Begriffssysteme

Ein Vergleich verschiedener Arten von Begriffssystemen und Entwurf des integrierenden Thema-Datenmodells

Version 1.0 (19.12.2003)

Studienarbeit im Diplomstudiengang Informatik
an der Humboldt Universität zu Berlin

von Jakob Voß

Betreuung: Rainer Eckstein

Zusammenfassung: Begriffssysteme tauchen als Thesauri, Klassifikationen, Nachschlagewerke, Begriffsnetze, Ontologien etc. in verschiedenen Fachwissenschaften zur Ordnung und Repräsentation von Wissen auf. Diese Arbeit gibt einen Überblick über die verschiedenen Arten von Begriffssystemen und ihre Bestandteile. Dies sind im wesentlichen Begriffe, Bezeichnungen und Relationen. Die häufigsten Bestandteile und ihre Anwendungsfälle werden erklärt, darunter die Terminologische Kontrolle, Begriffskombination, Notationen, die wichtigsten Relationsarten und Regeln sowie natürlich sprachliche Anteile. Es werden verschiedene Datenformate für Begriffssysteme vorgestellt. Die gemeinsamen Strukturen und Bestandteile verschiedener Begriffssysteme werden in dem gemeinsamen Datenmodell „Thema“ zusammengefasst, das anhand einer XML-Repräsentation des Modells erläutert wird.

Schlagworte: Begriff, Thesaurus, Klassifikation, Ontologie, Terminologie, Taxonomy, Nachschlagewerk, Semantic Web, XML, RDF

Vorwort

Sowohl die Informatik als auch die Bibliotheks- und Dokumentationswissenschaft beschäftigen sich unter anderem mit der Strukturierung von Informationen. Während die Informatik Methoden zu ihrer Speicherung und Verarbeitung bereitstellt und dabei mit einer bemerkenswerten Geschwindigkeit neue Techniken entwickelt, um mit der wachsenden Datenflut fertig zu werden, kann die Dokumentation eine langjährigen Erfahrung in der Ordnung und Erschließung großer Wissensmengen aufweisen. Mit Hilfe von nach dokumentarischen Regeln erstellten Katalogen ist in diesen Sammlungen – z. B. in Bibliotheken – auch eine Suche nach Themen und Inhalten möglich, während dies zum Beispiel im Internet nicht der Fall ist. Dies liegt zum einen daran, dass keine einheitlichen Metadaten für alle Dokumente vorliegen (ein kleiner Schritt in diese Richtung ist Dublin Core) und zum anderen, dass herkömmliche Suchmaschinen nur nach Zeichenketten und nicht nach Themen suchen.

Um eine „intelligentere“ Suche im Internet zu ermöglichen, gibt es seit einiger Zeit Bestrebungen, die Inhalte von Webseiten mit maschinenlesbaren Informationen über deren Bedeutung auszustatten. Durch Verknüpfungen soll so ein „semantisches Netzwerk“ entstehen, in dem sich intelligente Computerprogramme selbständig auf die Suche nach Informationen machen oder sogar im Auftrag Geschäfte abschließen können. Eine wichtige Grundlage spielen dabei die so genannten Ontologien – das sind Systeme von definierten Begriffen, mit deren Hilfe Vereinbarungen über die Bedeutung von einzelnen Inhalten getroffen werden können.

Solche Begriffssysteme gibt es jedoch bereits seit vielen Jahren innerhalb der Dokumentation, wo mit Schlagwortketten, Klassifikationen, Thesauri und anderen Dokumentationssprachen die Themen bzw. Inhalte von Dokumenten in einer formalen Sprache kodiert werden, so dass sie bei einer sachbezogenen Suche wiedergefunden werden können. Die dabei verwendeten Systeme sind zwar strukturell meist einfacher als Ontologien und in ihrer technischen Umsetzung nicht auf dem neuesten Stand; sie werden aber dafür umfangreich und mit langjähriger Erfahrung in der Praxis eingesetzt, während Ontologien bisher eher Forschungsgegenstand sind oder in Bereichen eingesetzt werden, wo sie nicht über herkömmliche Techniken hinausgehen.

Den Anstoß für diese Arbeit gab die Erfahrung als Student beider Disziplinen – der Information und der Bibliotheks- und Dokumentationswissenschaft – dass trotz oft gemeinsamer Ziele viel zu wenig Zusammenarbeit gibt und Entwicklungen nebeneinander vorbei laufen. Während in der Informatik die theoretischen Grundlagen der Dokumentation so gut wie gar nicht bekannt sind, anstatt dass auf diese aufgebaut wird, fehlt es in der Bibliothekswissenschaft oft an der zeitgemäßen Umsetzung und Weiterentwicklung erprobter Methoden.

Ich habe es mir mit dieser Arbeit zum Ziel gesetzt, zunächst einmal die theoretischen Grundlagen aufzuarbeiten und verschiedene Arten von Begriffssystemen nebeneinander zu stellen, um ihre gemeinsamen Bestandteile herauszuarbeiten. Bei der Suche nach weiteren Fachwissenschaften, die sich mit der Modellierung einzelner Denkeinheiten beschäftigen, habe ich den Eindruck bekommen, dass auch dort eher ein Neben- als ein interdisziplinäres Miteinander festzustellen ist. Dieser Umstand ist sicherlich nicht gänzlich vermeidbar, da er in der Natur des Wissenschaftsprozesses und der dort fortwährend stattfindenden Spezialisierung begründet liegt; Entwicklungen wie die des Semantic Web und der Dokumentation sollen jedoch gerade über diesen Umstand hinweghelfen und es ermöglichen, dass verschiedene Personen (und Computer) trotz unterschiedlicher Sprachen feststellen können, wann Sie sich über gleiche Dinge unterhalten und wann nicht.

Ich hoffe, mit meiner Arbeit einen kleinen Beitrag in diese Richtung geben zu können.

Inhaltsverzeichnis

1. Einleitung

Begriffssysteme gibt es in vielen Bereichen. In ihrem weitesten Sinne existieren sie seit Menschen damit begannen, ihre Vorstellungen in schriftliche Zeichen zu übertragen, damit andere Menschen sie wieder entziffern und in vergleichbare Vorstellungen umsetzen können. Bereits vorher – in der menschlichen Sprache – kommunizieren wir mit Begriffen. Da ein Gespräch jedoch die Möglichkeit einer direkten Rückfrage gibt und fast jede Aussage in den Kontext eines Gespräches eingebunden ist, müssen Begriffe dort weit weniger konkret definiert sein. Noch genauer müssen Begriffe für die Verarbeitung mit Computern gefasst sein, die es erstmals ermöglicht, Begriffe nicht nur in statischen, sondern in dynamischen Systemen abzulegen und darin automatisch verarbeiten zu lassen. Vor allem in diesem Bereich haben in den letzten Jahren bestimmte Arten von Begriffssystemen unter der Bezeichnung *Ontologien* und im Rahmen der Idee des *Semantic Web* einen starken Aufschwung erfahren und sind in der Informatik Thema der aktuellen Forschung. Aber auch andere Wissenschaften beschäftigen sich bereits länger mit verschiedenen Arten von Begriffssystemen.

Beispielsweise sind in der Dokumentationswissenschaft seit Jahrzehnten Klassifikationen und Thesauri zur Erschließung der Inhalte von Dokumenten bekannt, wo eindeutig definierte Begriffe als Schlagworte und vollständige Dokumenten miteinander in Beziehung gesetzt werden. Auch Enzyklopädien und Handbücher beinhalten einzelne Themen und Begriffe, die in ihnen erklärt und definiert werden. All diese Systeme, die in dieser Arbeit hinsichtlich ihrer Form (2.1) und Struktur (2.2) untersucht werden sollen, haben die gleiche Grundlage – nämlich begrifflich geordneten Vorstellungen (bzw. Vorstellungen, die mit einem Begriffssystem geordnet werden sollen) im menschlichen Geist und Sprache. Es ist deshalb anzunehmen und festzustellen, dass die unterschiedlichen, sich mit Begriffssystemen beschäftigenden Wissenschaften bei der Untersuchung und praktischen Anwendung von Begriffssystemen voneinander profitieren können. Doch zunächst einmal:

1.1. Was sind Begriffssysteme?

Um mit einer Definition zu beginnen: *Begriffssysteme* (auch *Konzeptsysteme*) sind Systeme von unterscheidbaren Konzepten, die mittels Relationen in Beziehung zueinander gesetzt werden und in einer natürlichen, visuellen und/oder formalen Sprache formuliert werden können.

Es sei angemerkt, dass diese Definition keinen Anspruch auf Allgemeingültigkeit hat – auch ist „Begriffssystem“ kein wissenschaftlich anerkannter Fachbegriff. Stattdessen werden jeweils unterschiedliche der hier dargelegten verwandten Systeme von Begriffen als Begriffs- oder Konzeptsystem bezeichnet. Häufiger sind jedoch speziellere Ausprägungen, die auch genauere Bezeichnungen besitzen. Neben Begriffssystemen lassen sich Begriffsrepräsentationssysteme feststellen, das sind Repräsentationsformalismen wie die menschliche Sprache, Logik- und Programmiersprachen oder Datenbanken, auf die im zweiten Teil dieser Arbeit eingegangen wird (Kapitel 3). Als Ergebnis wird mit dem *Thema-Datenmodell* ein integrierendes, XML-basiertes Datenmodell und -format vorgestellt, das die verschiedenen hier dargestellten Bestandteile und Konzepte in sich vereinigt.

Trotz aller Gemeinsamkeiten lassen sich bei Begriffssystemen verschiedene in der Praxis vorkommende Arten unterscheiden. Die wesentlichen Arten werden deshalb in Kapitel 2.1 vorgestellt. Eine grobe Einteilung kann in die Bereiche *Nachschlagewerke*, Systeme zur *Dokumentation* und *Datenbanken* vorgenommen werden. In der Informatik werden die meisten Begriffssysteme unter anderem als Ontologie, Datenbankschema oder Objekthierarchie bezeichnet.

1.2. Begriffssysteme in verschiedenen Fachwissenschaften

Begriffssysteme spielen in vielen Bereichen und Fachgebieten eine wichtige Rolle. Aus diesem Grund beschäftigen sich unterschiedliche Fachwissenschaften aus unterschiedlichen Blickrichtungen mit verschiedenen Arten von Begriffssystemen. Da der interdisziplinäre Blick über den Tellerrand des eigenen Faches schwierig ist, sind Überschneidungen keine Seltenheit, so dass die Urheberschaft über ähnliche Ideen und Systeme gleich mehrfach reklamiert werden kann.

Unabhängig davon besitzt *jede* Wissenschaft einen gewissen Konsens über die Bedeutung von Teilen ihres speziellen Fachvokabulars, ohne den ein Austausch von Wissen unmöglich ist. Mit solchen *Terminologien* (siehe 2.1.3) besitzen also alle Wissenschaften ihre Begriffssysteme. Diese können sowohl explizit über (zum Beispiel als Fachwörterbücher oder Klassifikationen) oder in einem Fachgebiet auftreten (zum Beispiel die Taxonomie der Lebewesen in der Biologie) als auch implizit als unausgesprochene *Paradigmen* und Grundannahmen (siehe [Kuhn] und 2.1.1).

Eine interdisziplinäre Terminologie *zum Thema Begriffssysteme* existiert bislang nicht und wäre auch lediglich auf einem sehr groben Niveau möglich – wohl aber gibt es sie in einzelnen Disziplinen (meist in Form von Wörterbüchern und Glossaren). Zwei Terminologien aus dem Dokumentationsbereich, die bereits Thesaurusstrukturen besitzen, enthalten [Neveling] und [Umstätter].

Eine kurze – zugegebenermaßen nicht sehr wissenschaftliche – Recherche mittels einer einfachen Suchanfrage bei Google[1] ergab folgende Trefferanzahlen zur Verwendung einiger Bezeichnungen für Begriffssysteme (jeweils einmal „Suche Seiten in Deutsch“ und einmal „im ganzen Web“):

	Deutsch	***Web***
Konzeptsysteme	24	26
Begriffssysteme	1.020	1.080
Terminologien	3.830	8.850
Ontologien	4.530	5.060
Thesauri	7.720	730.000
Klassifikationen	30.500	41.400

Schaut man sich die indexierten Seiten für die schon sehr spezielle Bezeichnung „Konzeptsystem“ an, so findet man Beiträge aus der Informatik, (Medizinischen) Dokumentation, Lexikographie, Kognitionswissenschaft, Betriebswirtschaft, Sprachwissenschaft, Psychologie, Literaturwissenschaft und Kulturwissenschaft. Unter „Begriffssysteme“ findet man vergleichbare Beiträge, auch aus den Bereichen Terminologielehre, Mathematik, Wissensmanagement und Philosophie.

Insgesamt lässt sich die theoretische Beschäftigung am besten in den Bereich der Informationswissenschaft einordnen, die auch als Oberbegriff der hier hier aufgeführten Wissenschaften verstanden werden kann – aber auch die kognitionswissenschaftlichen Disziplinen spielen eine Rolle.

Wichtige Institutionen sind neben vielen einzelnen Fachorganisationen unter anderem das *World Wide Web Consortium* (W3C) für den Bereich der Informatik, die *Internationale Gesellschaft für Wissensorganisation* (ISKO) aus dem Bereich der Terminologie- und Informationswissenschaft sowie verschiedene Normungsgremien, z. B. für den Bereich der Bibliothekswissenschaft.

1 Am 14.6.2003 unter http://www.google.de

2. Theoretischer Hintergrund

Um Begriffssysteme erkennen, beurteilen und voneinander abgrenzen zu können, ist zunächst eine theoretische Betrachtung notwendig. Dies ist um so wichtiger, da Projektbeschreibungen und Lehrbücher in der Regel in einem Fachgebiet verhaftet sind, und nur auf einzelne Aspekte eingehen. Im Folgenden werden zunächst die wichtigsten Arten von Begriffssystemen dargestellt und danach die wesentlichen in ihnen enthaltenen Bestandteile analysiert. Die dabei entwickelten Typologien stellen ausdrücklich eine pragmatische Einteilung dar; zum einen lassen sich Begriffssysteme sicher nach speziellen Gesichtspunkten auch anders aufteilen und zum anderen sind die Grenzen zwischen den Systemen stets fließend. Gerade deshalb ist aber eine Unterscheidung der prinzipiellen Möglichkeiten sinnvoll, denn in der Praxis herrscht (auch durch den unterschiedlichen Sprachgebrauch) oft einiges an Unklarheit –. beispielsweise in der Unterscheidung zwischen Wörterbuch und enzyklopädischen Nachschlagewerken oder zwischen Klassifikation und Thesaurus.

Nach der Betrachtung der Arten (2.1) und Bestandteile (2.2) wird noch kurz auf mögliche Arten der Darstellung von Begriffssystemen eingegangen (2.3) eingegangen.

2.1. Arten von Begriffs-, Konzept- und Ordnungssystemen

Im Folgenden sollen verschiedene Arten von Begriffs- Konzept- und Ordnungssystemen dargestellt werden, die sich in der Praxis antreffen lassen. Die Typologie beruht im Wesentlichen auf den unterschiedlichen Gebieten, in denen die Systeme verwendet werden. Hinsichtlich ihres Einsatzzweckes und ihrer Art lassen sich Begriffssysteme in drei Bereiche einteilen (Bild 1).

Die einfachste Form sind *Nachschlagewerke* (2.1.2), die meist sprachliche Definitionen und Erklärungen zu einzelnen Begriffen enthalten. Die *Dokumentation* benutzt verschiedene Arten von Ordnungs- und Begriffssystemen (2.1.4 ff.), die vorrangig zum Beschreiben und Wiederauffinden von anderen Objekten dienen. *Datenbanken* und formale Beschreibungen (2.1.10 f.) können komplexere Strukturen besitzen, erfassen größere Datenmengen und ermöglichen Speicherung und automatische Verarbeitung.

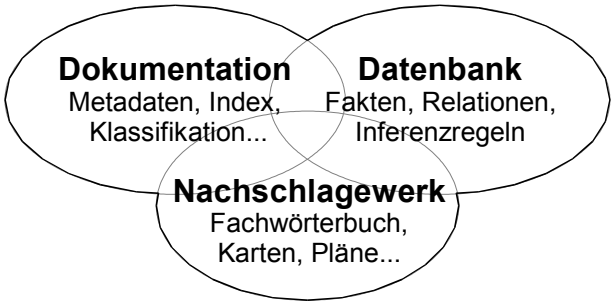


Bild 1: Einssatzzwecke und Arten von Begriffssystemen

Zwischen allen drei Bereichen gibt es Überschneidungen und allen dreien liegen (zum Teil philosophische) Konzeptsysteme als Vorstellung in den Köpfen von Menschen zugrunde.

2.1.1. Philosophische Begriffssysteme

Die Entlehnung des Begriffes *Ontologie* für formale Begriffssysteme innerhalb der Informatik (2.1.10) lässt zu recht vermuten, dass Begriffssysteme etwas mit Philosophie zu tun haben. Auch die unter 2.2.1 und 2.2.2 beschriebene Unterscheidung von Begriff und Benennung hat ihre sprach- und erkenntnistheoretischen Wurzeln in der Philosophie. Die Ontologie ist innerhalb der theoretischen Philosophie die Wissenschaft, die sich mit dem Sein beschäftigt. Obwohl die Ontologie nur ein Teilgebiet der Metaphysik ist, wird sie oft mit dieser gleichgesetzt. Der Begriff Ontologie kann auch für eine spezielle ontologische Theorie, d. h. für eine spezielle Annahme über das Wesen des Seins stehen, was direkt zum technischen Begriff der Ontologie in der Informatik führt. Jede Wissenschaft (ja selbst der Begriff der Wissenschaft selber) beruht auf einer Vorstellung darüber, wie das Universum aufgebaut ist und worüber daher zu kommunizieren Sinn macht. Die von den meisten

Menschen vertretene ontologische Position ist ein positivistischer "naiver Realismus", nach dem die Realität als real und erkennbar angenommen wird. Nach der skeptischen Position des Konstruktivismus wird die gesamte Welt inklusive aller Bedeutungen erst vom Menschen konstruiert (die konstruktivistische Positionen ist jedoch nicht, wie häufig angenommen wird, mit einer positivistischen Weltsicht unvereinbar).

Für den praktischen Umgang mit Ontologien und anderen Begriffssystemen sind philosophisch betrachtet zwei Tatsachen (sic!) interessant: Zum einen schafft jede Ontologie (Weltsicht) ihre eigene Welt. Jedes noch so fein ausgearbeitete Begriffssystem wird Lücken und subjektive Festsetzungen aufweisen. Dies zu ignorieren wäre genau so töricht wie die Meinung, das Lesen einer Enzyklopädie genüge, um die Welt zu verstehen. Zum anderen beruht jede Ontologie auf impliziten Annahme über die Welt, in der sie eingebunden ist und die nicht in ihr enthalten sein können. Diese impliziten Annahmen führen oft dazu, dass ein Begriffssystem von verschiedenen Menschen unterschiedlich interpretiert wird. Die Unmöglichkeit einer vollständigen und endgültigen Beschreibung der Welt, und damit die Kontextgebundenheit jeder Ontologie, kann auch philosophisch begründet werden. Von Außen mag es möglicherweise so erscheinen, als seien in der langen Geschichte der Philosophie bereits alle metaphysischen Probleme gelöst oder als unlösbar erledigt. Innerhalb der wissenschaftlichen Philosophie geben die Bereiche Metaphysik und Ontologie jedoch immer wieder Anstoß für Fragestellungen und bilden ein lebendiges Forschungsgebiet, wie z.B. die Gründungen der internationalen Zeitschrift *Methaphysica* im Jahre 2000 belegt.

