzdem ist man immer wieder erstaunt, wie selbstbewusst viele Fachleute zugeben, Teile der Fachterminologie nicht nur nicht zu kennen, sondern ihre Präzisierung sogar abzulehnen. Der Hintergrund dafür ist einerseits verständlich, weil schon in Prüfungen Wert darauf gelegt wird, die vorhandene Terminologie nicht nur einfach auswendig zu lernen, sondern durchaus kritisch zu hinterfragen. Darüber hinaus gibt es viele Definitionen, in denen die Unschärfe, die zu einem Begriff gehört, ein wichtiges Charakteristikum dieser Definition sein kann. Andererseits zeigt sich an dem Mangel an definitorischer Schärfe in einer Wissenschaft gerade deren Wissenschaftlichkeit.

Semantik und Pragmatik

Mit dieser Erfahrung vermag ein Experte auch den Wechsel von der Pragmatik zur Semantik zu vollziehen. Denn aus semiotischer Sicht könnten wir auch sagen, dass der Laie im Sinne von C.W. Morris (1938), der Pragmatik als "*relation of signs to interpreters*" definierte, zunächst nur semiotische Pragmatik betreibt. Der Laie versucht von Kind an zu ergründen, welche *essentia* hinter diesem, von verschiedenen Personen verwendeten und wiederholt auftretenden Wort steht. Was ist das Wesentliche an diesem Wort und welche *existentia* verbirgt sich dahinter.

Die Pragmatik wird allerdings oft mit der Semantik verwechselt, die sich mit den "*relations of signs to the objects to which the signs are applicable*" beschäftigt. Bei der so definierten Semantik haben wir es also mit der entgegengesetzten Situation zu tun. Angenommen, wir befinden uns in einer Bibliothek und überlegen, wie wir dieses Objekt gegenüber ähnlichen, aber davon zu differenzierenden anderen Objekten, am besten nennen. Was ist für diesen Raum, in dem sich so viele Bücher auf Regalen befinden, charakteristisch? Ist schon dieser Raum eine Bibliothek oder erst das ganze Gebäude? Gehören auch andere Gebäude zu dieser Bibliothek, und wie verhält es sich bei einer Virtuellen Bibliothek, deren Gebäude nur noch unreal existieren? Hier beginnt nun der Experte zu definieren. Er analysiert Bibliotheken nach gewissen Funktionen, nach den Medien die sie enthalten, nach ihren Aufgaben, nach ihrer Architektur, ihrem Alter, ihren Zielgruppen, etc. Als Ergebnis erteilt er diesen durch Kategorien abgegrenzten Objekten Namen. Er benennt Objekte mit Hilfe eines möglichst einsichtigen und umfassenden semantischen Netzes, so dass jeder, der seiner Logik folgt, diese auch nachvollziehen kann und ein möglichst vollständiges Gesamtbild erhält. Denn das Wichtigste für jede Kommunikation zwischen einem Sender und seinem Empfänger ist zunächst, dass sie einen gemeinsamen Zeichensatz verwenden, d.h. in unserm Fall, dass sie beide das Wort Bib-

liothek und seine Komposita kennen. Damit vermögen sie zu kommunizieren, so wie zwei Computer kommunizieren können, die den gleichen ASCII-Zeichensatz und möglicherweise auch den gleichen *spell checker* haben. Diese Computer tauschen Nachrichten aus, sie verstehen auch die Zeichen im Sinne der Informationstheorie, sie verstehen aber nicht ihre Bedeutung. Auf diesen wichtigen Unterschied haben schon Shannon und Weaver in ihrem grundlegenden Werk „*The Mathematical Theory of Communication*“ hingewiesen.

Erst durch die Nutzung eines gemeinsamen semiotischen Thesaurus (Schwarz, I. und Umstätter, W. 1999), in dem jedes Wortfeld von allen anderen verwandten Wortfeldern durch Hierarchie, durch Assoziation, durch Homonymie, Synonymie, etc. umrissen ist, kann die Verbindung von Benennung und Begriff erzeugt werden. Voraussetzung dabei ist, dass die Seite der Pragmatik eine möglichst abbildungsgetreue Entsprechung der Semantik zeigt.

Während also Ontologien, insbesondere die zur Zeit moderne linguistische Form davon, bei genauer Betrachtung den pragmatischen Aspekt von Thesauruskonstruktionen darstellen, bietet der wirklich semantische Aspekt, die Möglichkeit einer Thesauruskonstruktion, wie sie die Wissenschaft dringend benötigt.

Interessanterweise ist unsere natürliche Sprache von der Begrenztheit, Unschärfe und Fehlerhaftigkeit des Wortschatzes von Kleinkindern und Laien bis zur Vielschichtigkeit beliebiger sozialer und beruflicher Gesellschaftsschichten geprägt. Sie ist ohne jede Übertreibung ein perfektes Vehikel um unserer allgemeinen Unkenntnis Ausdruck zu verleihen. Wenn wir beispielsweise nicht wissen, ob wir vor einer Tanne, Fichte oder Lerche stehen, sagen wir verallgemeinernd Nadelbaum. Wenn wir hinsichtlich des Unterschiedes zwischen Laub und Nadelbäume unsicher sind, können wir auch einfach nur von Bäumen sprechen. Zur Ergründung dieser Feinheiten beginnen Kleinkinder mit der Frage nach dem Namen dieses Dings da. Auch sie treiben Ontologie, und zeigen uns damit auf einfache Weise, wie man sie erwirbt.