Neben der Ontologie sind für die Betrachtung von Begriffssystemen mehrere weitere philosophische Disziplinen relevant. Dazu gehören fast alle Teilgebiete der Theoretischen Philosophie (Logik, Erkenntnistheorie und Wissenschaftstheorie) sowie die Sprachphilosophie und, sobald es um die praktische Anwendung von Begriffssysteme geht, auch die Ethik. Vor allem dort, wo vorgefertigte Systeme die Realität nicht adäquat wiedergeben (können) oder verschiedene Ansichten aufeinander stoßen, berührt dies fast immer auch ethische Belange (siehe dazu [Bowker]).

2.1.2. Nachschlagewerke

Nachschlagewerke sind verschiedenartige zur Fachliteratur zählende Sammlungen von Informationseinheiten, die nach bestimmten systematischen und formal Kriterien aufgestellt sind und sich durch ihre Struktur leicht nach bestimmten Teilen durchsuchen lassen. Trotz unterschiedlicher Ausprägungen der in einem Nachschlagewerk enthaltenen Informationseinheiten (Artikel, Definitionen, Wörter, Personen, Daten, Zeichnungen etc.) sind diese Begriffe – jedes Nachschlagewerk ist also ein Begriffssystem. Der Schwerpunkt eines Nachschlagewerks, im Vergleich zu anderen Begriffssystemen, liegt in seinem Zweck als Hilfsmittel zum direkten Auffinden von Informationen zum Wissenserwerb und in der menschlichen Zielgruppe. Somit lassen sich alle hier aufgeführten Begriffssysteme auch als Nachschlagewerk bezeichnen. Datenbanken, die primär der Sammlung und Weiterverarbeitung von Daten dienen, sind dagegen nicht unbedingt Nachschlagewerk, obwohl sie auch als solche fungieren können.

Da sich Nachschlagewerke an Menschen richten, bedienen sie sich fast immer der Sprache. Auch andere Medien wie Bilder und Zeichnungen oder Filme sind möglich. Diese können sowohl illustrativ als auch als konstitutive Elemente eines Nachschlagewerks auftreten (z. B. in Landkarten).

Einige Beispiele für Nachschlagewerke sind Enzyklopädien, Lexika und Wörterbücher, bibliographische Verzeichnisse, Kataloge, Telefonbücher, Ablaufdiagramme und Übersichtspläne.

Für die Digitalisierung von Enzyklopädien und Wörterbüchern wird schon seit einigen Jahren SGML oder XML verwendet. Das vorherrschende Format (zumindest für die Edition bereits gedruckt vorliegender Werke) ist in diesem Bereich die *Text Encoding Initiative* (TEI).

Enzyklopädien und Lexika

Bereits in der Antike und im Mittelalter (dort besonders im arabischen und chinesischen Raum) gab es umfangreiche systematische Wissenssammlungen. Mit der Aufklärung kommt der Begriff einer *Enzyklopädie* als Darstellung der Gesamtheit des menschlichen Wissens auf. Der Prototyp einer solchen Enzyklopädie ist die von 1751 bis 1772 in 18 Bänden erschienene »*Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres*« der so genannten Enzyklopädisten unter der Leitung von JEAN LE ROND D'ALEMBERT und DENIS DIDEROT.[2] Die einzelnen Artikel sind mittels Siglen, die zu Anfang des einzelnen Artikels die übergeordnete Wissenschaft anzeigen, und durch ein Verweisungssystem miteinander verbunden.

Mit steigendem Umfang von Enzyklopädien wurde ihre systematische Ordnung durch eine alphabetische Ordnung abgelöst. Die erweiterte Möglichkeit der multidimensionalen Strukturierung von enzyklopädischen Hypertexten hat sich noch nicht durchgesetzt, da die Erstellung von elektronischen Ausgaben herkömmlicher Enzyklopädien für die Verlage nicht finanzierbar ist. Angesichts der exponentiell wachsenden Informationsmenge – die Anzahl (wissenschaftlicher) Publikationen verdoppelt sich rund alle 15 Jahre – können universelle Enzyklopädien (wie der Brockhaus oder die Encyclopaedia Britannica) nur einen Überblick geben und auf weitere Literatur verweisen. Zur Spezialisierung gehört auch, dass zu ausgewählten Themen und Fachgebieten Fach- und Speziallexika erscheinen. Mit dem Internet ist die Idee einer 'Bibliotheca Universalis' wieder in scheinbar greifbare Nähe gerückt[3] – sei es in Form eines Nachschlagewerkes für Menschen (z. B. [Wikipedia]) oder für Maschinen (Semantic Web).

Die Bezeichnung „Lexikon" wird übrigens sowohl für Enzyklopädische Nachschlagewerke als auch für Wörterbücher verwendet. Obwohl es auch Zwischenformen gibt, sind jedoch beide klar zu trennen, da Enzyklopädien Begriffe (Konzepte) und Wörterbücher Wörter enthalten.

Wörterbücher

Ein Wörterbuch ist ein Nachschlagewerk mit einer meist alphabetisch geordneten Sammlung von lexikalischen Einheiten (Wörter, Phrasen, Morpheme), die sprachliche Informationen wie z.B. die Definition eines Wortes, den Hinweis auf synonyme Wörter oder die Entsprechung in einer Fremdsprache enthält. Bekannte deutsche Wörterbücher sind das „Grimmsche Wörterbuch" und der „Duden". Je nach Anwendung kann es viele unterschiedliche Wörterbücher geben (Rechtschreibwörterbücher, Übersetzungswörterbücher, Fachwörterbücher, Etymologische Wörterbücher, Reimwörterbücher, Abkürzungsverzeichnisse, Namensverzeichnisse, Konkordanzen u.v.a.m.).

Mit der Herstellung von Wörterbüchern beschäftigt sich die *Lexikographie* (Wörterbuchkunde). Sie greift dabei unter anderem auf Ergebnisse der linguistischen Disziplin der *Lexikologie* zurück, die die Beziehungen und Regeln zwischen den einzelnen lexikalischen Bestandteilen (Morpheme, Wörtern und feste Wortgruppen) einer oder mehrerer Sprachen untersucht. Eine gute und aufschlussreiche Einführung in die Wörterbuchkunde bietet [Engelbert].

2.1.3. Terminologien

Als *Terminologie* bezeichnet man im allgemeinen die Gesamtheit aller Begriffe und Benennungen (*Termini*) einer Fachsprache bzw. die Fachsprache selbst. Im Gegensatz zu allgemeinen Wörterbüchern besitzen die Begriffe einer Terminologie in der Regel eindeutige Benennungen . Terminologien können beispielsweise in Wörterbüchern, Schlagwortlisten, Glossaren, Thesauri oder

2 Inzwischen verfügbar unter http://encyclopedie.inalf.fr/

3 Peter Haber: *Der wiedererwachte Traum von der «Bibliotheca Universalis». Das totale Wissen im digitalen Zeitalter.* In: Neue Züricher Zeitung vom 24.1.2000, S. 25-26

terminologischen Datenbanken abgelegt sein. Sie können deskriptiv oder normativ erstellt werden Wissenschaftliche Terminologien unterstehen – zumindest in lebenden Fachgebieten – einem beständigen Wandel durch die Neuinterpretierung alter und Einführung neuer Termini. Für etablierte gesicherte Erkenntnisse und andere Zwecke gibt es auch normativ geregelte Terminologien, z. B. die Namensgebung der Chemischen Elemente und Verbindungen (Nomenklatur).

Die Entwicklung einheitlicher Terminologien ist meist die Aufgabe von internationalen Fachorganisationen und Normungsgremien. Formale Richtlinien für Terminologie sind u.a. in verschiedenen DIN und ISO-Standards festgelegt. Das verantwortliche Gremium innerhalb der ISO ist das ISO/TC37: „Terminology and other language resources". Kurzdarstellungen und Links zu dieser und anderen Institutionen, u.a. der „Gesellschaft für Terminologie und Wissenstransfer", gibt es auf der Seite des *International Information Centre for Terminology* in Wien (http://linux.infoterm.org/).

Ein wichtiges Anwendungsgebiet von Terminologien ist die Übersetzungsarbeit. Vor allem technische Übersetzungen, sind ohne Terminologiemanagement nicht möglich. Für die Erstellung von Normen, Anleitungen, Handbücher und Benutzerschnittstellen von Softwareprogrammen in verschieden Sprachen muss klar geregelt sein, welcher Begriff in welcher Sprache wie bezeichnet wird. In der Informatik ist dieses Problem als Lokalisierung (L10N) und Internationalisierung (I18N) bekannt. Microsoft, dessen Textverarbeitung Word in mittlerweile 35 Sprachversionen erhältlich ist, unterhält zu diesem Zweck eigens terminologische Datenbanken, die für die Übersetzer bindend sind.[4] Wie schwierig sich einmal etablierte Terminologien ändern lassen, zeigt beispielsweise die Umbenennung von „Verzeichnis" in „Ordner". Auch in der Computerlinguistik – z. B. zur automatischen Übersetzung – spielen Terminologien eine wesentliche Rolle. Daneben beschäftigt sich die Terminologiewissenschaft mit allgemeinen Fragen der Wissenschaftstheorie und Wissensstrukturierung und -organisation. Als Verwandte Disziplinen sind neben den Sprachwissenschaften undder Computerlinguistik die Sprachphilosophie von Interesse.

Das Aufstellen einer übergreifenden wissenschaftlichen Terminologie ist kompliziert, da Begriffe nicht nur in verschiedenen Fachwissenschaften sondern mitunter auch innerhalb einer Disziplin je nach Wissenschaftler unterschiedlich definiert und benannt werden. Dies hängt zum Teil mit dem Wesen von Wissenschaft selbst zusammen, in der ständig neue Begriff gebildet werden und erst mit der Zeit zu eindeutig benennbaren Konzepten reifen. Es muss aber auch zugegeben werden, dass zusätzlich persönliche Interessen und Abneigungen eine Rolle spielen, denn kein Wissenschaftler gibt gerne die auch mit Einfluss verbundene „Definitionshoheit" über ein Gebiet ab.

2.1.4. Metadaten und Kataloge

Unter Metadaten oder Metainformationen versteht man *Daten über Daten*, das heisst Angaben, die sich auf den Inhalt, Umfang, Ursprung oder andere Eigenschaften von andere Daten beziehen. Ein klassisches Beispiel sind die Angabe von Autor und Titel eines Buches. Obwohl die Bezeichnung „Metaten" noch relativ neu ist (die Marke Metadata® wurde noch 1991 als unbedenklich eingestuft), ist das Konzept seit der Antike aus dem Bibliothekswesen bekannt. In der der Bibliothek von Alexandria entwickelte im 3. Jh. v. Chr. KALLIMACHOS mit den so genannten Pinakes den ersten systematischen Bibliothekskatalog. In 120 Schriftrollen verzeichnete er zu den wesentlichen Werken Autor, Titel, Entstehungszeitraum und Zeilenumfang und die einleitenden Sätze. Die einzelnen Titel wurden nach wissenschaftlichen und literarischen Kategorien gemäss aristotelischem Vorbild geordnet und die Schriftrollen mit Etiketten versehen, mittels derer man die sie identifizieren konnte.

Der Begriff Metadaten wird vor allem für Informationen über Dokumente und Publikationen benutzt, während beispielsweise solche Angaben wie der Name und die Größe einer Datei eher

4 Siehe DIETER E. ZIMMER: *Die Multikulturmaschine*. In: Die Zeit vom 29.4.1999, Seite 37-38

„Attribute“ oder schlicht „Eigenschaften“ genannt werden. Metadaten lassen sich in einigen Fällen aus den zu beschreibenden Daten extrahieren (z. B. die Dateigröße durch Zählen von Bytes) in vielen Fällen ist dies jedoch nicht oder nur mit intellektuellem Aufwand möglich.

Formal- und Sacherschließung

In der bibliothekarischen Katalogisierung wird bei der manuellen Erzeugung von Metadaten zwischen Formalerschließung und Sacherschließung unterschieden. Die *Formalerschließung* beinhaltet objektive Informationen (Format, Autor, Erscheinungsjahr…), wahrend sich die *Sacherschließung* auf den Inhalt einer Publikation bezieht. Methoden der Sacherschließung sind die Formulierung von Zusammenfassungen (Referieren), die Vergabe von Stich- oder Schlagwörtern (Indexieren bzw. Beschlagwortung) oder die Einordnung in eine Klassifikation (siehe 2.1.5). Obwohl die Sacherschließung im Vergleich zu automatischen Verfahren (z. B. Indexierung zur Volltextsuche) teurer und aufgrund ihrer Subjektivität nie 100%ig reproduzierbar ist, dient sie als wesentlicher Bestandteil von Wissensmanagement. Im Ggs. zur automatischen Indexierung lassen sich bei der intellektuellen Zuordnung gezielt die wesentlichen Schlagwörtern oder Kategorien angeben und damit bei der Recherche nicht nur eine höhere Treffergenauigkeit (*precision*), sondern auch ein bessere Trefferquote (*recall*) erzielen. Die Verwendung eines kontrollierten Vokabulares stellt sicher, dass gleichen Themen mit gleichen Schlagworten verzeichnet werden. Ein Beispiel ist die Vergabe von englischen Keywords, mit denen sich auch fremdsprachige Texte finden lassen. Einfache Schlagwortlisten oder Indexe sind der Erste Schritt zum Erstellen eines Begriffssystems. Das größte deutschsprachige Schlagwortsystem ist die Schlagwortnormdatei (SWD) mit rund 1.300.000 Einträgen (davon 600.000 Deskriptoren), die mit etwa 140.000 Relationen verknüpft sind (Stand 2003). Sie wird von der Deutschen Bibliothek in Kooperation mit verschiedenen Bibliotheksverbünden verwaltet.

Metadatenformate

Zur Speicherung und zum Austausch von Metadaten müssen diese einem einheitlichen Format entsprechen. In Bibliotheken bestehen dazu seit Jahrhunderten Regeln, die verschiedene Aspekte vom Umfang der aufzunehmenden Metadaten über deren Ansetzung (Schreibweise, z. B. bei Namen) bis zu ihrer konkreten Speicherung in Form von Katalogkarten oder Datensätzen festlegen. Die Regeln werden in umfangreichen Regelwerken festgehalten und ständig aktualisiert. In Deutschland sind dies die *Regeln für die Alphabetische Katalogisierung* (RAK) und vor allem im englischsprachigen Raum die *Anglo-American cataloguing rules* (AACR). Zur Zeit gibt es die Diskussion über einen Umstieg von RAK auf AACR. Das in Deutschland verwendete Datenformat ist das *Maschinelle Austauschformat für Bibliotheken* (MAB2); internationaler Standard ist das von der Library of Congress verwaltete Format MARC21 (*MAchine-Readable Cataloging*),[5] das auch in einer XML-Repräsentation verfügbar ist. Eine Übersicht bibliothekarischer Datenformate gibt [Eversberg].[6]

Dem Nicht-Bibliothekar sind eher andere Formate wie BibTeX und Literaturverwaltungsprogramme wie EndNote und ProCite bekannt. Als kleinster gemeinsamer Nenner für den Austausch von Metadaten (nicht nur über Publikationen) hat sich der Dublin Core Standard durchgesetzt. In seiner einfachen Version als *Dublic Core Metadata Element Set* besteht es aus einer einfachen Terminologie aus 15 Datenfeldern wie Titel, Autor, Datum, Indentifikator etc., die alle mehrfach vorkommen können und sich bei Bedarf durch Schemata und Qualifiers [Kokkelink] genauer spezifizieren lassen. Die Verwendung von Dublin Core ist jedoch uneinheitlich, da es kein dazugehöriges Regelwerk gibt.

5 *Library of Congress Network Developement and MARC Standard Office*: http://lcweb.loc.gov/marc/
6 Siehe auch die gesamten *Material zu Katalogen und Datenbanken*: http://www.allegro-c.de/formate/

2.1.5. Klassifikation

Eine Klassifikation oder Systematik ist eine planmäßige Darstellung von Klassen, Kategorien oder anderen Konzepten, welche nach bestimmten Ordnungsprinzipien gestaltet ist. Die einzelnen Klassen werden in der Regel durch den Vorgang der Klassifikation, das heißt durch die Einteilungen von Objekten anhand bestimmter Merkmale, gewonnen und hierarchisch angeordnet. Klassifikationen werden zur Dokumentation und in der Wissenschaft (dort spricht man eher von Systematik) verwendet. Ziel einer Klassifikation ist es, einen Überblick über die darin geordneten Objekte zu verschaffen (Analyse) und die thematische Suche unter ihnen zu ermöglichen (Ordnung).

Beispiele für Klassifikationen sind die Biologische Systematik des SYSTEMA NATURAE von CARL VON LINNÉ, die Internationale Klassifikation der Krankheiten (ICD) und verschiedene Bibliotheksklassifikationen. In vielen Klassifikationen werden die einzelnen Klassen über eine eindeutige Notation bezeichnet. Die Identifikation der in einer Klassifikation abgelegten Objekte kann durch eine Signatur geschehen, die ggf. Teile einer Notation enthält (siehe 2.2.7 und das folgende Beispiel).

Beispiel für einen Eintrag einer Klassifikation (mit Notation)

In der Regensburger Verbundklassifikation (RVK)[7] gibt es die Klasse der mit der Notation `NU 3025` für die Geschichte der Humboldt Universität zu Berlin. Die dazu gehörende Klasseneinteilung ist:

N Geschichte

NU Geschichte der Wissenschaften und des Unterrichtswesens

NU 1500-7950 Geschichte der Wissenschaften

NU 2500-4250 Geschichte der wissenschaftlichen Institutionen

NU 2500-4215 Universitäten und Hochschulen

NU 3000-3329 Deutschsprachige Universitäten

NU 3025 Berlin/Humboldt Universität

Jedes Werk in einer Bibliothek erhält (mindestens) eine Notation. Die Regalaufstellung kann somit entsprechend der Einteilung der Klassifikation vorgenommen werden (*Systemtische Aufstellung*).

Die Beziehungen zwischen den Einträgen einer Klassifikation bestehen hauptsächlich aus Hierarchischen Relationen (2.2.8/S. 28). Vereinzelt gibt es auch Querverweise. Die meisten Klassifikationen sind streng *monohierarchisch* aufgebaut, d. h. eine Klasse kann nur eine Oberklasse haben. In vielen Systemen kann ein in die Klassifikation eingeordnetes Objekt aber (zusätzlich) mehreren Klassen zugeordnet werden. Das im OPAC der HU Berlin (http://opac.hu-berlin.de/) enthaltene Buch „*Kommilitonen von 1933*“ über die Vertreibung von Studierenden der Berliner Humboldt Universität ist beispielsweise zusätzlich zur Klasse `NU 3025` den Klassen `AL 50712` (*Geschichte des Hochschul- und Universitätswesen der Humboldt Universität*) und `NU 7100` (*Sonstige Geschichte der Studenten als Teil der Geschichte der Wissenschaften*) zugeordnet. Zur Beschreibung eines Objektes mit mehreren Begriffen siehe auch unter *Kombinierte Begriffe und Mapping* (2.2.5/S. 24).