Dagegen zeichnen sich Fachterminologien durch ein Höchstmaß an Sprachdifferenzierung, Präzision und Korrektheit aus. Sie sind semantische Konstrukte, die in erster Linie durch eine wissensbasierte Logik bestimmt werden. Als Musterbeispiel kann hier die Mathematik gelten, die im Laufe der Jahrtausende, auf Grund von wenigen Axiomen, ein in sich geschlossenes und beeindruckend komplexes Gedankengebäude erreicht hat. Wie beeindruckend es für Kepler gewesen sein muss, zu sehen dass man auf Grund dieser Mathematik so weitreichende Erscheinungen wie den Lauf von Planeten berechnen konnte, zeigt der Titel seines Buches „*Harmonices mundi*“ von 1619. Für die Entstehung von Wissen ist es also nichts Neues, dass wir Gedankengebäude errichten, von denen wir erst später erkennen, dass sie eine Entsprechung in der realen Welt haben.

Der Unterschied eines aus semantischer Sicht erzeugten Thesaurus und eines solchen aus pragmatischer Sicht kann wohl am besten an der binären Nomenklatur von Linné gezeigt werden, der sein *Systema Naturae* (1735) auf den Überlegungen von Camerarius (*De Sexu Plantarum* 1694) aufbaute. Dieser hatte die Sexualität der Pflanzen entdeckt und damit die Voraussetzung geschaffen, die gesamte Natur unter dem Blickwinkel der Vererbung zu betrachten. Das Ergebnis der Systematisierung unter diesem Aspekt musste, wie wir heute wissen, zu der Entdeckung der Evolution führen, weil die zugrundegelegte Systematik als Kriterium der Klassenbildung, die Verwandtschaft aller Lebewesen benutzte. Schon Linné selbst muss dies geahnt haben, weil er in der letzten Auflage seines Werkes den Hinweis auf die Erschaffung aller Lebewesen gestrichen hat. Er hatte dies Ergebnis aber keinesfalls von Beginn an vorausgesehen und auch nicht voraussehen können. Im Gegenteil, erst seine Art der

systematischen Betrachtung hatte zu einer Wissensstruktur geführt, die unser gesamtes Weltbild geradezu revolutioniert hat.

Während wir in der natürlichen Sprache beispielsweise das Wort Mensch benutzen, dessen Etymologie für unseren allgemeinen Sprachgebrauch weitgehend unwichtig ist, legt die binäre Nomenklatur dafür *Homo sapiens* fest. Dabei steht bemerkenswerterweise die *sapientia*, für ein Lebewesen mit Einsicht, Verständnis bzw. Weisheit. Wichtiger ist in unserem Zusammenhang hier allerdings die direkte Verknüpfung mit den Ober- bzw. Unterbegriffen und die dazugehörige Definition. So ordnete schon Linné den *Homo sapiens* der Ordnung der Primates mit Mensch, Affe und Halbaffe zu. Damit gewann das Wort Mensch, als Übersetzung von *Homo sapiens*, eine veränderte Bedeutung. Er wurde eindeutig zum Tier erklärt – was bekanntlich zu einem Sturm der Endrüstung führte.

Ganz ähnlich wirkte sich die semantische Entwicklung der Nomenklatur in der Chemie aus. Dort hat man zunächst alle bekannten chemischen Elemente nach ihren Eigenschaften gesammelt, hat sie danach geordnet und gelangte so zum Periodensystem, das sich dann ebenfalls als eine Wissensstruktur darstellte, aus der sich noch unentdeckte Elemente mit Ihren Eigenschaften vorhersagen ließen.

Auch dies ist ein Thesaurus mit klassisch wissenschaftlichem Hintergrund, in dem die Objekte der Chemie semantisch begründete Bezeichnungen wie beispielsweise Fe^{++} für zweiwertiges Eisen (*Ferrum*) erhalten. Die Nomenklatur ist durch Ober- und Unterbegriffe gekennzeichnet. Sie enthält aber darüber hinaus auch die Angabe von Valenzen zur Kennzeichnung von Redoxpotentialen, weil die Chemie in erster Linie durch die Reaktion bzw. Oxidation von Stoffen gekennzeichnet ist.

Semantische Thesauri sind demnach in erster Linie dadurch gekennzeichnet, dass Objekte nach einem grundsätzlichen Kriterium (Axiomatik der Mathematik, Sexualität in der Biologie, Reaktivität in der Chemie) systematisiert werden. Nomenklatur und Begrifflichkeit erfahren aus dieser systematischen Logik heraus ihre Verknüpfung. Dabei ist die Art der Vergabe der Namen zweitrangig. Entscheidend ist die Stellung im System. Auf diesem Wege entstehen also Systeme, die aus ihrer immanenten Logik heraus eine gewisse pragmatische Selbständigkeit gewinnen. Sie zeigt sich meist darin, dass das System an einigen Stellen Objekte vorausagt, die zunächst noch nicht beobachtet werden konnten. Sobald sich diese Objekte in der Realität allerdings verifizieren lassen, steigt die Wahrscheinlichkeit, dass so entstehende Gedankengebäude eine reale Entsprechung besitzen. Semantik und Pragmatik sind somit zwei Aspekte des selben semiotischen Thesaurus. In der Messung von Wissen (Umstätter, W. 1998) zeigt sich die Bedeutung der Voraussagbarkeit in ihrem Umfang und in ihrer Sicherheit als entscheidendes Kriterium des Wissens. Dabei ist allerdings, bei einem geringen Vorhersagevolumen, die Wahrscheinlichkeit zu berücksichtigen eine richtige Vorhersage durch Zufall zu treffen.

Beim pragmatischen Ansatz der Ontologien, beobachten wir den Versuch einer Rekonstruktion solcher Semantiken. Sobald diese aber nicht auf einer wissenschaftlichen Grundlage basieren, sondern auf der unscharfen und oft laienhaften natürlichen Sprache, zeigt die innere Logik dieser Nomenklatur viele Widersprüche, Unschärfen und variierende Bedeutungen eines Wortes. So kann der Begriff Mensch im allgemeinen Sprachgebrauch als Homonym mit äußerst unterschiedlicher Bedeutung angesehen werden. Mensch als Artbegriff, als Individualbegriff, als geistiges, tierisches, sexuelles, schöpferisches Wesen, als Teil der Gesellschaft, als Ergebnis einer Ontogenie bzw. einer Phylogenie, als Kunstwerk, etc.

Thesauri und ihre Darstellung

Wir können also einen Thesaurus semantisch konstruieren, indem wir existierenden Objekten unter bestimmten Aspekten Zeichen zuordnen, oder pragmatisch rekonstruieren, indem wir Zeichen, die wir erhalten, kategorisieren und systematisieren.

Während in den herkömmlichen dokumentarischen Thesauri die Vernetzungen von Benennungen zur Darstellung der Begrifflichkeit in geradezu simplifizierter Form aus einer einfachen Hierarchie mit Ober- und Unterbegriffen bestand, erlauben uns die Computer weitaus komplexere Organisationsformen – die allerdings noch weit von den neuronalen Netzen menschlicher und tierischer Gehirne entfernt sind.

Die grobe Vereinfachung bekannter Thesauri hatte ihren Grund in der lange Zeit notwendigen Darstellung von Thesauri auf dem Papier. Eine weitere Vereinfachung lag in der Zusammenfassung zahlreicher sogenannter Nichtdeskriptoren zu Äquivalenzklassen, durch die der Sprachumfang des Thesaurus massiv auf die verwendbaren Deskriptoren reduziert wurde. Durch siehe-Verweisungen konnte auch auf andere infragekommene Hierarchien aufmerksam gemacht werden.

Gerade die Entwicklung von Hypertext bzw. Hypermedia hat die Möglichkeit geschaffen beliebige Links zur mehrdimensionalen Vernetzung der Benennungen zu erzeugen. Sicher ist diese Zahl äußerst gering, gegenüber der Zahl an Synapsen in neuronalen Netzen, sie ist aber ein interessanter Schritt in die Richtung vernetzter Begrifflichkeit. Dabei wird klar, dass der Unterschied zwischen Benennung und Begriff darin liegt, dass die Benennung das Wort im Thesaurus ist, während sich seine Begrifflichkeit aus der Vernetzung ergibt. So erhalten wir z.B. automatisch eine Vorstellung von der Bedeutung des Wortes Fahrzeug, wenn wir in der hierarchischen Unterordnung, Wörter wie Auto, Flugzeug, Schiff oder Zug aufgeführt finden. Hier können wir auch von Wörtern, und nicht von Worten sprechen, weil sie in keinem direkten semantischen Zusammenhang stehen. Wenn wir beispielsweise hören, dass jemand von einem Wasserfahrzeug spricht und wir in unserem einfachen Thesaurus nur einen Unterbegriff „Schiff“ haben, der die Bedingung erfüllt, sich auf dem Wasser zu bewegen, so „wissen“ wir, dass er nur dieser Deskriptor in Frage kommen kann. Damit haben wir gleichzeitig die Möglichkeit alle Unterbegriffe, die zum Deskriptor Schiff gehören, wie Bug, Kiel, Reling, etc., geistig zu vererben.

Solche Thesauri können Computern in verschiedenster Form vorgegeben werden. Als Mono- und Polyhierarchien, als Graphen, als *Frames*, als sogenannte semantische Netze oder auch als Texte in natürlicher Sprache. Insbesondere in der letzten Darstellungsform sind die größten Schwierigkeiten zu erwarten, weil die natürliche Sprache des Menschen mit am schwersten formalisierbar ist. Trotzdem beginnt man in den letzten Jahren sich mit Hilfe der Ontologien auf diese Form besonders zu konzentrieren, weil das Verstehen menschlicher Äußerungen durch Computer eine besondere Herausforderung darstellt. Außerdem gibt es gute Gründe anzunehmen, dass die Sprache noch immer das beste Vehikel zur Übertragung des Wissens ist, das wir in unserem Gehirn erzeugen. Zum Dritten bietet es sich an, Computer, denen heute Gigabyte an Texten zur Verfügung stehen, aus diesen Texten lernen zu lassen. Sie können über intelligente Agenten sich bestimmte Nachrichten suchen und können beispielsweise aus Texten wie:

Wissen ist Macht.