Wie auch andere Begriffssysteme sind Klassifikationen und Systematiken keine starren Gebilde, sondern mit den in ihnen enthaltenen Objekte stetigen Änderungen unterworfen. Zur Klärung der Bedeutung einzelner Klassen setzt man u. a. Kommentare (so genannte „Scope Notes“) und Verweisen zwischen verwandten Klassen ein.

7 Die RVK ist frei im Internet zugänglich unter http://www.bibliothek.uni-regensburg.de/rvko_neu/

Zur Darstellung von Klassifikation und anderen (mono)hierarchischen Zusammenhängen werden Bäume verwendet. Beispiele für solche Darstellungen sind Stammbäume, Verzeichnisstrukturen, Inhaltsverzeichnis u.v.a.m. Während beispielsweise physische Objekte, die sich nur an einer Stelle befinden können, nur in monohierarchische Klassifikation aufgenommen werden, ist in elektronischen Medien eine stärkere Verlinkung nach dem Hypertextprinzip möglich. Untersuchungen haben jedoch gezeigt, dass sich Benutzer in hierarchischen Strukturen deutlich einfacher zurechtfinden, während sich Netzwerkstrukturen besser zur gezielten Informationssuche eignen.

2.1.6. Register

Ein Register ist eine lineare Anordnung von Bezeichnungen oder Begriffen und Verweisen auf (Text)stellen, in denen diese Begriff verwendet oder definiert werden. Eine andere Bezeichnung für Register ist *Index*. Die Ordnung in einem Register kann alphabetisch, systematisch oder gemischt vorgenommen werden. Beispiele für Register innerhalb von Dokumenten sind Verzeichnisse (Schlagwortregister, Inhaltsverzeichnis, Namensverzeichnis, Abkürzungsverzeichnis). Während einfache Register keine interne Gliederung oder lediglich eine Unterteilung nach Anfangsbuchstaben aufweisen (einstufiges Register), bilden Inhaltsverzeichnisse hierarchisch die gesamte Struktur eines Dokumentes ab. Die Grenzen zwischen Klassifikation und (mehrstufigen) Registern sind fließend — man kann die Klassifikation auch als Spezialfall eines Registers betrachten. Mittels Permutation (2.2.4) lassen sich auch einzelne Wortbestandteile von Bezeichnungen über ein Register zugänglich machen.

Die alphabetische Ordnung von Einträgen ist übrigens alles andere als trivial. In der Lexikopgraphie wird zwischen glattalphabetischer, initialalphabetischer, nestalphabetischer und striktalphabetischer Anordnung unterschieden ([Engelbert], S. 125). Zusatzbuchstaben wie die deutschen Umlaute können entweder an einer eigenen Stelle eingefügt („ä“ nach „a“), anderen Buchstaben gleichgeordnet („ä“=“a“) oder transliteriert werden („ä“=„ae“). Auch die Behandlung von Sonderzeichen, Leerzeichen, Groß/Kleinschreibung und Ziffern muss geklärt werden.

2.1.7. Glossar

Ein *Glossar* ist eine Wörterliste mit Erklärungen, die meist Bestandteil eines (Fach)textes ist und die Bedeutung von erklärungsbedürftigen Wörtern festlegt. Im Gegensatz zu einem Wörterbuch enthält ein Glossar vorrangig Definitionen; darüber hinaus können auch weitere lexikographische Informationen – beispielsweise zur Sprache und Herkunft (Etymologie) – enthalten sein. Auch Verweise zwischen verwandten Begriffen sind möglich. Die Definitionen in einem Glossar sollen anders als in größeren Nachschlagewerken wie etwa einem Lexikon möglichst kurz und eindeutig sein, da ein Glossar in der Regel eine bestimmte *Terminologie* aus Fachwörtern definiert. Die Definitionen sind jedoch nicht immer trennscharf genug, so dass ein Begriffe mitunter auch unter mehreren Einträgen in einem Glossar vorkommen kann. Die Grenzen zwischen Glossar und Lexikon sind fließend, so werden manche Glossare auch als *Definitionslexikon* bezeichnet.

Beispiel für ein einfaches Glossar

- **Begriff:** eine Denkeinheit die als Konzept in der menschlichen Vorstellung existiert und nicht an eine bestimmte Sprache gebunden ist.
- **Benennung:**siehe Bezeichnung
- **Bezeichnung:** Repräsentation eines Begriffs mit sprachlichen (Benennung) oder anderen Mitteln.

2.1.8. Thesaurus

Ein Thesaurus ist in der Dokumentationswissenschaft ein kontrolliertes Vokabular, dessen Begriffe durch Relationen miteinander verbunden sind. Die Bezeichnungen „Thesaurus“ stammt aus dem Griechischen (thesauros=Schatz, Schatzhaus) und wurde früher für den (oder ein Sammelwerk des) Wortschatz(es) einer Sprache benutzt. Bekannt sind unter anderem der *Thesaurus Linguae Graecae* und *Thesaurus Linguae Latinae*. Auch Synonymwörterbücher wie beispielsweise *Roget's Thesaurus of English Words & Phrases* oder in Textverarbeitsungsprogrammen integrierte Programme werden als Thesaurus bezeichnet. Mit einem eigentlichen (dokumentarischen) Thesaurus haben sie gemein, dass beide der Wortfindung dienen. In letzterem stehen jedoch die Worte als Benennungen (Deskriptoren) eindeutig für bestimmte Begriffe. Thesauri haben sich wie Schlagwortkataloge als geeignete Mittel zur Sacherschließung (siehe) und zur Recherche in Dokumenten erwiesen.

Die Thesaurusnormen DIN 1462 bzw. das internationale Äquivalent ISO 2788 sehen folgende Relationsarten und die dazugehörige Abkürzungen vor:

	DIN 1462	***ISO 2788***
Äquivalenzrelation	**BF** - Benutzt für Synonym	**UF** - Used for
	BS - Benutze Synonym	**USE/SYN** - Use synonym
Hierarchische Relation	**OB** - Oberbegriff	**BT** - Broader Term)
	UB - Unterbegriff	**NT** - Narrower term
	Ggf. wird zwischen generischer und partitiver Hierarchie unterschieden	
Assoziative Relation	**VB** - Verwandter Begriff	**RT** - Related term
		TT - Top term

Die Relationen dienen in erster Linie dazu, bei der Indexierung und Recherche die passende Benennungen für einen gesuchten Begriff zu finden. Unterschiedliche Schreibweisen, Synonyme oder als gleichbedeutend behandelte Quasi-Synonyme, Abkürzungen, Übersetzungen etc. werden durch *Äquivalenzrelationen* miteinander in Beziehung gesetzt und in vielen Fällen eine Bezeichnung als so genannte *Vorzugsbenennung* ausgewählt. *Hierarchische Relationen* verknüpfen allgemeine mit spezielleren Begriffen oder Verbinden Teile mit einem größeren Ganzen. Hierarchien lassen sich auch für die Recherche einsetzen, indem bei der Suche nach einem Oberbegriff auch die jeweiligen Unterbegriffe gefunden werden. Im Beispielthesaurus würde eine Suche nach „Amphibienfahrzeug“ beispielsweise auch Dokumente finden, die lediglich mit dem Schlagwort „Schwimmwagen“ versehen sind. *Assoziative Relationen* werden schließlich dort eingesetzt, wo Begriffe auf irgend eine andere Art und Weise zusammenhängen als durch Äquivalenz oder Hierarchie. Es wird empfohlen, bei Bedarf die Relationen genauer zu differenzieren und eigene Relationen einzuführen.

Beispiel für einen Thesaurus (Auszug)			
Kraftfahrzeug OB *Fahrzeug* UB *Amphibienfahrzeug*	**Wasserfahrzeug** OB *Fahrzeug* UB *Amphibienfahrzeug*	**Amphibienfahrzeug** BF *Schwimmwagen* OB *Kraftfahrzeug* *Wasserfahrzeug*	**Schwimmwagen** BS *Amphibienfahrzeug*

Im Gegensatz zu einer Klassifikation treten in einem Thesaurus häufig Polyhierarchien auf, d.h. ein Begriff kann mehrere Oberbegriffe haben. Es ist auch nicht Aufgabe eines Thesaurus, ein Wissensgebiet systematisch in seiner Gesamtheit zu strukturieren. Der Menschliche Drang, Objekte durch Klassifikation zu ordnen, kann die Erstellung eines Thesaurus sogar sehr behindern.

Obwohl die Gemeinsamkeiten zu anderen Begriffssystemen wie beispielsweise Semantischen Netzen auf der Hand liegen (viele Ontologien sind nur wenig mehr als um spezielle Relationen erweiterte Thesauri), sind Thesauri relativ unbekannt, während sich die monohierarchische Ordnung fast überall findet. Dies mag zum einen daran liegen, dass sich die freie Graphenstruktur eines Thesaurus schwieriger überblicken und verwalten lässt als eine Klassifikation.

Zum anderen muss man zugeben, dass Bibliotheken, Museen und Dokumentationseinrichtungen, in denen Thesauri im großen Stil eingesetzt werden, der Entwicklung in der Informatik oft um Jahre hinterherhinken. Die schlimmstenfalls nur in gedruckter Form vorliegenden Thesauri lassen nur mit Fantasie erahnen, dass es sich eigentlich im Begriffsnetze handelt, die für eine Hypertextdarstellung geradezu prädestiniert sind. So kommt es, dass sich aktuelle Entwicklungen wie XML, RDF und Ontologien (noch) nicht in den Normen und der Standardliteratur zu Thesauri widerspiegeln (in der aktuellen Forschung ist dies zumindest eher der Fall). Auf der anderen Seite gehen gleichzeitig die nicht zu unterschätzenden in der Praxis mit Thesauri gemachten Erfahrungen nur sehr spärlich in die Entwicklung verwandter Techniken im Rahmen des Semantic Web ein.

Mit diesen Beschränkungen ist das einzige deutschsprachige Handbuch zu Thesauri [Wersig], das schon etwas älter ist (1975), in Bereichen trotzdem noch sehr lehrreich. Eine kürzere Darstellung des Aufbaus und der Anwendungspraxis eines Thesaurus findet sich beispielsweise in [Wolters].

Das in Kapitel 3.2 vorgestellte *Thema*–Datenmodell basiert im wesentlichen auf der Struktur eines Thesaurus, die durch Konzepte aus anderen Begriffssystemen und Datenformaten ergänzt wurde.

2.1.9. Mind Maps, Konzeptuelle Karten und Semantische Netze

Eine der Hauptaufgaben bei der Entwicklung eines Begriffssystems ist es, sich über die in ihm enthaltenen Begriffe klar zu werden. Dazu werden Begriffe gesammelt, in Beziehung gesetzt und voneinander abgegrenzt. Ein aus der Lerntheorie stammender Ansatz für diesen Prozess sind die so genannten *Mind Maps*. Dabei handelt es sich um graphische Darstellungen von Konzepten, Begriffen, Ideen und Strukturen, die in der Regel hierarchisch von einen Hauptbegriff ausgehend in ein konzentrisches Baumdiagramm eingezeichnet werden. Mind Maps werden häufig in einem Brainstorming-Prozess eingesetzt, um sich einen Überblick über ein Themengebiet zu verschaffen und eignen sich auch gut als Mnemotechik und für didaktische Zwecke.

Das Konzept von Mind Maps geht auf den Psychologen Tony Buzanan und sein 1971 erschienenes Buch „An Encyclopedia of the Brain and Its Use“ zurück. Eine Verallgemeinerung der Baumstruktur von Mind Maps auf allgemeine Graphen führt zu *Konzeptuelle Karten* (*conceptual maps*) die bereits bereits in der 1960ern von Joseph Novak eingeführt wurden. Ein sehr ähnlicher aber aus einer anderen Richtung stammender Ansatz sind *Konzeptuelle Graphen* (*conceptual graphs*). Sie basieren unter anderem auf logischen Theorie von Charles Peirce. Eine neuere Bezeichnung für Konzeptuelle Graphen ist *Semantische Netze* (siehe auch 2.1.10). Konzeptuelle Graphen bzw. Semantische Netze sind graphische Begriffsnetze, die eine definierte Semantik aufweisen, d. h. die mit Linien und Pfeilen dargestellten Beziehungen zwischen einzelnen Begriffen besitzen eine definierte Bedeutung. Bei hinreichender Formalisierung lassen sich Konzeptuelle Graphen in logische Aussagen überführen (siehe [Sowa] und http://users.bestweb.net/~sowa/cg/). Einen Vergleich von Konzeptuellen Graphen den Grundlagen des Semantic Web zieht Tim Berners-Lee in [TBL_b].

Beispiele

Mehrere Beispiele für kleine Semantische Netze finden sich im Kapitel 2.2.8. Ein komplexes, quasi-formales Beispiel ist die *Unified Modeling Language* (UML) bzw. Mit ihr ausgedrückte Modelle. In Kapitel 3.2 befindet sich eine UML-Repräsentation des Thema-Datenmodells.

Ein Beispiel, das sich je nach Sichtweise unter anderem als Datenbank, Terminologie, Semantisches Netz oder Thesaurus bezeichnen lässt ist das *WordNet*. WordNet ist ein seit 1985 am Cognitive Science Laboratory der Universität von Princeton entwickelter Wortschatz der Englischen Sprache.

WordNet besteht aus einer lexikalische Datenbank, die semantische Beziehungen zwischen den Wörtern enthält. Diese sind nach psycholinguistischen Erkenntnissen entworfen, da das WordNet ursprünglich entwickelt wurde, um natürlichsprachliche Texte für Computer verstehbar zu machen. Die Datenbank, die frei durchsuchbar und mitsamt Software kostenlos verfügbar ist, wird auch für andere Zwecke eingesetzt. Ihre Struktur ähnelt einem erweiterten Thesaurus.

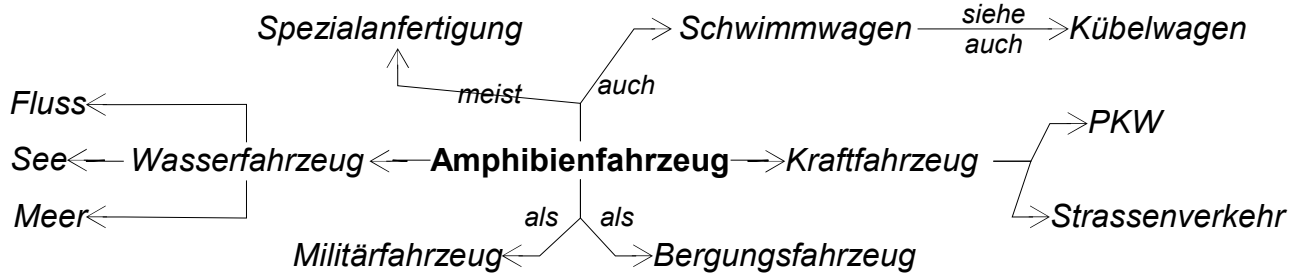


Bild 2: Beispiel für eine Mind Map

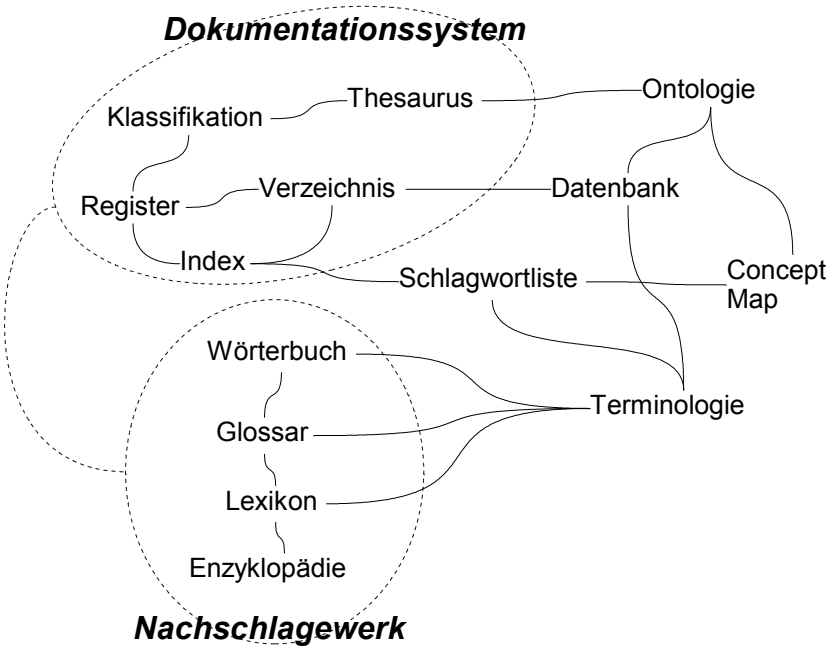


Bild 3: Beispiel für eine Konzeptuelles Netz

2.1.10. Ontologien, Wissensrepräsentation und das *Semantic Web*

Unter einer *Ontologie* versteht man in der Informatik ein formal definiertes System von Konzepten und Relationen zwischen ihnen. Zusätzlich enthalten Ontologien (zumindest implizit) Regeln. Eine andere Bezeichnung für eine Ontologie ist die eines Semantischen Netzes. Im engeren Sinne bezieht man die aus der Philosophie stammende Bezeichnung Ontologie oft auf Begriffssysteme, die anderen Begriffssystemen zugrunde liegen oder die die Struktur anderer Begriffssysteme beschreiben. Beispielsweise kann eine Ontologie die Begriffe „Ort“ und „Zeit“ enthalten, auf die in einem einem konkreten semantischen Netz, das z. B. einzelne Termine enthält, „Termin“ verwiesen wird. Eine Ontologie, die anderen Ontologien zugrunde liegt, wird auch *Meta-Ontologie* genannt. Ob eine Ontologie eine Meta-Ontologie ist oder nur ein einfaches Semantisches Netz, hängt jedoch vom Standpunkt ab, da sich die Verschachtelung beliebig (auch rekursiv, wie beispielsweise in der Beschreibung von RDFS in RDFS selbst) fortsetzen lässt.

Ontologien spielen vor allem im Bereich der Künstlichen Intelligenz, beim Wissensmanagement (*Knowledge Management*) und in der Datenmodellierung eine Rolle und dienen grob gesagt der Wissensrepräsentation und -verarbeitung. Die Erstellung einer Ontologie ist in etwa vergleichbar mit der Konzeption einer Datenbankstruktur – in gewisser Weise lassen sich auch Datenbankschemata als Ontologien auffassen. Eine einfache Formen einer Ontologien mit einem beschränkten Satz von Relationen ist ein Thesaurus. Die Abgrenzung von Ontologien zu anderen Begriffssystemen ist – auch angesichts des inflationären Gebrauches des Wortes „Ontologie“ – nicht einfach. Als herausragendes Merkmal lässt sich jedoch feststellen, dass in Ontologien die Relationen (und ggf. auch die konkreten, mit Relationen ausgedrückte Aussagen) selber wieder Begriffe des Begriffssystems sind und dass Ontologien logische Aussagen und Regeln enthalten (können), die in einer formalen Sprache festgelegt sind. Für die Formulierung von Ontologien existieren verschiedene logik-basierte Sprachen, die weitgehend den unter unter 2.2.9 genannten Sprachen für Regeln entsprechen. Die am verbreitetetsten Vertreter sind RDF/RDFS (3.1.4), DAML+OIL bzw. die daraus entwickelte Web Ontology Language (OWL) und der Topic Maps Standard (3.1.5).