Information ist Wissen in Anwendung.

Richtig zu wissen bedeutet „durch Gründe zu wissen“.

Wissen ist nicht definierbar.

Ich weiß, dass ich nichts weiß.

Wissen als "Erkenntnis der Wahrheit aus ihren ersten Ursachen.

Die Götter haben sicheres Wissen - *epistème*, die Menschen haben nur Meinungen – *doxa*.

pragmatisch rekonstruierte Thesauri erstellen. Es ist klar, dass sie dazu Programme brauchen, die ihnen helfen die Aussagen auf Redundanz zu prüfen, auf Widersprüchlichkeit, auf Anakoluten und auf Zusatzinformationen, aus denen sie lernen können. Wenn sie die wahre Bedeutung der Worte und ihrer Aussagen verstehen wollen, müssen sie ihren genauen Zusammenhang und ihre Begründung kennen. Erst daraus kann festgestellt werden, worin die eigentliche Bedeutung der Aussage liegt und wie zuverlässig sie ist. Ein inzwischen schon klassisches Beispiel ist das im Internet gut vertretene *CYC*-Projekt von D.B. Lenat. In ihm wird der Versuch unternommen, die Bedeutung von Worten aus vorgegebenen Texten, durch Programmvorgaben zu ermitteln. Obwohl auch hier von Semantik gesprochen wird, handelt es sich bei diesem Ansatz eindeutig um einen, im Sinne der Semiotik, pragmatischen Ansatz.

Trotzdem sollte im Rahmen dieser Aktivitäten nicht vergessen werden, dass es außerhalb dieses im eigentlichen Sinne pragmatischen Ansatzes auch den semantischen gibt, der gerade aus wissenschaftlicher Sicht nicht weniger interessant ist.

Wie wir bereits eingangs sahen, ist „Verstehen“ und „Wissen“ nicht synonym, wenn das Verstehen nicht auf Begründung beruht. Schon in unserem einfachen Thesaurusbeispiel des Wasserfahrzeugs, zeigt uns die Erfahrung, dass es neben Schiffen auch andere Wasserfahrzeuge wie Flöße, U-Boote oder Surfbretter gibt. Dies ist einerseits eine Frage der Definition und damit eine Frage der Differenzierung des Thesaurus, andererseits ist es eine Frage der Logik, die diese Differenzierung notwendig macht bzw. ausschließt. So können wir zur besseren Identifizierung des Wortes Schiff definieren, dass Schiffe einen Antrieb haben müssen, dass sie oberhalb des Wasserspiegels schwimmen müssen oder auch, dass sie eine Mindestgröße haben müssen, wenn sie nicht zu den Schiffsmodellen gerechnet werden sollen. Unterlässt man eine solche Präzisierung, so könnte auch ein treibender Baumstamm als Schiff bezeichnet werden, was auf der Seite der Pragmatik erhebliche Schwierigkeiten in einer Kommunikation hervorrufen würde. In einer Reihe von Fällen können wir die Unterteilung von Begriffen auch so gestalten, dass die Logik weitere Unterteilungen verbietet. Dies gilt beispielsweise für sich ausschließende binäre Teilungen, wie Sein und Nichtsein, positiv oder negativ, etc.

Wir müssen aus dieser Perspektive heraus zwei Arten von Thesauri unterscheiden. Solche die in erster Linie dazu dienen die Bedeutung des Gesendeten zu verstehen und solche, die uns ein vertieftes Verständnis im Sinne von Wissen über das Gesendete, und damit übrigens auch über den Sender selbst, vermitteln. Im Prinzip ließe sich die Ver- und Entschlüsselung von Zeichen auf der Ebenen der Informationstheorie ebenso als Thesaurus ansehen, hier sprechen wir allerdings, zur besseren Unterscheidung von Kodierung und Dekodierung und nicht von Thesauri, obwohl die Zeichen in ganz entsprechender Weise hierarchisch angeordnet sind. Es sei nur an die Anordnung von Morseschlüsseln oder an die binäre Hierarchie der ASCII-Zeichen zur Datenübertragung bei Computern erinnert. An der Unterscheidung von Kodierung und Dekodierung lässt sich auch sehr schön die Parallele zur Semantik und Pragmatik erkennen. Sie sind durch die Kommunikation untrennbar miteinander verbunden.

Zusammenfassend lässt sich sagen: Die Systemimmanenz der Semantik ist die Voraussetzung zur pragmatischen Rekonstruktion der ontologischen Hermeneutik einer Aussage, die ein Sender an seinen Empfänger schickt.

Kommen wir in diesem Zusammenhang noch einmal auf die Darstellung von Wortbedeutungen und Wissen in Mono- und Polyhierarchien, Graphen, *Frames*, sog. semantischen Netzen oder auch in Texten der natürlichen Sprache zurück. Dass Monohierarchien meist unzureichend sind Wissen darzustellen liegt auf der Hand. Sie können nur in speziellen monokausalen Fällen Verwendung finden.

Betrachten wir beispielsweise eine bestimmte Person, die zum Arzt kommt:

Aus biologischer Sicht ist diese Person ein *Homo sapiens* und gehört in das Tierreich, weil für sie die Definition von Leben, von Tier, von Wirbeltier, etc. zutrifft.

Sobald wir diese Person etwas näher beschreiben wollen, müssen wir eine Reihe anderer Aspekte mit berücksichtigen. Dazu könnte die Anatomie, die Psychologie, die Ökonomie, die Soziologie, etc. zählen. Für jede dieser perspektivischen Betrachtungen ist eine eigene Hierarchie aufzustellen, in der diese Person auf ganz verschiedenen Ebenen der Polyhierarchie erscheinen.

Anstelle einer solchen, recht komplexen neuronalen Darstellung bietet sich die Nutzung von sogenannten *Frames* und *Slots* an, in denen die Person innerhalb eines Rahmens, mit Hilfe einzelner Ausschnitte beschrieben wird, wobei jeder Frame mit zahlreichen anderen *Frames* vernetzt werden kann. Dabei könnte man sich die *Frames* auch als in höchstem Maße vereinfachte Neuronen in einem neuronalen Netz vorstellen.

Wichtig bei der Beschreibung innerhalb eines Rahmens ist, dass die einzelnen Ausschnitte einer gewissen Systematik und Vereinheitlichung unterliegen, wie sie sich auch aus den Hierarchien ergeben. Die sog. *Slots* können auch als normierte Felder in einem Dokument gesehen werden, deren Inhalte aus den Hierarchien entnommen werden.

Beispiel:

Individuum A

Biologische Zugehörigkeit = *Homo sapiens*

Geschlecht = männlich

Alter (in Jahren) = 32

Größe (in cm) = 170

Gesundheit = krank

Befund = Bakterielle Infektion

Diagnoseergebnisse = ...

Ein solcher *Frame*, und diese Feststellung ist dokumentarisch nicht uninteressant, unterscheidet sich zunächst keinesfalls von einem herkömmlichen Dokument, wie wir es in Dokumentationssystemen zahlreich finden. Erst wenn die enthaltenen Informationen so verknüpft sind, dass sich aus ihnen begründete Entscheidungen gewinnen lassen (von Experten oder von einer Inferenzmaschine), wird aus einer klassischen Informationsbank eine Wissensbank. Wenn also beispielsweise Symptome der Krankheit eine bestimmte Diagnose untermauern, kann von Wissen gesprochen werden. Im Prinzip handelt es sich dabei um ein semantisches Netz, das uns die Möglichkeit gibt, einer Krankheit den richtigen Namen zu geben, so wie er sich aus dem Gesamtbefund ergibt.

Auch unsere natürliche Sprache erlaubt die Abbildung von Graphen, semantischen Netzen bzw. *Frames*. Sie ermöglicht Sätze, in denen die selbe Information enthalten sein kann, wie in den *Frames*.

Beispiel:

Das bezeichnete Individuum ist A. Ihr Geschlecht ist männlich. Das Alter beträgt 32 Jahre. etc.

Dabei sind alle Angaben zu einer Person in einem Dokument zusammenzufassen. Trotzdem tritt hier das Problem der natürlichen Sprache auf, einen Mangel an Präzision zu haben. Die Wissenschaft hat daher die natürliche Sprache durch eine erhebliche Zahl an Zusatzzeichen erweitert. Dies betrifft in erster Linie die Mathematik mit ihren Plus-, Minus-, Multiplikations-, Gleichheits- und anderen Zeichen. Dazu kommen logische Operatoren, wie UND, ODER, NICHT u.a. Auch die anderen Wissenschaftsdisziplinen haben feste Vereinbarungen entwickelt und teilweise sogar normiert, deren Verwendung eindeutig vorgeschrieben ist. Dabei ist an die Vielzahl der Programmierbefehle in der Informatik zu denken, die durchaus auch in diesen Bereich fallen. Wenn wir beispielsweise eine WENN-DANN-Schleife haben, die gerade bei Expertensystemen eine große Rolle spielt, so finden wir darin eine Festlegung, die uns bei genauer Betrachtung die Möglichkeit, kausale Bedingungen in Programmen explizit darzustellen.

Beispiel:

WENN Person A die Symptome der Krankheit B, C, D... aufweist, DANN ist sie zu X Prozent Wahrscheinlichkeit von Bakterium Z infiziert.

Auch die Dokumentation kennt in bescheidenem Umfang spezielle Zeichen wie UB = Unterbegriff, OB = Oberbegriff s. = Siehe Verweis oder auch die Einrückung in einer Tabelle zur Darstellung von Hierarchieebenen. Sie hat damit eine formalisierte Sprache, um Relationen auszudrücken.

Wir haben daher zunächst die Wahl, einen Thesaurus in der herkömmlichen dokumentarischen Struktur aufzubauen und darzustellen oder auch in Graphen, in *Frames* bzw. in anderen programmierbaren Strukturen. Wir können ihn allerdings ebenso in einer fachlich begründeten, aber durchaus natürlichen und verständlichen Sprache darstellen.