Eine der größten Ontologien ist die Wissensdatenbank Cyc (vom englischen *encyclopedia*). Sie wird seit 1984 beständig weiter entwickelt und ist als OpenCyc[8] auch in einer freien Version verfügbar. Das ambitionierte Ziel des Cyc-Projektes war es, das gesamte menschliche Alltagswissen in einem formalen System abzubilden (*Knowledge Representation*). Cyc besteht aus einer Menge von einfachen Regeln (Wasser macht Objekte, die sich in ihm befinden nass), die es ermöglichen sollen, dem Computer ein gewisses Maß an „gesundem Menschenverstand“ einzuprogrammieren. Bspw. kann ein Programm mit Hilfe der Cyc-Ontologie aus der Aussage, dass jemand im Meer schwimmt und dass man sich beim Schwimmen im Wasser befindet, Schlussfolgern, dass die betreffende Person dabei nass wird. Die Cyc-Ontologie enthält rund 100.000 Begriffe und 1.000.000 Aussagen. Cyc findet vor allem in natürlichsprachlichen (englischen) Systemen Anwendung. Die Cyc-Ontologie ist in einer eigenen Sprache (CycML) beschrieben, die sich auch nach DAML exportieren lässt.

Ontologien haben mit der Idee des so genannten Semantic Web innerhalb der letzten Jahre einen Aufschwung erfahren. Dies hat jedoch nicht unbedingt zu einer Klärung des Begriffes Ontologie beigetragen. In vielen Fällen handelt es sich bei den als Ontologien bezeichneten Strukturen lediglich um Klassifikationen oder Thesauri. Von der in RDF als *Reification* bezeichneten Möglichkeit von Relationen über Relationen und von Regeln wird (unter anderem aufgrund ihrer Komplexität) relativ wenig Gebrauch gemacht, obwohl gerade diese Merkmale Ontologien von anderen Begriffssystemen abheben. Der Vorteil von modernen Ontologien ist jedoch, dass sie in der Regel in einer formalen Sprache abgefasst sind, die sich (in der Regel in einer XML-

8 http://www.opencyc.org/ bzw. http://www.cyc.com/

Repräsentation) leicht weiterverarbeiten lässt. Die Referenzierbarkeit und Wiederverwendbarkeit wird durch die konsequente Verwendung von URI URL und Standards wie XLink gewährleistet. Diese sind jedoch eher technischer als konzeptueller Natur. Einer der Hauptnutzen der mit der Idee des Semantic Web angestoßenen Entwicklungen ist die Öffnung von lokal begrenzten Systemen und die Ablösung von proprietären Formaten.

Das Semantic Web ist als Gesamtheit von über das Internet zugänglichen Ressourcen, die eine semantische Struktur besitzen und durch Metadaten beschrieben werden. Diese basieren auf Ontologien, über die Agenten und Suchmaschinen selbständig Informationen sammeln und auswerten können. Die in Schichten aufgebaute Architektur des Semantic Web wurde 2000 von Tim Berners-Lee auf der XML 2000 Konferenz vorgestellt (siehe Bild 7) und 2001 in *Scientific American* einer breiteren Öffentlichkeit bekannt gemacht [TBL_a].

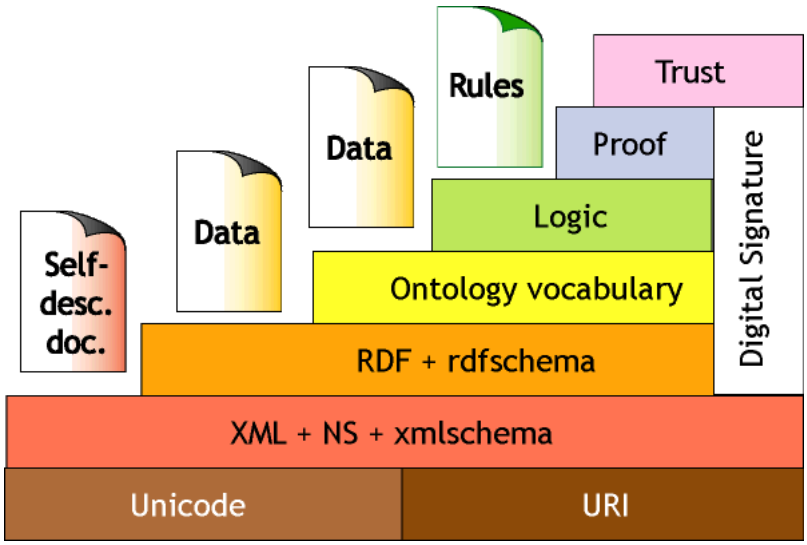


Bild 4: Die Architektur des Semantic Web. Von Tim-Berners-Lee auf der XML 2000 Konferenz vorgestellten Grafik

Die unteren Schichten der Technik des Semantic Web (Unicode, URI, XML, Namespaces und XML Schema) haben sich bereits als Standards für viele Anwendungen durchgesetzt und auch RDF und RDF Schema sind relativ etabliert, ebenso wie die konkurrierenden Topic Maps, die sich jedoch in RDF übersetzen lassen. Zur Wissensrepräsentation (*Ontology vocabulary*) ist inzwischen ist inzwischen DAML+OIL bzw. die Ontology Web Language (OWL) die am weitesten verbreitete Sprache, wobei daneben noch immer zahlreiche andere Sprachen existieren. Der größte Handlungsbedarf besteht noch bei den logischen Aussagen und Regeln (*Logic*) und darüber liegenden Ebenen.

Bisher mangelt es auch noch an einfachen Programmen zur Erstellung und Nutzung von Ontologien im Semantic Web (gesucht ist eine so genannte *„killer application"*). Da die unter Koordination des Word Wide Web Consortiums entwickelten Datenformate Techniken noch relativ neu und teilweise in der Entwicklung nicht abgeschlossen sind, haben sie sich in existierenden Softwareprogrammen noch nicht durchgesetzt. Die bestehende Software lässt sich grob einteilen in

- Ontologie-Editoren (Protégé, OntoEdit, OilEdit…)
- Thesaurus Software (siehe http://www.willpower.demon.co.uk/thessoft.htm)
- MindMap Software (siehe http://www.mindmap.ch/software.htm)
- Terminologie Software (siehe http://www.iim.fh-koeln.de/dtp/)

Zur Erstellung von Nachschlagewerken lassen sich verschiedene Programme benutzen. Ein viel ver-

sprechender Ansatz für Hypertexte ist die Software der Wikipedia-Enzyklopädie [Wikipedia].

Eine grundsätzliche Schwierigkeit liegt in der Architektur des Semantic Web selbst. So lassen sich die im Modell aufeinander aufbauenden Schichten *RDF* (Resourcen und Relationen), *Ontology* (gemeinsame Begriffe) und *Logic* (Regeln für Inferenzmechanismen) nicht so sauber trennen, wie es in der Informatik mit Abstraktionsebenen gewohnt ist. Die in folgender Interpretation dargestellten Ebenen lassen sich besser zu einer Ebene der Semantik (bzw. übergeordnet der Semiotik) zusammenfassen, aus denen sich Ontologien und andere Begriffssysteme zusammensetzen.

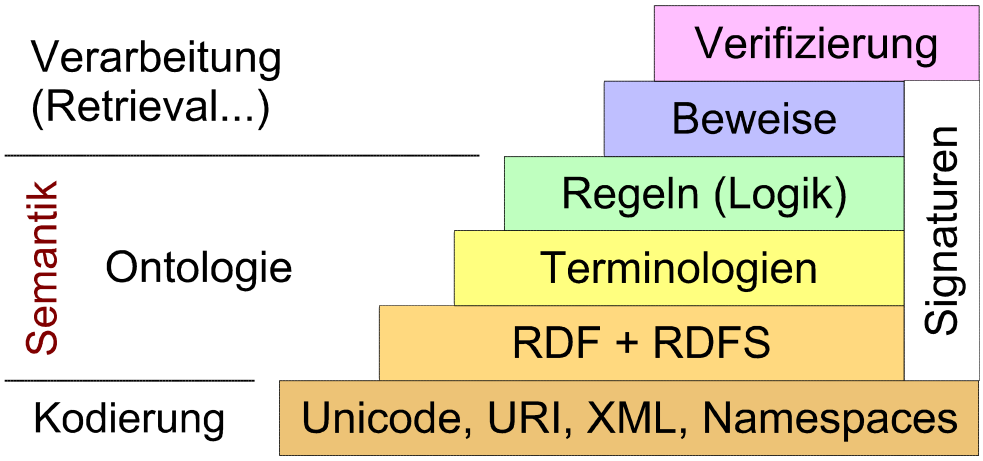


Bild 5: Die Architektur des Semantic Web. Leicht modifizierte Interpretation.

Trotz der bestehenden Schwierigkeiten ist das Semantic Web und die damit verbundenen Technologien ein lebhaftes und praxisrelevantes Forschungsfeld. Vor allem die technische Umsetzung (Ebene der Kodierung) und damit verbundenen Möglichkeiten der Vernetzungen ermöglichen einen qualitativen Fortschritt bei der Entwicklung und beim Einsatz von Begriffssystemen.

Es ist wünschenswert, dass sich Wissenschaftler aus anderen mit Begriffssystemen beschäftigten Gebieten (1.2, vor allem auch Geisteswissenschaften) stärker als bisher an der fortschreitenden Entwicklung beteiligen und ihre Erfahrungen einbringen, was für beide Seiten Vorteile bringen kann.

2.1.11. Mathematische Strukturen

Die Grundlagen der Definition von Ontologien und anderen *formalen* Begriffssystemen sind mathematischer Natur. Sobald logische Aussagen und Schlussfolgerungen über Begriffssysteme gemacht werden können, liegt diesen ein mathematisches Modell zugrunde. Ebenso wie die philosophischen Grundlagen von Begriffssystemen sind deren mathematischen Eigenschaften in der Praxis jedoch nicht immer direkt von Bedeutung, da sie implizit vorausgesetzt werden. Für die Modellierung interessante Disziplinen innerhalb der Mathematik sind unter anderem

- die Verbandstheorie (untersucht hierarchischer Strukturen)
- die Kategorientheorie (untersucht Klassen von Objekten anhand möglicher Relationen)
- und die Graphentheorie (untersucht Graphen, das sind Netzen miteinander verbundener Objekte)

Neben der theoretischen Betrachtung, die Aufschluss über die generellen Eigenschaften und Möglichkeiten eines konkreten Begriffssystems gibt, können mathematische Verfahren in Form von Algorithmen beispielsweise dazu dienen, aus bestehenden Begriffssystemen neue Infomationen abzuleiten (Inferenz) oder aus verschiedenen Quellen mittels Methoden der Statistik (z. B. über neuronale Netze[9]) automatisch Begriffe zu extrahieren.

9 Vgl. beispielsweise Andrei Krell: *Neuronale Netze und deren Eignung zur Klassifikation im Data Mining.* (Diplomarbeit) Chemnitz 26.9.2001

2.2. Allgemeine Bestandteile und Konzepte

Beim einem Vergleich der verschiedenen im letzten Kapitel dargestellten Begriffssysteme lassen sich neben den jeweiligen Unterschieden eine Reihe von Gemeinsamkeiten feststellen. Aus ihnen sollen im Folgenden die fundamentalen konzeptuellen Bestandteile von Begriffssystemen aufgeführt werden. Beispielsweise enthalten alle Begriffssystemen *Begriffe*, ganz gleich, ob sie in Form von Einträgen in einem Lexikon, als Deskriptoren in einem Thesaurus oder als Klasse einer Klassifikation auftreten. Daneben sind verschiedene Formen von *Bezeichnungen*, *Relationen* und *Regeln* von tragender Bedeutung. Sie bilden die grundlegenden Bausteine eines Begriffssystems. Eine wesentliche Eigenschaft vieler Begriffssysteme ist die Trennung von Begriffen (2.2.1) und Bezeichnungen (2.2.2). Neben einfachen Relationen (2.2.8), die sich in bestimmte Relationsarten einteilen lassen, können mehrere Begriffe durch Kombination (2.2.5) verbunden sein.

Die einzelnen hier genannten Bestandteile sind natürlich je nach Art des Begriffssystems mehr oder weniger stark ausgeprägt oder zum Teil auch gar nicht vorhanden. Auch von den unterschiedlichen Bezeichnung für gleiche Konzepte darf man sich nicht verwirren lassen.

2.2.1. Begriffe

»Denn eben, wo Begriffe fehlen,
Da stellt ein Wort zur rechten Zeit sich ein.
Mit Worten läßt sich trefflich streiten,
Mit Worten ein System bereiten,
An Worte läßt sich trefflich glauben.
Von einem Wort läßt sich kein Jota rauben.«
(Goethe, Faust I, Schülerszene)

Die einzelnen Begriffe (englisch *concept*) eines Begriffssystems werden je nach Einsatzzweck des Systems und der Fachwissenschaft, in der sie entwickelt wurden, u. a. als Konzept, Klasse, Objekt, Ressource, Knoten, Eintrag, Element, Entität u.v.a.m. bezeichnet. Es sei darauf hingewiesen, dass in der Umgangssprache vermehrt *Begriff* synonym für *Wort* verwendet wird ("Ein Begriff aus dem Englischen") und deshalb im Zweifelsfall die Bezeichnung *Konzept* verwendet werden sollte.

Bei den Konzepten oder Begriffen handelt es sich um unterscheidbare Denkeinheiten, die zunächst nur in der Vorstellung von Menschen vorhanden sind. Der Unterschied zwischen *Begriffen* als Konzepten in der menschlichen Vorstellung, *Bezeichnungen* als Platzhalter für Begriffe in einer Sprache und den bezeichneten *Objekten* selber wird oft mit Hilfe des so genannten *semiotischen Dreiecks* dargestellt (wobei die einzelnen Bestandteile der Dreiecks je nach Autor sehr unterschiedliche Namen haben können).

Die Aufgabe eines Begriffssystems ist es, diese Konzepte definiert in Beziehung zu setzen und durch ein System von Bezeichnungen so in Sprache auszudrücken, dass sie von verschiedenen Menschen verstanden werden. Die grundsätzliche Schwierigkeit dieses Vorgehens besteht darin, dass keine Sprache die Vorstellung eines Menschen in seiner Gesamtheit ausdrücken kann (die Geschichte der gescheiterten Versuche, eine solche Sprache zu finden, hat UMBERTO ECO in seinem 1995 erschienen Buch *Die Suche nach der vollkommenen Sprache* dargestellt [Eco]). Kein Lexikoneintrag, z. B. für den Begriff "Eisenbahn" wird je alles enthalten, was man sich unter "Eisenbahn" vorstellen bzw. damit assoziieren kann. Dies liegt schon allein daran, dass sich Vorstellungen ständig ändern und von individuellen Erfahrungen geprägt sind. Ein Begriffssystem muss sich deshalb auf die "objektiven" Bestandteile eines Begriffes beschränken und diese

festhalten. Die Frage, ob die vorgestellten und in Zeichen aus Schrift und Sprache dargestellten Begriffe realen Objekten aus der "wirklichen" Welt entsprechen, ist eine rein philosophische und sei dahingestellt. In der Regel geht man pragmatisch frei nach Platon davon aus, dass die Begriffe konkreten Ideen entsprechen, die auch unabhängig vom Menschen existieren. So setzt die Beschreibung des Begriffes "Eisenbahn" in einem Begriffssystem voraus, dass es so etwas wie *die* Eisenbahn an sich gibt, die objektiv beschrieben werden kann.

2.2.2. Bezeichnungen und Benennungen

Obwohl ein Begriffssystem eigentlich von Begriffen handelt, kann es nur Sprache (in Form von Zeichen) enthalten, da uns kein anderes Kommunikationsmittel zur Verfügung steht (zumindest solange der Telepathie ein Wunsch bleibt). *Alle* Bestandteile eines Begriffssystems lassen sich nur durch Sprache ausdrücken. Die direktesten Bestandteile sind die Bezeichnungen für die einzelnen Begriffe. Eine Bezeichnungen ist ein schriftliches, bildliches, akustisches oder anderes Zeichen, das für einen bestimmten Begriff steht. Wenn die Bezeichnung ein natürlichsprachlicher Ausdruck ist, spricht man auch von *Benennung* (englisch *term*). Nach DIN 2342 ist eine Bezeichnung eine "*Repräsentation eines Begriffs mit sprachlichen oder anderen Mitteln*", die äquivalente internationale Norm ISO 1087 spricht einfach von "*any representation of a concept*". Im umgangssprachlichen Gebrauch wird die Bezeichnung (sic!) "Begriff" oft für eine Benennung anstatt für den Begriff selber verwandt. Während es sich bei einer Eisenbahn (Begriff) um ein Schienen gebundenes Fahrzeug handelt, ist "Eisenbahn" (Bezeichnung) ein Wort mit 9 Buchstaben, das im Deutschen als Benennung für eine Eisenbahn verwendet wird.

Ein Begriff kann viele verschiedene Bezeichnungen haben (Eisenbahn, railroad/way, ferrocarril...). Gleichzeitig kann eine Bezeichnung für beliebig viele unterschiedliche Begriffe stehen. Dies liegt daran, dass die Bezeichnung eines Begriffes durch eine Zuweisung vorgenommen wird, die sich als Ergebnis des tatsächlichen Sprachhandelns über einen Zeitraum entwickelt hat. Es entsteht so eine Relation zwischen einem Begriff und seiner Bezeichnung, die nur in ihrem Kontext sinnvoll ist. So steht es zum Beispiel jedem frei, seinen Hund "Eisenbahn" zu nennen. Obwohl Bezeichnungen relativ willkürlich sind, beeinflussen sie über die Vorstellung, die sie in den Köpfen von Menschen erzeugen, die Begriffe, für die sie stehen. Die eigentliche Problematik liegt nämlich, wie schon erwähnt, darin, dass die Begriffe keine festen Einheiten bilden, sondern für die Verwendung in einem Begriffssystem erst geklärt werden müssen. Erst wenn klar ist, ob mit Eisenbahn die Eisenbahn als Transportmittel (englisch *railway*) oder der konkrete Zug (englisch *train*) gemeint ist, lässt sich sagen, wofür das Wort "Eisenbahn" steht. Die Festlegung, welche Bezeichnung in einem Begriffssystem für welchen Begriff steht, wird als *terminologische Kontrolle* bezeichnet. Wenn jede Benennung eindeutig einem Begriff zuzugeordnet ist, handelt es sich um ein *kontrolliertes Vokabular*.

In einem Begriffssystem kann es verschiedene Arten von Bezeichnungen geben. Für die maschinelle Verarbeitung ist die Zuweisung eines *Identifikators* notwendig (siehe 2.2.6). Andere nicht-natürlichsprachliche Bezeichnungen sind *Notationen* (2.2.7). Für den Menschen sind jedoch bestimmte Wörter (Benennungen) die entscheidende Art der Bezeichnung. Hierbei kann es verschiedene Arten von Benennungen geben. Häufig anzutreffen sind verschiedene Schreibweisen, alternative Benennungen (*railway – railroad*), wobei eine der Benennungen als *Vorzugsbenennung* zur präferierten Verwendung vorgeschrieben werden kann, sowie Abkürzungen und Übersetzungen.