Unsere natürliche Sprache besticht zunächst durch ein Höchstmaß an Ökonomie, auf das schon Zipf (1949) in seinem Buch „*Human Behavior and the Principle of least Effort*“ hingewiesen hat. Diese Ökonomie wird in erster Linie durch die Syntax erreicht, durch die wir beispielsweise komplexe Vorgänge beschreiben, um ihnen dann einen einfachen Namen zu geben. Außerdem hat sie den großen Vorteil, dass das enthaltene Wissen in verständlicher Form in den Computer eingegeben werden kann, von ihm verarbeitet und in ebenso verständlicher Form wieder ausgegeben werden kann. Der allgemeine Trend in der Informatik, auch die Programmierung immer stärker einer verständlichen natürlichen Sprache anzupassen, in der kryptische Befehle immer weiter verschwinden, ist bereits seit längerem zu beobachten.

Schwarz, I. und Umstätter, W. (1998) haben bei einer früheren ISKO-Tagung über das Prinzip der Objektdarstellung mit Hilfe der SGML als einer natürlichen Sprache, die mit einer Metasprache dem Computer verständlich gemacht werden kann, berichtet. In diese Philosophie der SGML, in der wir Wörter, Sätze und ganze Dokumentteile bestimmten Kategorien zuordnen können, so dass beispielsweise Aussagen wie:

Beispiel:

Heute um 10.12 Uhr erschien Person A (32), männlich, alleinstehend mit den folgenden Symptomen, die auf eine bakterielle Infektionskrankheit schließen lassen.

Dieser Satz kann mit Markup-Zusätzen so versehen werden, dass über ein Volltextretrievalsystem problemlos alle wichtigen Angaben gezielt abgefragt und an eine Inferenzmaschine weitergeben werden könnten.

Beispiel:

Heute, am <Datum>9.9.1999 </Datum>um <Uhrzeit>10.12</Uhrzeit> Uhr erschien <Name>Person A</Name> <Alter>(32) </Alter>, <Geschlecht>männlich</Geschlecht>, <Familienstand>alleinstehend</Familienstand> mit den folgenden Symptomen, die auf eine <Vermuteter Befund>bakterielle Infektionskrankheit</Vermuteter Befund> schließen lassen.

Wenn wir diese Idee nun auf einen semiotischen Thesaurus anwenden wollen, so können wir in einzelnen *Frames* beliebige Benennung einsetzen, sie definieren, über *Scope Notes* die Randbedingungen und weitere Zusatzinformationen angeben, ihre sogenannte *History* kurz umreißen, sie zu Äquivalenzklassen zusammenschließen und sie mit anderen Benennungen in anderen *Frames* beliebig hierarchisch, assoziativ oder auch logisch vernetzen.

Wir gehen dabei von einer Systemgrundlage aus, die auf Axiomen basiert und aus denen wir den Rest des Gedankengebäudes logisch aufbauen. Ob diese Axiomatik zuverlässig und tragfähig ist, ergibt sich erst aus dem logischen Gesamtgebäude, das sich aus dieser Axiomatik durch die systemimmanenten Schlussfolgerungen extrapolieren lässt.

Es soll im folgenden kurz angerissen werden, wie man sich ein solches System vorstellen könnte. Ein in sich geschlossenes Konzept kann nur in einem aufwendigen gedanklichen Gesamtgebäude für bestimmte Problembereiche erzeugt werden.

Außerdem soll an dieser Stelle auch daran erinnert werden, dass es auf dieser Grundlage möglich wäre das Digitale Lehr und Handbuch (Umstätter, W. 1995) zu verfassen, dessen Gliederung sozusagen ein semiotischer Thesaurus ist, der schrittweise von verschiedenen Autoren vervollständigt und verbessert werden kann. Die dabei zugrundeliegende Struktur erfordert allerdings eine konsequente Logik und eine Kenntnis des bereits erarbeiteten, wobei die Dynamik des Systems die Beseitigung von Fehlern jederzeit zulässt.

Wählen wir als Beispiel das Wort Wissensorganisation. In Boolescher Logik setzt es sich aus Wissen UND Organisation zusammen. Es ist somit einerseits die Einschränkung des Wortes Wissen durch das Wort Organisation und andererseits als Unterbegriff von Organisation zu sehen, durch die Einschränkung, dass dann nur Wissen organisiert wird. Hierarchisch wäre es somit unter diesen beiden Aspekten einzuordnen.

Beispiel:

Wissensorganisation

Oberbegriff = Wissen

Oberbegriff = Organisation

Beide Oberbegriffe können entsprechend verlinkt werden.

Zur besseren Einordnung, und weil eine Objekthierarchie an Vererbung gebunden ist, wären damit zunächst die Begriffe Wissen und Organisation hinsichtlich ihrer Auswirkung auf die

Wissensorganisation zu prüfen. Verfolgen wir hier zunächst nur die Benennung Wissen, so müssten wir ein Thesauruselement über das Wort Wissen anlegen. Die darin enthaltenen Aussagen müssen dann in allen weiteren Schlüssen widerspruchsfrei Anwendung finden und sie dürfen grundsätzlich nicht mit den anderen logischen Folgerungen im System kollidieren. Eine solche Logik lässt sich erfahrungsgemäß am besten durch mathematische Gleichungen darstellen, die ihrerseits als Definitionen verstanden werden können, weil sie einen Begriff, durch die Relation zu anderen Begriffen festlegt. Damit wäre sofort zu prüfen, ob andere Begriffe, die von der Definition betroffen sind, dazu im Einklang stehen.

Beispiel:

Wissen

Definition: Wissen ist begründete Information

Oberbegriff: Information - da begründete Information eine Teilmenge von Information ist; außerdem Verlinkung mit dem Wort Information.

Oberbegriff: Begründung - da Handlungen, Informationen, Urteile, etc. begründet werden können; außerdem Verlinkung mit dem Wort Begründung.

Unterbegriff: Wissensorganisation - Verlinkung mit dem Wort Wissensorganisation.

Verwandter Begriff: Wissenschaft, im Sinne von methodischer Problemlösung - Verlinkung mit dem Wort Wissenschaft.

Information

Definition: Dieser Begriff bezieht sich auf die genauere Bezeichnung Mittlerer Informationsgehalt (H), der durch die Gleichung

$$H = - \sum p_i \times \log p_i,$$

definiert ist. Damit besteht eine direkte Beziehung zu der fundamentalen Größe der Entropie. In der Gleichung bedeutet p_i = Wahrscheinlichkeit, ein bestimmtes Zeichen aus einem vorgegebenen Zeichensatz gewählt zu haben.

Information erweist sich somit als ein rein wahrscheinlichkeitstheoretischer Begriff.

Diese Definition gilt grundsätzlich nur unter der Voraussetzung des Sender-Übertragungskanal-Empfänger-Modells, mit einem Zeichensatz der bei Sender und Empfänger identisch ist. Das Modell stammt, ebenso wie die dazugehörige Theorie von Shannon, C. und Weaver, W. aus dem Jahre 1949.

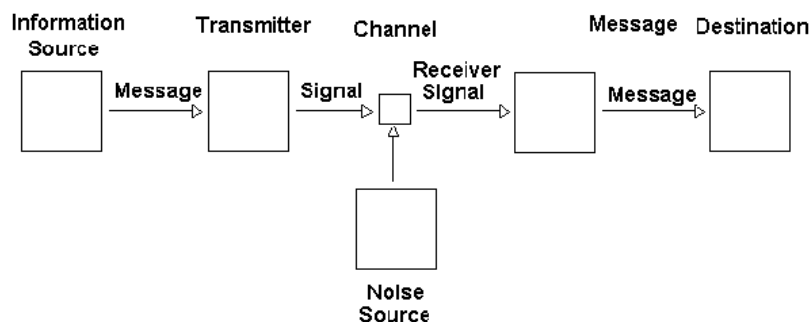


Abb. 1.: Informationsmodell von W. Weaver (S.7) und C. Shannon (S. 34)

Oberbegriff: Sein - da alles Sein von dem wir erfahren können nur über Information zu uns gelangen kann; außerdem Verlinkung mit dem Wort Sein.

Oberbegriff: Form - da Information ursprünglich als das einer Form innewohnende verstanden wurde; außerdem Verlinkung mit dem Wort Form.

Unterbegriff: Sender

Unterbegriff: Übertragungskanal

Unterbegriff: Empfänger

Unterbegriff: Wissen - Verlinkung mit dem Wort Wissen.

Quelle: Shannon, C and Weaver, W.: The Mathematical Theory of Communication
Illinois Books, edition (1963)

Mit Markup-Zusätzen erschiene im Prinzip:

<Benennung><Title>Information</Title></Benennung>
Dieser Begriff bezieht sich auf die genauere Bezeichnung <präzisierte Benennung> Mittlerer Informationsgehalt <präzisierte Benennung> (><Zeichen für präzisierte Benennung> H </Zeichen für Benennung>), der durch die Gleichung
<Definition> $H = - \sum p_i \times \text{ld } p_i$ </Definition>
definiert ist. Damit besteht eine direkte Beziehung zu der fundamentalen Größe der
<link>Entropie</link>.
In der Gleichung bedeutet <Abkürzung für Bedeutung von Var1> p_i </Abkürzung für Bedeutung von Var1> = <Bedeutung von Var1> Wahrscheinlichkeit ein bestimmtes Zeichen aus einem vorgegebenen Zeichensatz gewählt zu haben. </Bedeutung von Var1>
Information erweist sich somit als ein rein wahrscheinlichkeitstheoretischer Begriff.
Diese Definition gilt grundsätzlich nur unter der Voraussetzung des <WENN> <Bedingung><link>Sender-Übertragungskanal-Empfänger-Modells,</link></Bedingung>
<Wahr><UND> mit einem <Bedingung><link> Zeichensatz </link></Bedingung>
<Wahr> der bei Sender und Empfänger identisch ist <DANN><Begriff gültig>. <History>Das Modell stammt, ebenso wie die dazugehörige Theorie von <Zitat1>Shannon, C. und Weaver, W. </Zitat1> aus dem Jahre 1949</History>
. . .
<Quelle>Quelle:<Autor>Shannon, C and Weaver, W. </Autor>: <Titel>The Mathematical Theory of Communication </Titel> <Verlag>Illinois Books, </Verlag> edition (<Erscheinungsjahr>1963</Erscheinungsjahr>) <Quelle>

Damit wird unter anderem eine WENN-DANN-Schleife notwendig, die ihrerseits über die ursprüngliche SGML-Philosophie hinausgeht und die XML mit ihren Hyperlinks und den Möglichkeiten einer Programmierung erforderlich macht.