Als vorstellbare Objekte können Bezeichnungen selber wiederum Begriffe sein (z. B. in einem etymologischen Wörterbuch). Dies ändert jedoch nichts an der grundlegenden Unterscheidung zwischen Begriff und Benennung, der man sich in jedem Begriffssystem bewusst sein sollte.

2.2.3. Homonyme und Qualifikatoren

Oft kommt es vor, dass eine Bezeichnung für mehrere Begriffe stehen kann. Eine *Bank* kann beispielsweise eine Sitzgelegenheit oder ein Geldinstitut sein. Solche mehrdeutigen Benennungen nennt man *Homonyme*. Der umgekehrte Fall, wenn mehrere Bezeichnungen für den selben Begriff stehen, sind *Synonyme* (siehe 2.2.5). Bei den Homonyme werden gelegentlich noch die gleich lautenden Bezeichnungen (z. B. *Mohr – Moor*) als *Homophone* und gleiche Schreibweisen (z. B. *Wach-stube – Wachs-tube*) als *Homographen* unterschieden. Homonyme, die durch Bedeutungsverschiebung aus der unterschiedlichen Interpretation eines Wortes entstanden sind, nennt man auch *Polyseme*. Bei genauerer Betrachtung kann fast jeder Begriff je nach Kontext in verwandte Teilbegriffe zerlegt werden. Das Geldinstitut *Bank* kann zum Beispiel konkret für eine Gebäude, eine Geschäftsform oder eine Firma stehen. Der Mensch kann in der Regel mit solchen Mehrdeutigkeiten umgehen, da die konkrete Bedeutung (hoffentlich) aus dem Zusammenhang ersichtlich ist. In der alltäglichen Kommunikation ist eine Unterscheidung zwischen Begriffen und Benennungen deshalb selten explizit notwendig. Im wissenschaftlichen Diskurs treten jedoch oft Missverständnisse auf, weil Begriffe in verschiedenen Fachsprachen oder sogar von verschiedenen Autoren sehr unterschiedlich definiert werden (z. B. *Ring* als Schmuck, in der Mathematik, Kunst, Astronomie).

Formale Begriffssysteme sollten möglichst keinen Platz für Interpretationsmöglichkeiten geben – eine Bezeichnung soll in vielen Fällen auch für sich stehend immer das gleiche bedeuten, indem sie auf einen eindeutigen Begriff verweist. Wenn in einem System ungewollt Homonyme auftreten, müssen sie deshalb durch *Homonymzusätze* (*Qualifikatoren*) voneinander unterscheidbar gemacht werden. Diese bestehen nach DIN 1463 aus in runden Klammern angehängten Zusätzen. In den *Regeln für den Schlagwortkatalog* [RSWK] der Schlagwortnormdatei (SWD) werden nach §10 statt runden Klammern Winkelklammern (< und >) benutzt, also bspw. zur Bezeichnung “Absatz”:

`Absatz`	*für den Absatz von Waren, weil dies innerhalb des Einsatzzweckes der SWD (Literaturverschlagwortung) die häufigste Verwendungsform ist*
`Absatz <Text>`	
`Absatz <Schuh>`	*wobei die Bezeichnung* `Schuhabsatz` *vorzuziehen ist*

Die in §306 der RSWK festgelegten Regeln sind übrigens ein lehrreiches Beispiel für den kontrollierten Umgang mit Homonymen in einem Begriffssystem.

Bevor jedoch unnötig Homonymzusätze eingeführt werden, sollte man überprüfen, ob Homonymprobleme nicht auch durch das Ausweichen auf andere Bezeichnungen gelöst werden können. Statt `Bank (Sitzgelegenheit)` kann man auch festlegen, dass eine Bank zum Sitzen immer `Sitzbank` genannt wird. Ob `Eisenbahn (Spielzeug)` oder einfach `Spielzeugeisenbahn` die bessere Bezeichnung ist, hängt wie bei fast allen Entscheidungen nicht zuletzt von jeweiligen Fachgebiet und Einsatzzweck des Begriffssystems ab.

Die korrekte Zuordnung von Begriff und Benennung lässt sich allgemein durch die Angabe eines Kontextes lösen. Generell sollten Homonymzusätze möglichst selber eindeutig definierte Begriffe sein. Zusätzlich kann vereinbart werden, dass die Qualifikatoren zum Beispiel ein bestimmtes Fachgebiet bzw. eine bestimmte Fachsprache bezeichnen sollen (Ring <Umgangssprache>, Ring <Mathematik>, Ring <Astronomie>...). Eine andere Möglichkeit der Angabe von Kontexten sind Quellennachweise (siehe 2.2.10). Im XML Topic Map Standard wird die eindeutige Benennung von Begriffen (dort *topics*) durch Zuordnung von so genannten *scopes* gewährleistet (siehe 3.1.5).

Geklammerte Zusätze als Teil von Bezeichnungen werden nicht nur zur Unterscheidung von Homonymen, sondern auch bei der Permutation eingesetzt.

2.2.4. Permutationen

Permutationen sind verschieden Formen der Schreibweise einer Benennung, die sich durch das Vertauschen der einzelnen Elemente zusammengesetzter Wörter ergeben. Dadurch lassen sich auch einzelne Wortbestandteile in einem alphabetischem Register besser zugänglich machen. Bspw. werden in Namensverzeichnissen üblicherweise die Nachnamen vorangestellt, da sie das wesentliche Suchkriterium bilden (in Island werden allerdings aufgrund der Namensgebung, bei der der Familienname der Kinder aus dem Vornamen des Vaters gebildet wird, die Einträge im Telefonbuch nach den Vornamen sortiert). Beispiele für Permutationen (in verschiedenen Schreibweisen) sind:

```
Spielzeugeisenbahn          wird zu  Eisenbahn (Spielzeug-)
Terminologische Kontrolle   wird zu  Kontrolle (Terminologische)
Sitzgelegenheit             wird zu  Gelegenheit, Sitz-
```

Zur automatischen Erzeugung von Permutationen müssen Permutationsregeln festgelegt sein, die bestimmen, wo und auf welche Weise Bezeichnungen permutiert werden. Unter Umständen sind unerwünschte Trennungen durch spezielle Auszeichnungen zu unterbinden bzw. Permutationen, die nicht automatisch erzeugt werden, zu erzwingen. Ein Beispiel für solche Permutationsregeln findet sich im GOS-Thesaurus-Handbuch [Wolters]. Obwohl die Möglichkeit der Volltextsuche einen direkten Zugriff auf Bezeichnungen ermöglicht, haben gute Register mit permutierten Einträgen noch immer ihre Berechtigung, vor allem wenn es darum geht, sich einen schnellen Überblick zu verschaffen. Aus Unkenntnis oder aufgrund des Mehraufwandes wird jedoch in den meisten Fällen auf die Permutation verzichtet und lediglich eine einfache alphabetische Liste angeboten.

2.2.5. Synonyme, Vorzugsbenennungen und Begriffskombination

Wie schon erwähnt, kommt es häufig vor, dass es zu einem Begriff mehrere mögliche Bezeichnungen gibt. Synonymie ist ebenso wie Polysemie immer eine Frage des Kontextes. Ob zwei Bezeichnungen synonym sind und damit ihre Begriffe übereinstimmen, hängt also auch vom Verwendungszusammenhang des Begriffssystems ab. So differenziert beispielsweise ein Fachlexikon innerhalb seines Faches viel stärker als bei den Begriffen aus angrenzenden Gebieten. Ein anderes Beispiel ist der einfache Dublin-Core Standard, der mit seiner Beschränkung auf 15 Metadatenfelder zwangsläufig Kategorien, die in bibliothekarischen Datenformaten wie MAB und MARC unterschieden werden, als synonym zusammenfassen muss (die Einführung von Dublin Core Qualifiers [Kokkelink], die eine genauere Spezifizierung der Dublin Core Felder ermöglichen, sind ein Versuch, diesem Umstand Rechnung zu tragen). Bei einer solchen bewussten Zusammenlegung von Begriffen spricht man auch von *Quasi-Synonymen*. In multilingualen Thesauri werden mitunter verschiedene Arten von Synonymie unterschieden (voll und partiell) [DIN1463-2]/ [ISO5964]**.** Die Bestimmung von Synonymen ist auch die wesentliche Aufgabe beim Mapping von verschiedenen Begriffssystemen. In einem linguistischen Thesaurus, der u. a. zur Wortfindung beim Formulieren von Texten dient, bilden Verweise zu Synonymen (und Antonymen) die Hauptbestandteile.

Unabhängig davon können bei Bedarf verschiedene Arten von Synonymie unterschieden werden (verschiedene Schreibweisen, Flexion, Abkürzungen, Übersetzungen). Die durch Flexion und verschiedene Schreibweisen entstehenden Synonyme (Foto, Photo, Fotos, Photos) werden in der Regel nicht einzeln aufgenommen. Stattdessen werden zur Erkennung und Angleichung von Wortstämmen computerlinguistische Verfahren eingesetzt (*Stemming*) bzw. zur Vermeidung von unterschiedlichen Schreibweisen *Ansetzungsregeln* eingeführt. Diese können z. B. besagen, dass der Singular dem Plural vorzuziehen ist (siehe RSWK, § 9 und § 303).

Vorzugsbenennungen, Kontexte und Rollen

In den meisten herkömmlichen Begriffssystemen, die in irgendeiner Form auf Papier ausgedruckt werden müssen (Lexika, Glossare, Thesauri), gibt es zu jedem Begriff eine so genannte *Vorzugsbenennung*, auf die von den anderen Benennungen des selben Begriffes verwiesen wird. In elektronischen Systemen ist die Herausstellung von Vorzugsbenennungen weniger notwendig. In einem elektronischen Lexikon könnte beispielsweise unter "Apfelsine" und "Orange" der gleiche Artikel zugänglich sein und in einem Retrievalsystem können bei einer Suche nach "Orange" gleichzeitig alle Vorkommen von "Apfelsine" ermittelt werden. Schon der Übersichtlichkeit halber ist es jedoch zur Darstellung ratsam, bei Synonymen eine Benennung als Vorzugsbenennung auszuwählen und von den anderen Benennungen auf diese zu verweisen. Die kollaborativ erstellte Internetenzyklopädie Wikipedia [Wikipedia] arbeitet beispielsweise mit so genannten *Redirects*, das sind Artikel, von denen automatisch zu einem anderen Artikel weitergeleitet wird.

Welches die Vorzugsbenennung eines Begriffes ist, kann bei einer genaueren Kennzeichnung der synonymen Bezeichnungen auch vom Verwendungszusammenhang (Kontext) abhängen. Im Kontext der englischen Sprache ist beispielsweise "car" die gewählte Bezeichnung für einen Begriff, der im Deutschen "Auto" genannt wird. In jedem Fall müssen die Kontexte (z. B. die möglichen Sprachen) selber definierte Begriffe sein. Im Topic Maps Standard lassen sich dazu *Scopes* angeben, die gleichzeitig einen Namensraum aufbauen, in dem jede Bezeichnung nur einmal vorkommen darf.

Ein anderes Beispiel für kontextabhängige Bezeichnungen sind die Benennungen von Relationen. Die Instanzrelation zwischen einer Klasse A und einer ihrer Instanzen b kann sich beispielsweise sowohl als A—*class-of*→b als auch als b—*instance-of*→A ausgedrückt werden. Die konkrete Relation zwischen A und b wird im einen Fall als *class-of* und im anderen Fall als *instance-of* bezeichnet.[10] Die beiden Richtungen einer Relation werden auch als *Rollen* bezeichnet.

Kombinierte Begriffe und Mapping

Durch Kombination von einzelnen Begriffen lassen sich praktisch unendlich viele beliebig komplex zusammengesetzte Begriffe erzeugen ("Weserdampfschifffahrtskapitän", "Leiter der Poststelle des Verwaltungsamtes"). Diese Begriffe können entweder als ein zusammenhängender Begriff als Ergebnis der Kommbination einzelner Begriffe in ein Begriffssystem aufgenommen werden (*Präkombination*) oder lediglich ihre Einzelbestandteile (*Postkombination*). Der Vorteil von Präkombination liegt in der Genauigkeit, während man bei Postkombination mit deutlich weniger Begriffen auskommt, da komplexe Sachverhalte aus einfacheren Begriffen zusammensetzen werden ("Kapitän / Dampfschifffahrt / Weser"). Wie Erfahrungen mit dem Anfang der 50er Jahre von MORTIMER TAUBE entwickelten UNITERM-System gezeigt haben, hat die Reduktion auf Grundbegriffe jedoch ihre Grenzen, da die Unschärfe beim Retrieval zunimmt (beispielsweise könnte "Baum / Stamm" sowohl für Baumstamm als auch für Stammbaum stehen, da die Reihenfolge der Begriffe in einer Kombination nicht festgelegt ist). Begriffskombinationen treten häufig beim Aufstellen einer *Konkordanz* zwischen zwei Begriffssystemen aus unterschiedlichen (Fach)sprachen oder mit unterschiedlicher Genauigkeit auf. Der Abgleich verschiedener Begriffssysteme wird als *Matching* oder *Mapping* bezeichnet. DOERR und FUNDULAKI [Doerr] haben vorgeschlagen, die verschiedenen in ISO 5964 angegebene Äquivalenzbeziehungen so zu interpretieren, dass sie sich durch Boolesche Operatoren ausdrücken lassen. Dabei können drei grundlegende Fälle auftreten:

10 Es liesse sich einwenden, dass es sich bei *instance-of* und *class-of* um zwei verschiedene Relationen handele. Da die beiden Relationen immer paarweise entgegengesetzt auftreten, macht es jedoch keinen semantischen Unterschied, sie als zwei Rollen einer Relation zusammenzufassen. Näheres zur generischen Relation auf Seite 29.

1. *exact equivalence* (Maloideae = Apfelartige)
2. *broader equivalence* (Maloideae = Apfel OR Birne OR Eberesche OR Weißdorn)
3. *narrower equivalence* (Maloideae = Pflanze AND Frucht)

Wie man mit einem Blick in ein beliebiges zweisprachiges Wörterbuch feststellen kann, gibt es nur selten eine vollständige Übereinstimmung zwischen zwei Wörter in Form einer 1-zu-1-Beziehung. Dies gilt in etwas geringerem Umfang allerdings auch für die bezeichneten Begriffe. Wenn nicht eine eindeutige Übersetzung angebbar ist (1. Fall), kann ein Begriff entweder für einen oder mehrere mögliche speziellere Begriffe stehen (2. Fall) oder erst durch die Kombination mehrerer weniger speziellen Begriffe ausgedrückt werden (3. Fall, im Beispiel bieten diese nur eine sehr grobe Übereinstimmung). Zusätzlich können die Beziehungen durch Klammerung kombiniert werden. Jede Äquivalenzbeziehung lässt sich durch Umformung in eine disjunktive Normalform bringen.

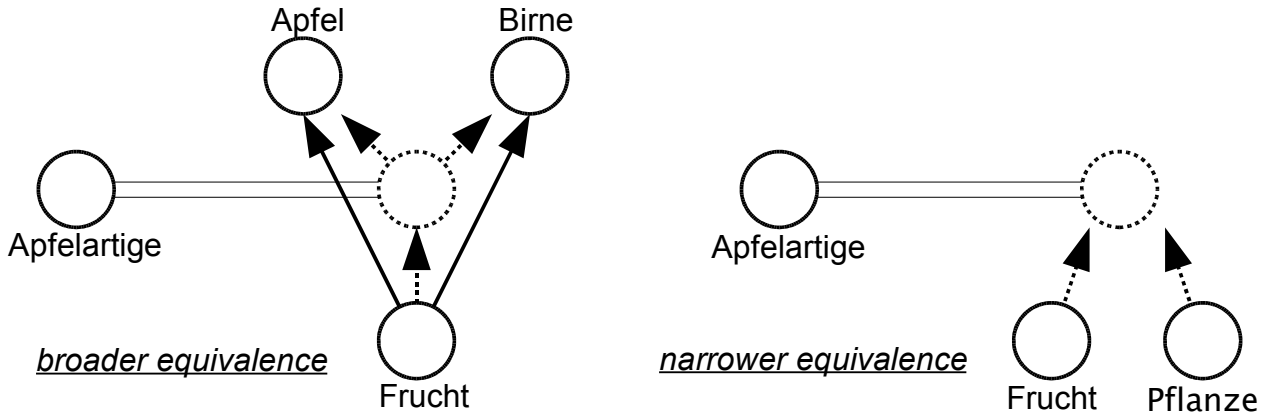


Bild 6: Neben der Übereinstimmung gibt es zwei Fälle von Äquivalenz zwischen Begriffen zweier Begriffssysteme

Ein Beispiel für die Verwendung von Begriffskombinationen ist das MACS-Projekt [MACS], in dem die Möglichkeit eines Mapping zwischen den drei großen zur Indexierung in Bibliotheken verwendeten Vokabularien *Schlagwortnormdatei* (SWD), *Library of Congress Subject Headings* (LCSH) und *Répertoire d'autorité-matière encyclopédique et alphabétique unifié* (RAMEAU) untersucht wurden. Mit Hilfe einer manuell erstellten Konkordanz lässt sich so mit einem Begriff in verschiedensprachigen Bibliothekskatalogen suchen. Auch in diesem Projekt zeigte sich, dass exakte Übereinstimmungen zwischen Begriffen in der Minderheit sind. So entspricht bspw. „Theater AND Kulturpreis“ (SWD) in den LCSH sowohl „Theatre – Financial awards“ als auch „Theatre – Non-financial awards“. Die Suche nach „Theatre – Financial awards“ wird daher im Deutschen als „Theater AND Kulturpreis“ umgesetzt (narrower equivalence) [Hoppen].

Mappings lassen sich auch durch automatische Verfahren ermitteln. Einen Überblick über Schwierigkeiten des Mappings von Thesauri gibt [Doerr_a]. Verfahren zum Mapping von Ontologien sind beispielsweise der SMART-Algorithmus [Noy] und das GLUE-System [Doan].

Auch die Verschlagwortung von Dokumenten zur sachlichen Erschließung ihres Inhalts kann als eine Art von Mapping verstanden werden – und zwar von einem gesamten Dokument auf ein begrenztes Vokabular von einzelnen Begriffen.

2.2.6. Identifikatoren

Für die interne Verarbeitung muss jeder Begriff mit einem eindeutigen Identifikator versehen sein. Das kann zum Beispiel in einer Datenbank ein ID-Feld mit einer fortlaufenden Identifikationsnummer sein. In RDF übernehmen URIs die Aufgabe der eindeutigen Identifikation von einzelnen

Elementen. Identifikatoren besitzen in der Regel keine eigene Bedeutung. Denkbar ist lediglich die Verwendung eines Hashwertes aus dem Erstellungsdatum, der Versionsnummer oder ähnlichen Eigenschaften eines Begriffes, die ab seiner Erzeugung nicht mehr verändert werden können. Wenn im Identifikator weitere Informationen kodiert werden sollen, spricht man eher von einer Notation.