Das hier stark vereinfachte Beispiel soll drei Punkte deutlich machen:

1. Die sich zunehmender Beliebtheit erfreuende XML ist geeignet, um semiotische Thesauri in einer völlig neuen Konstruktion aufzubauen.
2. Sie führt dazu, dass der logische Zusammenhang der einzelnen Begriffe in einem Text sehr viel stringenter erfolgen kann als bisher.
3. Sie erlaubt es, auf Grund der Verbindung von Programmierungen und Recherchemöglichkeiten begründete Informationen aus solchen Thesauri zu gewinnen, die in ihrer Komplexität bisher nicht möglich waren.
4. Die Thesauri als solche stellen sich in ihrer Benutzeroberfläche mit Text, Grafik oder Ton multimedial und nutzerfreundlich dar, so dass die dahinter liegende Komplexität jederzeit im *Backtracking* abfragbar ist.

Auch wenn hier nicht verschwiegen werden soll, dass diese Idee noch nicht endgültig ausgereift ist, dass es noch erheblicher Entwicklungsarbeit bedarf, um die angesprochene Komplexität zu bewältigen, so soll die Mitteilung dieser Überlegungen dazu beitragen, ein Bewusstsein für die Möglichkeit zu schaffen, dass sich Wissen auch auf diesem Wege evolutionär strukturieren d.h. organisieren lässt, in dem es sich gegenseitig beeinflussend wächst. Jeder Thesaurusbegriff, seine Definition, seine Stellung im System und seine Präzision beeinflussen

das gesamte Gedankengebäude in einer abnehmenden interdisziplinären Fernwirkung. Dabei kann davon ausgegangen werden, dass wir zwar wie bisher unsere Gedanken und Überlegungen niederschreiben, die dann in der Welt 3, im Sinne Poppers, ihren Platz in den Bibliotheken dieser Welt finden. Neu ist aber, dass sich die Wissenschaftler dieser Welt in der Digitalen Bibliothek am Aufbau von Wissensbanken beteiligen können, wie sie in Form von Büchern und Zeitschriftenaufsätzen nicht möglich waren. Damit führen diese Überlegungen auch zu definitorischen Unterscheidungen und Präzisierungen, die bisher noch nicht in dieser Deutlichkeit auftraten.

Sich in diese Problematik hineinzudenken bedarf es zweifellos gewisser geistiger Anstrengungen und eines erheblichen zeitlichen Aufwandes, der sich aber lohnt, wenn man zu der Einsicht gelangt, dass die sogenannte *Big Science* nur auf diesem Wege wirtschaftlich betrieben werden kann, weil diese Form der Wissensproduktion und –organisation eine massiver Rationalisierung darstellt, so wie Bibliotheken schon immer eine organisatorische Rationalisierungsmaßnahme bei der Wissensproduktion darstellten, und die sogenannte Online-Dokumentation des letzten Jahrhunderts antrat, um überflüssige Doppelarbeit zu verhindern.

Literatur

- Morris, C.W.: Foundation of the theory of signs. Chicago. (1938)
- Peirce, C. S. (1967). Einige Konsequenzen aus vier Unvermögen. In C. S. Peirce: Schriften I: Zur Entstehung des Pragmatismus. Frankfurt am Main.
- Schwarz, I. und Umstätter, W.: Zum Prinzip der Objektdarstellung in SGML
In: Herausforderung an die Wissensorganisation: Visualisierung, multimediale Dokumente, Internetstrukturen.
Hrsg.:Czap, H.; Ohly, P. und Pribbenow, S., Ergon Verl., Würzburg. S.173-179 (1998)
- Schwarz, I. und Umstätter, W.: Die vernachlässigten Aspekte des Thesaurus: dokumentarische, pragmatische, semantische und syntaktische Einblicke.
Nachr. f. Dok. 50 (4) S. 197-203 (1999)
- Shannon, C and Weaver, W.: The Mathematical Theory of Communication
Illinois Books, edition (1963)
- Umstätter, W.: Nutzen der Indexierung bei Online-Datenbanken
14. Online- Tagung der DGD Proceedings, Frankfurt am Main DGD-Schrift (OLBG-13)
2/92 S.403-420 (1992)
- Umstätter, W.: Die Rolle der Dokumentation bei der Entstehung der Digitalen Bibliothek und ihre Konsequenzen für die Bibliothekswissenschaft.
Published in: Nachr. f. Dok. 46 (1) S.33-42 (1995)
- Umstätter, W.: Die Messung von Wissen
Nachr. f. Dok. 49 (4) S.221-224 (1998)
- Zipf, G.K.: Human Behavior and the Principle of least Effort. Reading, Mass. Addison-Wesley (1949)