2.2.7. Notationen

Eine Notation ist ebenso wie ein Identifikator eine eindeutige Bezeichnung eines Eintrags in einem Begriffssystem. Notationen können damit gleichzeitig als Identifikatoren dienen. Neben der Identifizierung eines Eintrags lassen sich jedoch mit einer Notationen weitere Informationen über den Eintrag ausdrücken. Dies geschieht in kodierter Form mit Zahlen, Buchstaben und/oder anderen Zeichen (siehe Beispiel zur RVK im Kapitel 2.1.5). Eine *Signatur* ist in der Bibliothekswissenschaft eine Notation für den Standort eines Exemplars eines Werkes innerhalb einer Bibliothek. In manchen Fällen lassen Teile der Signatur Rückschlüsse auf das Fachgebiet und den Standort, den Anfang des Autorennamens, die Auflage oder andere Statusinformationen eines Exemplars zu.

2.2.8. Relationen

Neben den sprachlichen Definitionen und Erläuterungen enthalten *Relationen* zwischen Begriffen den eigentlichen semantischen Inhalt eines Begriffssystems. Relationen sind definierte Verknüpfungen zwischen Begriffen. Analog zur Begriffsbildung müssen wichtige Relationen ermittelt und festgelegt werden. Relationen sind selber definierte Begriffe (beispielsweise muss festgelegt sein, ob es „Oberbegriff“ oder „Übergeordneter Begriff“ heißt und was die Aussage A-ist-Oberbegriff-von-B bedeutet). Sie können auch Bestandteil weiterer Relationen sein. Letzteres ist vor allem in komplexeren Ontologien und Expertensystemen mit Inferenzmechanismen der Fall.

Ohne Relationen sind Begriffssysteme nur ungeordnete Sammlungen von Objekten und Bezeichnungen. Aus einer Liste von Bezeichnungen, wie sie z. B. in Volltextsuchmaschinen durch automatische Indexierung erzeugt wird, lassen sich noch keine verlässlichen Rückschlüsse auf die mit den Bezeichnungen gemeinten Begriffe ablesen.

Umfang und Art der Relationen machen die wesentlichen Unterschiede zwischen verschiedenen Arten von Begriffssystemen aus. In den meisten Systemen sind die möglichen Relationen sehr beschränkt. Ein einfacher Thesaurus nach DIN 1463 (siehe 2.1.8) unterscheidet beispielsweise nur Äquivalenz- Hierarchische- und Assoziationsrelation während es in einer Enzyklopädie (abgesehen von den vielfältigen in natürlicher Sprache dargelegten Beziehungen) nur die Assoziationsrelation in Form von Verweisen auf andere Begriffe gibt. Zur Übersicht von möglichen Relationsarten ist eine Einteilung nach ihren Eigenschaften hilfreich. Neveling und Wersig unterscheiden in [Neveling] Relationen mit folgenden Eigenschaften (seien R und S Relationen und x,y,z Begriffe).

- Transitive Relation $\forall x, y, z : xRy \wedge yRz \Rightarrow xRz$ z. B. R=„ist Teil von“
- Reflexive Relation $\forall x : xRx$ z. B. R=„ist identisch mit“
- Symmetrische Relation $\forall x, y : xRy \Leftrightarrow yRx$ z. B. R=„ist Gegenteil von“
- Asymmetrische Relation $\forall xRy \Rightarrow \neg yRx$ z. B. R=„ist Oberbegriff von“
- Antisymmetrische Relation $\forall x, y : xRy \wedge yRx \Rightarrow xSy$ z. B. R=„ist größer als“ und S=„ist gleich groß“
- Inverse Relation (zu S) $\forall x, y : xRy \Rightarrow ySx$ z. B. R=„ist Oberbegriff von“ und S=„ist Unterbegriff von“

Auch andere mathematische Relationseigenschaften (z.B. Irreflexivität) sind möglich. Die formalen Eigenschaften von Relationen sind relationsimmanente Regeln, die sich auch mit einer Regelsprache ausdrücken und überprüfen lassen (siehe 2.2.9). Zur pragmatischen Unterscheidung von Relationen ist folgende Unterteilung der wichtigsten Relationsarten hilfreich:

- Hierarchische Relationen (Über/Unterordnung)
  - Generische Relation (Abstraktion)
    - Instanzrelation (Klasse/Instanz)
    - Vererbungsrelation (Ober/Unterklasse)
  - partitive Relation (Teil/Ganzes)
- Ordnungsrelation (Folge)
- Eigenschaftsrelation (Attribut)
- Koordinative Beziehungen (Zusammenhang)
  - Synonymie (Äquivalenz)
  - Antonymie (Gegensatz)
  - Assoziationsbeziehungen (Verwandtschaft)

Je nach Anwendung spielen einzelne Relationsarten eine größere oder geringere Rolle und ihre Unterschiede werden demnach feiner differenziert oder ignoriert. Neveling und Wersig [Neveling] unterscheiden beispielsweise zusätzlich noch die *Definitionsrelation, Instrumentelle Relation, Kausale Relation, Beeinflussungsrelation* und *Komplementärrelation*. Instanz- und Vererbungsrelation werden (wie überhaupt die verschiedenen hierarchischen Relationen) in Nachschlagewerken, die der reinen Navigation in Begriffsbeständen dienen, selten unterschieden (siehe Seite 29).

Nach der formallogischen Definition von Relationen als Prädikate können diese auch ein- oder mehrstellig sein (beispielsweise im Topic Maps Standard mit *Associations*). Diese Fälle lassen sich jedoch wie im RDF-Datenmodell (dort als *N-Tripel*) auf dreistellige Relationen der Form Subjekt-Prädikat-Objekt zurückführen. Auf diese Weise sind auch Relationen über Relationen und damit Aussagen über Aussagen möglich. Die im folgenden Beispiel aus einfachen Relationen zusammengesetzte Aussage „Nietzsche schreibt in *Die fröhliche Wissenschaft*, dass Gott tot sei“, könnte auch mit einer dreistelligen Relation „Zitat“ ausgedrückt werden, die die Bestandteile Autor (*Nietzsche*), Quelle (*Die fröhliche Wissenschaft*) und Aussage (*Gott ist tot*) enthält.

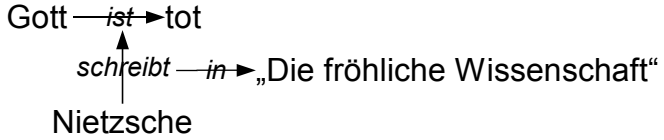


Bild 7: Komplexe Aussagen lassen sich aus einfachen zweistelligen Relationen zusammensetzen

Die einzelnen Relationsarten werden im Folgenden kurz erklärt. Außerdem werden mit Facettierung und Gruppierung zwei weitere Konzepte angesprochen, die rein formal auch als Relationen gelten, obwohl sie in der Regel nicht als solche bezeichnet werden.

Hierarchische Relationen

»…und in Holland aber auch. Vor allen Dingen aber in den Niederlanden […] Und natürlich in den gesamten Beneluxstaaten. Also in den Staaten: In Luxemburg, Belgien, Holland, Niederlanden…«
(Helge Schneider: *Ansprache* – auf der CD *New York*)

Rein formal lässt sich mit jeder gerichteten Relation (das heißt jede nicht-symmetrische Relation) eine Hierarchie aufbauen (die dabei aber auch zirkulär sein kann). Umgekehrt ist jede hierarchische Relation auch eine gerichtete Relation. Im engeren Sinne wird für eine hierarchische Relation jedoch vorausgesetzt, dass ein übergeordneter Begriff alle untergeordneten Begriffe in irgendeiner Art „enthalten" soll. Kreise sind somit ausgeschlossen, da in der Regel kein Begriff sich selbst als Teil enthalten kann. Es mag jedoch Spezialfälle geben, in denen dies nicht so ist.

Ein weiteres Kennzeichen, das in hierarchische Relationen weitaus stärker als in Ordnungsrelationen ausgeprägt ist, ist die Möglichkeit der Verzweigung. Ein Oberbegriff kann in jedem Fall mehrere Unterbegriffe und ggf. auch mehrere Oberbegriffe haben. Wenn ein Begriff lediglich einen Oberbegriff haben kann, spricht man von einer *Monohierarchie* und im anderen Fall – wenn ein Begriff mehrere Unterbegriffe besitzen kann – von einer *Polyhierarchie*. Zur Darstellung von Monohierarchien eignen sich Bäume. Damit lassen sich Monohierarchien auch wesentlich leichter überschauen als Polyhierarchien (siehe 2.1.5). In Bild 8 ist eine kleine Hierarchie dargestellt. Der untere Teil ist rein monohierarchisch (da nur jeweils ein Elternteil aufgenommen ist) und der obere Teil ist mit je zwei Vorfahren polyhierarchisch. Da die Oberen Begriffe (Vater, Mutter, Großvater, Großmutter) jeweils nur einen Nachfahren haben, ist die Darstellung noch vergleichsweise übersichtlich. In Polyhierarchien können neben der Unüberschaubarkeit weitere weitere Probleme wie das aus der objektorientierten Modellierung bekannte *Diamond Problem* auftreten (im vorliegenden Bild beispielsweise wenn Vater und Mutter von dem gleichen Großvater gezeugt worden wären). Der Mensch neigt dazu, Dinge monohierarchisch zu ordnen, das heißt zu klassifizieren. Klassifikationen sind nicht nur notwendig, weil sie leichter zu erfassen sind, sondern auch, weil sie die physikalische Ordnung von Dingen ermöglichen, die nur an genau einer einzigen Stelle aufbewahrt werden können. Für Begriffssysteme in elektronischen Medien gilt diese Beschränkung nicht mehr. Ein Hypertext muss beispielsweise nicht eine eindeutige Kapitelstruktur besitzen. Da Begriffssysteme nicht wie Bücher linear gelesen werden, bietet sich eine polyhierarchische Verlinkung an – bei einer notwendigen Visualisierung von Relationen haben allerdings auch diese Medien hinsichtlich ihrer Übersichtlichkeit ihre Grenzen im 2- oder 3-dimensionalen Raum.

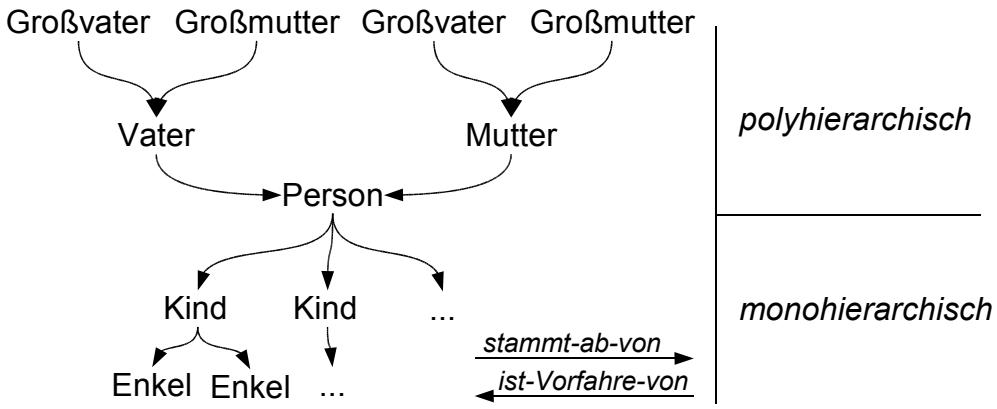


Bild 8: Abstammung als Beispiel einer hierarchischen Relation

Das Beispiel in Bild 8 zeigt übrigens auch, dass die Richtung von gerichteten Relationen relativ unbedeutend ist, weil es auf die Bedeutung der einzelnen Rollen ankommt (zu Rollen siehe 2.2.5).

Die Unterordnung kann in Bezug auf sehr viele Unterteilungskriterien (*Aspekte*) erfolgen (beispielsweise Verwaltung und Delegation, Abstraktion von Begriffen, Aufteilung in kleinere Bestandteile). Zwei häufige Fälle, unter die sich viele andere hierarchische Relationen unterordnen lassen (sic!) sind die *Generische Relation* und die *Partitive Relation* (siehe unten).

In einer Hierarchie sollten Unterbegriffe auf einer Ebene vergleichbare Begriffe sein. Ansonsten kann es bei der Betrachtung zu Fehlern wie die des eingangs zitierten Musikers und Komikers Helge Schneider kommen, der bei einer Aufzählung der von ihm erfolgreich bereisten Gebiete, die Beneluxstaaten in einer Reihe mit einzelnen Ländern nennt und beim Hinabsteigen zur Eben der einzelnen Beneluxländer in einen Zirkel gerät. Unterschiedliche Verwendungen hierarchischer Relationen (Polydimensionalität) sollten voneinander abgegrenzt werden. Diese Gruppierung geschieht durch *Aspekte* oder *Facetten* (in der Dokumentationswissenschaft wurden ursprünglich auf den Inder RANGANATHAN zurückgehende Facetten in Klassifikationen benutzt, um durch Kombination aspektbezogene Klassen zu erzeugen – bspw. mittels so genannter Anhängezahlen in der Dezimalklassifikation (UDC)). Ein Beispiel für die Unterteilung nach verschiedenen Gesichtspunkten ist Bild 9, auf dem die Bundesrepublik Deutschland (nicht vollständig) eingeteilt ist

a) nach Bundesländern (partitiv)
b) geographisch in Ost- und Westdeutschland (partitiv)
c) historisch in die BRD vor und Gesamtddeutschland nach der Wiedervereinigung (generisch).

Die im unteren Teil gepunktet dargestellten Relationen bestehen ebenso aus einer Aufteilung nach Bundesländern, wodurch diese in eine Polyhierarchie eingebunden sind.

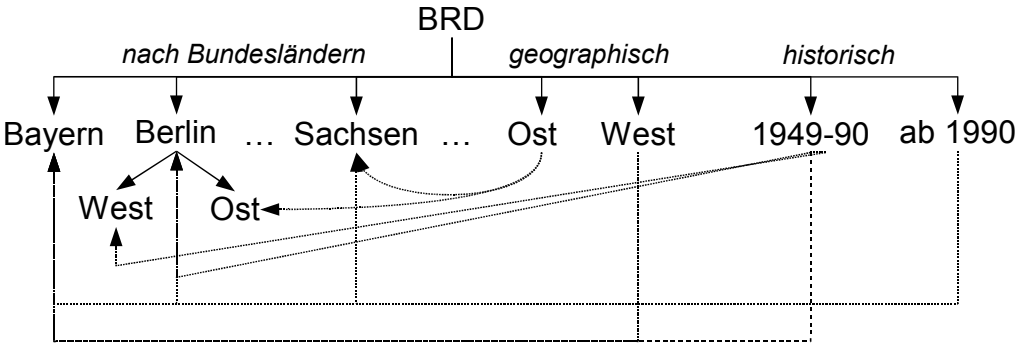


Bild 9: Hierarchische Relationen nach Aspekten unterteilt. Die gepunkteten Linien bilden zusätzliche hierarchische Unterteilungen (Polyhierarchie) nach Gebieten (partitiv).

Generische Relation

Die Generische Relation (auch *Abstraktionsrelation*) ist eine hierarchische Relation zwischen zwei Begriffen, von denen der untergeordnete Begriff alle Merkmale des übergeordnete Begriffes übernimmt und mindestens ein zusätzliches Merkmal aufnimmt bzw. verändert. In der Informatik spricht man bei diesem Vorgang von Vererbung oder Instanziierung. Die Unterscheidung zwischen Vererbungs- und Instanzrelation ist nur mit einem dahinter stehenden Klassenkonzept und den darin vorkommenden Vererbungsregeln sinnvoll. Dabei werden Eigenschaften (*Attribute*) (auch ohne Belegung mit konkreten Werten) an von einer übergeordneten Klasse (*Gattung*) an Unterklassen (*Art*) weitervererbt und erst in den Instanzen endgültig mit konkreten Werten (*Ausprägungen*) belegt. Ein Kreis – zum Beispiel auf einem Blatt Papier – hat als Instanz der Klasse *Kreis* einen

konkreten *Durchmesser*. Die Klasse *Kreis* lässt sich als Unterklasse von *Körper* auffassen, die als eine Eigenschaft von ihm die *Ausdehnung* erbt – in der Instanz eines konkreten Kreises wird diese *Ausdehnung* durch die Eigenschaft des *Durchmessers* erfüllt. Diese, aus der Ideenlehre von PLATON stammende und in der Objektorientierten Modellierung umgesetzte Vorstellung lässt sich jedoch nicht in jedem Begriffssystem so einfach durchsetzen wie angenommen.

Zum einen muss eine klare Übereinkunft darüber bestehen, was Eigenschaften sind und wie diese vererbt werden. Wenn aus einer generischen Relation automatisch Schlüsse gezogen werden können sollen, müssen diese *Vererbungsregeln* strikt eingehalten werden.[11] Da dies nur durch Regeln und Inferenz gewährleistet werden kann, ist ohne diese Mechanismen eine Unterscheidung zwischen Vererbungs- und Instanzrelation zwar für die Begriffsbildung hilfreich, aber nicht notwendig.

Zum anderen hängt die Unterscheidung von Klasse und Instanz vom Kontext ab. Ein Kreis als geometrische Form kann beispielsweise auch als Instanz behandelt werden, wenn in einem System keine konkreten Kreise vorkommen. Die Unterscheidung zwischen Klasse und Instanz ist somit eher eine pragmatische, bei der die Bezeichnung „Instanz“ andeutet, dass es sich um ein Objekt handelt, von dem keine weiteren Objekte abgeleitet werden. Die Instanzrelation ist somit einfach eine nicht-fortgeführte generische Relation, d. h. jeder Begriff ohne generische Unterbegriffe kann als Instanz betrachtet werden.

Partitive Relation

Die Partitive Relation (auch „Teil-Ganzes-Relation“, *Bestandsrelation*, *Aggregation*) wird verwendet, wenn ein Begriff in Einzelbestandteile zerlegt werden kann. Dies tritt beispielsweise im Bereich materieller Objekte und bei hierarchischen Verwaltungsstrukturen auf (siehe Bild 9). In der Regel dürfen sich die einzelnen Teile nicht überschneiden, so dass zwischen ihnen eine konträre Gegensatzbeziehung (Antonymie) festgestellt werden werden kann (siehe unten).

Eigenschaften

Die Eigenschaftsrelation (*Attributive Relation*) ist eine asymmetrische Relation, die einem Objekt ein Attribut zuordnet. Ein Beispiel ist die Benennung von Begriffen. Die Eigenschaftsrelation ist die grundlegende Relation in (relationalen) Datenbanken, wo die Attribute aus den einzelnen Feldern eines Tabelleneintrages bestehen.

Ordnungen

Ordnungs- oder Folgerelationen gibt es überall dort, wo Begriffe in Reihenfolge gebracht oder miteinander verglichen werden. Beispiele sind die alphabetische Ordnung von Begriffen nach ihren Bezeichnungen oder der Ablauf der Wochentage. Wie letzteres Beispiel zeigt, können Ordnungen auch zirkulär sein (auf Sonntag folgt wieder Montag). Ein Begriff kann auch mehrere Vorgänger oder Nachfolger gleichzeitig haben (z. B. beim Durchlaufen einer Hierarchie oder bei Abspaltungen und Zusammenführungen einer Entwicklungsreihe). Die wesentliche Eigenschaft einer Ordnungsrelation ist ihre Gerichtetheit. Da alle nicht-symmetrischen Relationen gerichtet sind, können viele Relationen als Folge aufgefasst werden (zum Beispiel alle hierarchischen Beziehungen). Ein Merkmal von Ordnungen ist auch, dass mit ihnen prinzipiell beliebig viele gleichrangige Begriffe in Reihe verbunden werden können (dies unterscheidet sie beispielsweise von der Eigenschaftsrelation). Dass die Richtung von Ordnungen mitunter willkürlich, ist zeigt das Beispiel in Bild 8/S. 28.

11 Zur Einführung von Vererbungsregeln in RDF siehe den kurz vor seiner Verabschiedung stehenden Working Draft *RDF Semantics* des W3C: http://www.w3.org/TR/rdf-mt/ vom 23.1.2003

Synonymie

Mittels Synonymie oder Äquivalenzrelation werden Begriffe verknüpft, die innerhalb eines bestimmten Kontextes als gleich angesehen werden. Neben der exakten Äquivalenz (Identität), die innerhalb eines Begriffssystems eigentlich nur auf der syntaktischen Ebene auftritt, kann es verschiedene Arten von Synonymie geben (Näheres unter 2.2.5).

Antonymie

Die Antonymie/Gegensatzrelation ist eine symmetrische Relation. In der philosophischen Logik unterscheidet man drei Fälle.[12] Einige Gegensatzpaare können auch mehreren davon angehören.

a) **Kontradiktorischer Gegensatz (Komplement):** Das Komplement einer Aussage oder eines Urteils in Form einer logische Verneinung. Für den kontradiktorischen Gegensatz gilt in der zweiwertigen Logik der Satz vom ausgeschlossenen Dritten, das heisst genau einer von zwei antonymen Begriffen ist wahr und der andere unwahr – eine andere Möglickeit gibt es nicht.
b) **Konträrer Gegensatz (Antonymie):** Die Gegenüberstellung von gleichartigen Begriffen einer Klasse (zum Beispiel Schwarz/Weiss, Apfel/Birne…). Im Gegensatz zum kontradiktorischen Gegensatz können auch mehr als zwei Begriffe konträr einander gegenübergestellt werden (zum Beispiel die drei Grundfarben Rot, Grün und Blau).
c) **Polarer Gegensatz:** Verhältnis zweier entgegengesetzter aber zusammenhängender Teile eines Ganzen (zum Beispiel Nord- und Südpol, Positiv und Negativ, Vorder- und Rückseite)

Antonymien in ihrer Form als konträre Gegensätze lassen sich anhand beliebiger Merkmale zwischen relativ vielen Begriffen ziehen und lassen sich alleine nicht für logische Schlussfolgerungen benutzen. In der Regel wird der Gegensatz anhand eines bestimmten Merkmals festgemacht (zum Beispiel zwischen „Mann“ und „Frau“ das „Geschlecht“). In Klassifikationen stehen die Unterteilungen einer Oberklasse oft in einem konträren Verhältnis, sofern sich die Unterklassen gegenseitig ausschließen, so dass ein Begriff nur in eine von ihnen eingeordnet werden dürfen. Dabei spricht man jedoch eher von *Nebenordnung*. Die Antonymie ist ein häufiger Spezialfall der Assoziationsrelation.

Assoziation

Mit einer Assoziationsrelation („siehe-auch-Relation“) werden Begriffe miteinander verbunden, die als irgendwie miteinander zusammenhängend ausgewiesen werden sollen. In vielen Systemen ist diese Relation symmetrisch. Mit ihr werden Verwandtschaftsbeziehungen zwischen Begriffe, die „immer zusammen gedacht werden sollen“, ausgedrückt. Bei der Erstellung eines Begriffssystems sind zunächst die meisten Relationen assoziativ. Auch automatische Verfahren können selten mehr als einfache Assoziationen zwischen Begriffen ermitteln. Viele Assoziationen lassen sich bei genauerem Betrachten näher bestimmen, indem speziellere Relationen eingeführt werden; in rein psychologisch begründeten Fällen (zum Beispiel „Liebe“ und „Rot“) kann jedoch nicht mehr gesagt werden, als dass die Begriffe irgendwie als zusammengehörig empfunden werden. Da das Menschliche Denken stark assoziativ abläuft, lassen sich bestimmte Zusammenhänge nur schwer formalisieren. In Nachschlagewerken, die für Menschen gedacht sind, ist dies weniger ein Problem als wenn Beziehungsnetze mit Methoden der Künstlichen Intelligenz ausgewertet werden sollen. Für einige Zwecke liefern jedoch auch rein statistische Methoden gute Ergebnisse, so dass auf die Semantik verzichtet werden kann (beispielweise Webseitenrankings anhand ihres Verlinkungsgrades oder Hinweise auf verwandte Produkte, die von Kunden gemeinsam erworben wurden).

12 Quelle: *Meyers großes Taschenlexikon*, 7.Auflage, 1999 (unter „Gegensatz“)

Gruppierung

Neben den verschiedenen Arten von Relationen zwischen Begriffen gibt es in vielen Begriffssystemen durch Gruppierungen ein weiteres Strukturierungsmittel. Durch Gruppierung werden ähnlich wie bei der partitiven Relation Begriffe zu einem neuen Gesamtobjekt zusammengefasst. Beispielsweise bilden auf der Ebene des gesamten Werkes alle Begriffe eines Begriffssystems ein neues Objekt – und zwar das gesamte Begriffssystem selbst. Auch darunter können Gruppierungen auftreten, beispielsweise die einzelnen Kapitel und Unterkapitel eines Dokumentes. Die Gruppierung findet in der Regel auf einer höheren Ebene als die (hierarchische) Verknüpfung von einzelnen Begriffen statt. So sind in einem enzyklopädischen Lexikon die Einträge normalerweise nach Anfangsbuchstaben zusammengefasst. Möglicherweise gibt es zwar einen Eintrag zum Buchstaben ‘A’; dieser Eintrag ist jedoch nicht der Oberbegriff aller Einträge mit dem Anfangsbuchstaben ‘A’, sondern Teil dieser Gruppe von Einträgen. Auch ein Verweis auf “Band 7” zeigt nicht auf einen unter ‘B’ stehenden Lexikoneintrag namens “Band 7”. Gleichwohl lassen sich ‘A’ und “Band 7” als einzelne Konzepte und damit Begriffe auf einer höheren Ebene verstehen.

Prinzipiell lässt sich die Gruppierung auch als Spezialfall einer hierarchischen Relation auffassen. In einer Ontologie liesse sich dazu eine Relation namens “definiert-innerhalb-von” verwenden, die alle Bestandteile der Ontologie mit der Ontologie selber verknüpft.

Die Aufteilung in Gruppen ist in der Regel streng monohierarchisch. Außerdem sollte der Übersichtlichkeit halber klar zwischen Gruppen und Einzelbegriffen unterschieden werden.

2.2.9. Regeln

»Und er kommt zu dem Ergebnis:
“Nur ein Traum war das Erlebnis.
Weil”, so schließt er messerscharf,
“nicht sein kann, was nicht sein darf.”«

(Christian Morgenstern, Die unmögliche Tatsache)

In einem Begriffssystem sind Regeln Vereinbarungen dessen Struktur. Sie bestimmen, welche Arten von Begriffen, Bezeichnungen und Relationen zwischen ihnen bestehen können oder müssen. Ein Beispiel für eine solche Regel ist die Forderung, dass Begriffe eindeutige Bezeichnungen haben müssen oder dass ein Begriff A genau dann Unterbegriff von B ist, wenn B Oberbegriff von A ist. Es sind aber auch komplexe Anforderungen möglich. Aufgrund von Regeln können Inferenzmaschinen durch Schlussfolgerung weitere Informationen aus einem Begriffssystem ableiten. Dies ist bspw. in Expertensystemen der Fall, deren Aufgabe im *Semantic Web* selbständig agierende Computerprogramme (*Agenten*) übernehmen sollen. Während herkömmliche Begriffssysteme mit einem einfachen Satz von festen Regeln auskommen, können in Ontologien Regeln aus beliebigen logische Aussagen über Begriffe und Relationen bestehen und selbst Teil einer Ontologie sein. In diesem Fall sind Regeln selbst wieder Begriffe, die ein eigenes Begriffssystem bilden.

Dazu müssen sie in einer formalen Regelsprache formuliert sein. Es existieren eine Vielzahl dieser zum Teil konkurrierenden Sprachen (u.a. Prolog, KL-ONE, Loom, F-Logic, CyCML, Ontolingua, KIF, CGIF, O-Telos...). In der von Tim-Berners Lee vorgeschlagenen Architektur des Semantic Web (Bild 5/ S. 19) baut die Regel der Ebene (*Logic*) auf die Ebene der Terminologien (*Ontology vocabulary*) auf. Dies ist in so fern problematisch, als dass Ontologien auch ohne Regeln von ihrer Struktur her schon (implizite) Regeln enthalten. Die meisten Regelsprachen, wie beispielsweise

RuleML[13], eine sich noch in der Entwicklung befindliche Sprache, die in XML und/oder RDF formuliert werden kann und als mögliche Sprache für den Einsatz im Semantic Web behandelt wird, bauen deshalb auf die Ebene der Daten anstatt auf Ontologien auf. Die vom W3C favorisierte aus den Sprachen DAML und OIL entwickelte *Web Ontology Language* (OWL)[14] beinhaltet auch die Möglichkeit von Regeln. Allerdings werden diese mit Mitteln der Beschreibungslogik formuliert, was nicht für alle Anwendungen praktikabel ist (vgl. VOLZ, DECKER und OBERLE, die dies in [Volz] kritisieren und einen Ansatz zur Integration von OWL und Logiksprachen vorstellen). Es bleibt also abzuwarten, in wie weit sich einzelne Regelsprachen breiter durchsetzen werden. Da Regeln aus ganz unterschiedlichen logischen Paradigmen formuliert werden können (Aussagenlogik, Prädikatenlogik, Modallogik, Frame-basierte Systeme, Fuzzy Logik...) ist es fraglich, ob das Konzept einer einzigen Regelsprache überhaupt möglich und sinnvoll ist. Überdies gelten für Regeln, die selbst Bestandteil eines Begriffssystems sein sollen (z. B. in einer RDF-basierten Regelsprache formulierte Regeln für RDF-Daten) die gleichen Beschränkungen wie für in herkömmlichen Programmiersprachen formulierte Programme: und zwar (neben anderen, komplexitätstheoretischen Grenzen der Berechenbarkeit) das aus der theoretischen Informatik bekannte Halteproblem, nach dem es kein Programm geben kann, das bei Eingabe des Programmcodes eines beliebigen anderen Programms entscheiden kann, ob dieses Programm stoppt oder nicht.

Auch aus diesem Grund werden Regeln eher auf einer Meta-Ebene formuliert — zum Beispiel in Form von Algorithmen zur Verarbeitung des Begriffssystems oder in umgangssprachlichen Normen und Standards. In der Praxis beziehen sich die meisten Regeln sowieso lediglich auf einzelne Relationen oder sie legen Wertebereiche (Datentypen) fest. Relationsimmanente Regeln (siehe 2.2.8) sind implizit in Relationsarten enthalten und verlangen, dass bestimmte Eigenschaften (Symmetrie, Transitivität, etc.) eingehalten werden. In (objektorientierten) Klassensystemen werden Regeln für die Vererbung und Instanziierung vorausgesetzt. Ausformuliert gilt für die Vererbungsrelation

$$C \text{ isSuperclass } D \wedge C \text{ hasProperty } P \Rightarrow D \text{ hasProperty } P$$

und für die Instanzrelation

$$c \text{ instanceOf } C \wedge C \text{ hasProperty } P \Rightarrow \exists\, p : p \text{ instanceOf } O \wedge c \text{ hasProperty } p$$

In Worten ausgedrückt bedeutet dies, dass eine Klasse alle Eigenschaften von ihrer Oberklasse erbt und dass bei Instanzen alle Eigenschaften mit konkreten Werten belegt werden müssen.

Ein grundsätzliches Problem von Regeln ist, dass sie sich wesentlich schwieriger veranschaulichen und begreifbar machen lassen als Begriffe und Relationen (man vergleiche beispielsweise Bild 7/S. 27 mit den eben genannten Formeln oder denen auf S. 26). Auch Änderungen, die im Verlaufe der Anwendung einer Ontologie regelmäßig notwendig sind, sind bei Regeln aufgrund ihrer Komplexität in ihren Auswirkungen schlecht überschaubar.

Zudem lässt sich fragen, ob Regeln überhaupt immer formalisierbar sind. Viele Regeln entspringen sozialen Übereinkünften, die sich nicht in strenge Vorschriften pressen lassen („Ausnahmen bestätigen die Regel“). Das Ignorieren dieser Tatsache führt zu bürokratischen Vorschriften, die zu teilweise widersinnigen Schlussfolgerungen führen können. Aus der Objektorientierten Modellierung ist beispielsweise das so genannte *Diamond Problem* bekannt, dass auftreten kann, wenn eine Eigenschaft von der selben Basisklasse auf unterschiedlichen Pfaden mehrfach geerbt wird.

Grundsätzlich hat die Kritik an der so genannten starken KI sicher auch im Bereich des Semantic Web ihre Berechtigung und für jedes Begriffssystem gilt, dass es nur eine mögliche Sicht der Welt erzeugt, die andere Gesichtspunkte ausblendet (siehe [Bowker]).

13 http://www.dfki.uni-kl.de/ruleml/ (RuleML Design in der Version 0.8 vom 3. September 2002)
14 http://www.w3.org/TR/owl-features/ (W3C Working Draft vom 31. Mai 2003)

2.2.10. Quellennachweise und Literatur

Quellennachweise können in Begriffssystemen zwei verschiedene Zwecke erfüllen: Zum einen dienen sie wie in wissenschaftlichen Publikationen als Beleg für die angeführten Inhalte. Dies soll der Überprüfbarkeit und als Verweis auf Hintergründe, genauere Ausführungen und verwandte Gebiete dienen. Diese Rolle spielen Quellennachweise zum Beispiel in Lexika und Fachwörterbüchern, die eine Zusammenstellung von direkter Information und Nachweisen bilden. In weiten Teilen des modernen Wissenschaftssystems dienen Literaturverweise als Zitierungen auch als eine (nicht unumstrittene!) Methode zur Messung von wissenschaftlicher Leistung über einen so genannten *Citation Index*. Solche Begriffssysteme (Referenz- und Bibliographiedatenbanken) bestehen sogar hauptsächlich aus Literaturverweisen.

Zum anderen ist die Angabe einer Quelle zugleich Verweis auf ihren Verwendungszusammenhang. Da fast alle Begriffe kontextgebunden sind (zum Beispiel Fachbegriffe im Kontext ihres Fachgebietes), ist die Quellenangabe nicht reiner Zusatz sondern auch *Bestandteil* der Definition eines Begriffes. Viele philosophischen Begriffe sind beispielsweise erst im Zusammenhang des Werkes eines bestimmten Philosophen verstehbar. Quellenangaben dienen also auch als Angabe eines Kontextes.Oft überschneidet sich sich diese Funktion mit der Angabe eines Homonymzusatzes, um die Bezeichnung eines Begriffes eindeutig zu machen (vgl. 2.2.3).

Häufig werden zur Erklärung eines Begriffes etymologische Herleitungen seiner Bezeichnung herangezogen. In der Regel haben sie jedoch eher erläuternden Charakter oder für Mnemotechniken.

2.2.11. Natürlichsprachliche Bestandteile

Natürlichsprachliche Bestandteile sowie Bilder und andere Medien können sowohl konstitutiven als auch rein erläuternden Charakter haben. In einem Lexikon liegt fast der gesamte Inhalt natürlichsprachlich vor, während sich die Begriffe in einem einfachen Thesauri oder einer Ontologie mehr durch die mit ihren Bezeichnungen bei Benutzern assoziierten Bedeutungen und aus dem Zusammenhang ihrer Relationen ergeben. Zusätzliche Texte haben lediglich erläuternden Charakter und sind meist nur in Form von kurzen Kommentare (zum Beispiel als so genannte *Scope Notes* in einem Thesaurus) möglich. Die Texte bestehen dabei aus einfachen Zeichenketten ohne weitere Textformatierungen. In anderen Nachschlagewerken können natürlichsprachliche Erläuterungen und Definitionen auch zusätzliche Textelemente und Formatierungen, wie Listen, Abbildungen und Tabellen, beinhalten. Die Gestaltung von Texten hängt zwar auch von ästhetischen Gesichtspunkten ab, zumindest in Sachtexten sollten sie jedoch als Mittel zum Zweck die inhaltlichen Aussagen eines Textes unterstützen. Es sei darauf hingewiesen, dass die hier sprachlich genannten Zusätze zu einzelnen Begriffen in einem Begriffssystem auch aus anderen Medien wie Bilder, Filem und Töne bestehen bzw. diese enthalten können können.

Die zusätzlichen (natürlichsprachlichen) Bestandteile eines Begriffssystems, können unterschiedliche Zwecke erfüllen. Dazu gehören unter anderem Kommentare, Definitionen, Beispiele und Erklärungen. Auch zusätzliche Angaben zur Verwaltung (beispielsweise das Datum der letzten Änderung) sind möglich. Damit nicht der eine Begriff einen „Kommentar“ enthält, der zweite eine „Anmerkung“ und der dritte einen Text, der gar nicht weiter gekennzeichnet ist, sollten die unterschiedlichen Arten der textuellen Bestandteile aus einem kontrollierten Vokabular entstammen. Vor allem wenn mehrere verschiedene Anmerkungen, Illustrationen, Definitionen, Erklärungen etc. zu einem Begriff möglich oder notwendig sind, ist ihre Typisierung notwendig.

2.3. Darstellung und Benutzung von Begriffsystemen

Neben Inhalt und Struktur bildet die Darstellung und der Umgang mit einem Begriffssystem (das heisst die *Benutzerschnittstelle*) das Kriterium für seinen Zweck und Nutzen in der Praxis. Die Darstellung kann unter anderem rein textlich, tabellarisch oder grafisch, gedruckt oder als Hypertext bis hin zu dreidimensionalen virtuellen Räumen erfolgen. Die Navigation und Suche kann mittels Suchanfragen oder über Browsing-Strukturen in alphabetischen und systematischen Verzeichnissen oder interaktiv erfolgen, indem der Benutzer geführt wird und Dialoge dynamisch an seinen Wissensstand- und Bedarf angepasst werden – bis hin zu Agenten, die selbständig in Begriffssystemen navigieren und so gewünschte Informationen aus extrahieren – der Fantasie sind keine Grenzen gesetzt.

Die Art der Darstellung hängt natürlich neben der Art des Begriffssystems auch stark vom Medium (Papier gebunden oder elektronisch) vom Umfang und von ergonomisch-ästhetischen Aspekten ab. Bei vielen nicht selten in größeren Arbeitsabläufen eingebetteten Systemen ist es sogar angebracht, sich *zuerst* den Fragen der Darstellung zu widmen und *danach* mit der Klärung von Begriffen, Benennungen und Relationen zu befassen. In jedem Fall muss im ersten Schritt geklärt werden, wer die Zielgruppe eines Begriffssystems ist und wie die einzelnen Anwender damit umgehen bzw. was sie damit machen wollen. Diese Frage, die im Falle eines Nachschlagewerkes noch einfach zu beantworten ist, macht im Falle von Ontologien und Datenbanken möglicherweise den eigentlichen Kern eines Systems aus. Auch gilt die einfache Regel, dass eine schlechte Einbindung in Benutzeroberflächen und andere Programme den Nutzen einen noch so gut durchdachten Begriffssystems bis zur Nutzlosigkeit schmälern kann.

Die Möglichkeiten der Darstellungen von und Interaktion mit Begriffssystemen bietet sicherlich ausreichend Stoff für weitere Arbeiten. Sowohl zur *Informationsvisualisierung* als auch zum Design von Benutzerschnittstellen existiert umfangreiche Literatur.

Die grundlegenden Darstellungsarten, die sich in der einen oder anderen Form in vielen Systemen wiederfinden, ergeben sich aus den in Kapitel 2.1 beschriebenen Formen von Begriffssystemen:

1. Nachschlagewerke: *(Hyper)text*
2. Klassifikationen: *Bäume*
3. Register: *Listen*
4. Thesauri und Semantische Netze: *Graphen*
5. Ontologien: *Interaktive Suchsysteme*

3. Datenformate und -modelle

Zur formalen Beschreibung von Begriffssystemen existiert eine Vielzahl an Datenformaten je nach Anwendungsgebiet. Insgesamt ist in den letzten Jahren eine allgemeine Hinwendung zur Verwendung von XML festzustellen – die konkreten Formate sind jedoch, je nachdem, aus welcher Fachdisziplin sie stammen und für welche Aufgabe sie eingesetzt werden, sehr unterschiedlich ausgeprägt und nicht aufeinander abgestimmt. Im Zuge der weltweiten Vernetzung von Informationen wird der Austausch von Ontologien und anderen Begriffssystemen eine zunehmende Rolle spielen. Nicht zuletzt im Rahmen des so genannten genannten *Semantic Web* (siehe 2.1.10) ist deshalb eine verbesserte Interoperabilität wünschenswert.

Mit dem unter 3.1.4 vorgestellten RDF-Modell und darauf aufbauenden Ontologiesprachen sind bereits Standards geschaffen worden, die ein einheitliches Datenformat darstellen sollen. Dabei handelt es sich jedoch um mehrheitlich theoretische Arbeiten, die nur zum Teil auf den bereits existierenden Begriffssystemen und ihren Anwendungen und Datenformaten aufbauen.

Die theoretische Betrachtung verschiedener Arten von Begriffssystem und ihrer Bestandteile im vorangegangenen Kapitel hat gezeigt, dass es trotz aller Unterschiede gemeinsame Strukturen gibt. In diesem Kapitels wird ein integrierendes Datenmodell vorgestellt (3.2), dass die wesentlichen Bestandteile und Konzepte verschiedener Begriffssysteme abbildet. Die Arbeit beinhaltet eine Repräsentation des Datenmodells mittels einer XML-DTD (3.3). Zunächst soll jedoch eine Übersicht über die bereits existierenden existierenden Datenformate gegeben werden (3.1).

3.1. Bestehende Datenformate

Die folgende Übersicht von existierenden Datenformaten für verschiedene Arten von Begriffssystemen bzw. ihre Bestandteile kann bei der sich ständig vergrößernden Vielfalt von Formaten naturgemäß nicht vollständig sein. Sie ist deshalb nicht als Nachschlagewerk sondern als Hinweis auf wesentliche Formate und Konzepte zu verstehen. Auf der technischen Seite liegt der Schwerpunkt bei offenen XML-basierten Formaten, die die freie Weiterverarbeitung von Daten erleichtern. Auf der inhaltlichen Seite wird zunächst auf Textformate eingegangen (3.1.1), da diese bei der Betrachtung von Begriffssystemen oft vernachlässigt werden. Anschließend werden verschiedene Formate für Thesauri (3.1.2) und Terminologien (3.1.3) genannt. Etwas ausführlicher wird danach auf RDF (3.1.4) und Topic Maps (3.1.5) eingegangen. Nicht näher eingegangen wird an dieser Stelle auf konkrete Metadatenformate (Dublin Core, BibTeX, MARC, MAB, METS, ONIX etc.), deren Betrachtung für die Erstellung vieler Begriffssysteme auch hilfreich sein kann.

3.1.1. Texte und Textformate

Die meisten Informationen in und über einzelne Begriffssysteme liegen nach wie vor als natürlichsprachlicher Text vor. Da Menschen Texte nur ungern in formal logischen Formeln schreiben, wird dies auf absehbare Zeit auch so bleiben. Obwohl formale Begriffssysteme gerade ein Versuch sind, die natürliche Sprache durch eine künstliche Syntax und Semantik zu ersetzen, sind zusätzliche Erklärungen, Definitionen, Anmerkungen und Beispiele zumindest für den Menschen unerlässlich. Die meisten Datenformate und Sprachen zur Beschreibung von Begriffssystemen bieten deshalb die Möglichkeit, Kommentare und Anmerkungen zu einzelnen Elementen hinzuzufügen. Allerdings bestehen diese in der Regel nur aus einfachen Zeichenketten ohne weitere Formatierungen und sind auch nicht integraler Bestandteil des Begriffssystems.

Während die meiste Typographie und Gestaltung lediglich unterstützenden Charakter hat und somit oft als nicht Bedeutung tragendes Element eines Textes ignoriert werden kann, gehören einige

Auszeichnungen wie Hervorhebungen, Absätze, Listen, Zwischenüberschriften, Tabellen und Bilder als wesentlicher Bestandteil zu einem Text dazu.

Im folgenden werden einige Formate für Dokumente und Texte vorgestellt, die gleichzeitig als Teil eines allgemeinen Formates für ein Begriffssystem dienen können. Die Auswahl beschränkt sich auf die XML-basierten Formate XHTML, DocBook, TEI und DiML. Die nicht auf SGML/XML basierenden Formate sind entweder Ausgabeformate (PDF), oder proprietär (Word), so dass eine sinnvolle automatische Weiterverarbeitung der Texte nicht praktikabel ist. Dies gilt aufgrund seiner Komplexität (verschiedene Pakete, eigene Makros u.ä.) auch für das in einigen wissenschaftlichen Disziplinen stark verbreitete Satzprogramm LaTeX.

Neben den hier genannten Formaten sei auf das XML-basierte Dateiformat des freien Office-Paketes *OpenOffice* hingewiesen. Dieses eignet sich hervorragend als Zwischenschritt bei der Bearbeitung von Texten, die in ein spezielleres XML-Format überführt werden sollen. Weitere XML-Textformate sind unter anderem ISO 12083 und das *News Industry Text Format* (NITF). Beide finden jedoch nur in begrenzten Bereichen Verwendung.

XHTML

Die *Hypertext Markup Language* (HTML) wurde ursprünglich von Tim Berners Lee als Internet-format für Hypertexte im *World Wide Web* entwickelt. Eine einheitliche, syntaktisch korrekte SGML-Anwendung von HTML wurde aus verschiedenen Gründen nie richtig implementiert. Stattdesen wurde das Format von Browserherstellern und dem W3C beständig erweitert und entwickelte sich immer mehr zu einer umfangreichen visuellen Auszeichnungssprache für Webseiten. Inzwischen existiert mit XHTML eine XML-konforme Version und durch *Cascading Stylesheets* (CSS) lassen sich viele rein optischen Auszeichnungen aus dem HTML-Code heraushalten. HTML ist allerdings stark auf die Bildschirmpräsentation elektronischer Medien ausgerichtet. Die Auszeichnungssprache eignet sich daher nur beschränkt als Austauschformat, sondern ist eher als Ausgabeformat. Viele einfache Elemente aus HTML finden sich auch in anderen XML-Formaten.

DocBook

Das vor allem innerhalb der Informatik bekannte für SGML und XML in einer Document Type Definition (DTD) festgelegte Format *DocBook* ist ein offener Standard, der unter der Leitung von Norman Walsh von der *Organization for the Advancement of Structured Information Standards* (OASIS) gepflegt wird. Es eignet sich besonders zur Erstellung von Büchern und Artikeln im technischen Umfeld wie beispielsweise zur Softwaredokumentation. Da DocBook viele Elemente speziell für diesen Zweck enthält, ist das Format für allgemeine Texte etwas unübersichtlich und überladen. Mit *Simplified DocBook* existiert eine vereinfachte Version der DTD.

TEI

Die *Text Encoding Initiative* (TEI) ist eine 1987 gegründete Organisation (das *TEI Konsortium*) und gleichzeitig ein Dokumentenformat für die Kodierung und den Austausch von Texten, das die TEI entwickelt hat und weiterentwickelt. Das Format basiert auf SGML (inzwischen XML) und ist in einer DTD festgelegt (siehe http://www.tei-c.org/).

TEI hat sich zu einem de-facto-Standard innerhalb der Geisteswissenschaften entwickelt, wo es Geisteswissenschaften) z.B. zur Kodierung von gedruckten Werken (Editionswissenschaft) und zur Auszeichnung von sprachlichen Informationen (Linguistik) in Texten verwendet wird.

Die TEI-Version P3 wurde 1994 verabschiedet und ist 2002 durch die XML-Version P4 abgelöst worden. Gleichzeitig gibt es *TEI lite* mit einem verringertem Elementumfang. Die TEI-DTD ist aus

verschiedenen sachbezogenen Modulen aufgebaut, die beispielsweise Elemente für die Struktur eines Buches, zur Auszeichnung von Gedichten und Dramen, zur Markierung einzelner Zeilen und Seiten, für Tabellen, für kritische Anmerkungen und auch für Sprachkorpora, Terminologien und Wörterbücher enthalten.

Das Wörterbuch-Modul von TEI wird als ein Standard für wie die Erstellung von Wörterbüchern eingesetzt – beispielsweise die elektronische Ausgabe des Deutschen Wörterbuch von Jacob und Wilhelm Grimm (http://www.dwb.uni-trier.de/). Elektronische oder gedruckte Wörterbücher lassen sich mit eigenen Programmen wie dem an der Universität Trier entwickeltem SGML-Text und Satzprogramm TUSTEP oder mit Hilfe von XSLT aus den TEI-Quellen erzeugen.

DiML

Die *Dissertation Markup Language* (DiML) ist ein XML-basiertes Dokumentenformat zur Archivierung elektronischer Publikationen – speziell Dissertationen und Habilitationen – das am Rechenzentrum der Humboldt-Universität zu Berlin entwickelt wurde und seit 1997 (SGML) bzw. 2003 (XML) eingesetzt wird. Das Format basiert auf der 1987 von Juri Rubinski vorgelegten *Dissertations DTD* (ETD), einem Dokumentenformat (DTD) in SGML. Aus einer überarbeiteten Version entstand 1997 die *Dissertation Markup Language* (DiML). 1998 wurde die Prüfungsordnung der Humboldt Universität dahingehend geändert, dass Dissertationen und Habilitationen auch als elektronische Publikationen abgegeben werden können, sofern dafür eine spezielle Dokumentenvorlage verwendet wird. Mit Hilfe der in dieser enthaltenen Formatvorlagen können Dokumente nach SGML bzw. XML umgewandelt werden. Für die Verwendung der Dokumentenvorlage werden regelmäßig Autorenschulungen durchgeführt.

Die Entwicklung von DiML fand unter anderem im Rahmen des DFG-Projektes *Dissertationen Online* statt. Die SGML-DTD wurde 2002/03 weiter überarbeitet und daraus eine XML-DTD entwickelt. Diese ist aus verschiedenen Modulen aufgebaut, aus denen sich je nach Bedürfnis angepasste Dokumentenformate zusammenstellen lassen. Beispielsweise existieren eigene Module für einfache Textauszeichnungen, für Listen, für Tabellen, für Zitationen und für mathematische Formeln in MathML. Näheres siehe http://edoc.hu-berlin.de/diml/.

Die Konvertierung von Word-Dokumenten in das DiML-Dokumentenformat wird mit Hilfe des freien Office-Paketes *OpenOffice* und einigen XSLT-Skripten vollzogen. Zur Weiterverarbeitung von DiML-Dokumenten z. B. zur Darstellung in HTML existieren weitere Skripte (*diml-xsl*), die frei verfügbar sind. Das Thema-Format (3.3) basiert auf Teilen der DiML-DTD, so dass sich die betreffenden XSLT-Skripte von diml-xsl nutzen lassen.

Im Gegensatz zu den ebenfalls XML-basierten Dokumentenformaten DocBook, TEI und dem Dateiformat von OpenOffice ist DiML speziell auf die Archivierung wissenschaftlicher Publikationen ausgerichtet und mit zur Zeit 113 Elementen in 13 Modulen etwas überschaubarer.

3.1.2. Datenformate für Thesauri

Für den Austausch von Thesauri existieren verschiedene, meist proprietäre Formate. Thesaurus-Editoren bieten in der Regel die Möglichkeit, einen Thesaurus in einem oder mehreren Formaten zu exportieren. Die Tatsache, dass es bislang kein allgemein etabliertes XML-Format für Thesauri gibt, mag daran liegen, dass klassische Thesauri eine relativ einfache Struktur besitzen, so dass es relativ einfach ist, ein eigenes, proprietäres Format zu entwickeln. Auch bisher mangelnde Akzeptanz des OpenContent-Gedankens trägt dazu bei, dass dem Austausch von Thesauri nicht die Bedeutung beigemessen wird – schließlich könnte jemand anderes den eigenen Thesaurus weiterverwenden. So

sind selbst kostenlos zugängliche Thesauri wie Beispielsweise Eurovoc[15] nur in Ausgabeformaten wie PDF oder RTF erhältlich. Auch liegt ein einheitliches Format nicht unbedingt im Interesse der Hersteller von Thesaurus-Software, da viele Programme neben der Verwaltung gleichzeitig Funktionen zur Ausgabe und Darstellung bieten. Dieser Schritt ließe sich mit einem einheitlichen Datenformat unabhängig von der Software implementieren.

Aus diesem Grund existieren eine Vielzahl von proprietären Formaten. Zum Austausch haben einzelne Programme spezielle Import-Filter oder es wird das einfache Textformat genutzt. Für Terminologien und linguistische Thesauri existieren eigene Formate (beispielsweise für Textverarbeitungsprogramme mit einer Thesaurusfunktion – siehe u.a. http://thesaurus.kdenews.org/).

Textformat

Der kleinste gemeinsame Nenner für den Austausch ist dabei ein einfaches Textformat. Dabei werden die einzelnen Begriffe durch Leerzeilen getrennt und Relationen zu anderen Begriffen mit den entsprechenden Kürzeln eingeleitet werden. In der Regel werden die in DIN 1462 bzw. ISO 2788 und ISO 5964 festgelegten Relationsarten und dazugehörige Kürzel verwendet (siehe 2.1.8).

Zthes

Das „verbreitetste“ XML-Format für Thesauri ist *Zthes* (http://zthes.z3950.org/). Die DTD wurde aus einem Datenmodell zur Übertragung von Thesauri abgeleitet. Sie hat jedoch mehrere Nachteile:

- Die in dem zugrunde liegende Datenmodell enthaltene Beschränkung auf die in ISO 5964 festgelegten Relationen (und einer zusätzlichen Relation) lässt sich nicht validieren.
- Alle Relationen müssen jeweils doppelt (in beiden Richtungen) angegeben werden
- Die Verlinkung der Begriffe untereinander lässt sich nicht validieren

Aus diesem Grunde können leicht fehlerhafte Thesauri in Zthes formuliert werden. Ansonsten lassen sich Thesauri im Zthes-Format relativ einfach per XSLT in andere XML-Formate transformieren.

Normdatensätze

Für den Austausch von Normdatensätzen (*authorities*) wie beispielsweise der Schlagwortnormdatei gibt es eigene Unterformate der bibliothekarischen Austauschformate MAB und MARC. Da es sich lediglich um große kontrollierte Vokabularien zur Indexierung handelt, besitzen die Normdaten nur eine sehr gering ausgeprägte Thesaurusstruktur. Für das MARC-Format existiert bereits seit einiger Zeit die XML-Version (MARCXML), für MAB und die Daten der SWD bisher noch nicht.

Thesauri in RDF

In den letzten Jahren sind bereits Methoden zur Repräsentation von Thesauri in RDF vorgeschlagen worden (z.B. [Wilson]). Dabei handelt es sich eher um *proof-of-concept*-Arbeiten; aber es gibt auch einige praktische Beispiele (u.a. unter http://www.wam.umd.edu/~katyn/SemanticWebThesauri/).

Aufgrund der Komplexität von RDF eignen sich RDF-basierte Formate nur bedingt für den Austausch von reinen Thesauri, sondern sind eher zur Einbindung von bestehenden Thesauri in die Architektur des Semantic Web gedacht. Auch lässt sich fragen, ob speziell für Thesauri nicht eher der Ansatz von Topic Maps praktikabel ist.

Einen Überblick gibt Michael Wilson in seinem Vortrag „*A Thesaurus Interchange format for*

15 Eurovoc ist ein mehrsprachiger Thesaurus, mit dessen Hilfe die Dokumente in den Dokumentationssystemen der Institutionen der Europäischen Union indexiert werden: http://europa.eu.int/celex/eurovoc/

migrating to the Semantic Web“.[16] Er erwähnt dort, dass der Thesaurus-Standard ISO 5964 zur Zeit überarbeitet wird und wahrscheinlich ein Thesaurus-Format enthalten wird.

Weitere Formate und Projekte

Weitere verwandte bereits abgeschlossene Projekte und Formate sind

- Die *TML – Thesaurus-Markup Language* [Lee]
- Das *High-Level Thesaurus Projekt* (HILT) (http://www.ukoln.ac.uk/metadata/hilt/ und http://hilt.cdlr.strath.ac.uk)
- Das *LIMBER*-Projekt (siehe http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Miller/)
- Das *Virtual HyperGlossary*-Format (VHG) (http://www.vhg.org.uk/)
- Die *Vocabulary Markup Language (VocML)* (http://nkos.slis.kent.edu/VOCML-1.DOC und http://publish.uwo.ca/~craven/thewvocm.dtd)

3.1.3. Datenformate für Terminologien

Terminologische Datenbanken werden unter anderem bei der Übersetzung und zur Erstellung von Wörterbücher eingesetzt. Für den Austausch von Terminologischen Datenbanken gibt es verschiedene Datenformate, die zum größten Teil inzwischen auf XML-basieren:

- ISO 12200: Machine-readable terminology interchange format (MARTIF)
- TermBase Exchange (TBX)
- ISO 16642: Terminological Markup Framework (TMF)
- Open Lexicon Interchange Format (OLIF2)

Eine Übersicht über terminologische Standards findet sich bei der *Localization Industry Standards Association* (http://www.lisa.org/). Der in TEI enthaltene Teil zur Auszeichnung von Terminologien ist inzwischen veraltet. Das SALT-Projekt, in dem mit XLT-Format (*XML representation of Lexicons and Terminologies*) ein Versuch zur Kombination von MARTIF und OLIF entwickelt wurde, ist inzwischen abgeschlossen und die Ergebnisse fließen in die Weiterentwicklung der oben genannten Standards ein.

Verbreitete Programme zur Verwaltung von Terminologien sind unter anderem *Déjà Vu*, *MULTILIZER*, *MultiTrans* und *TRANS Suite 2000*.[17]

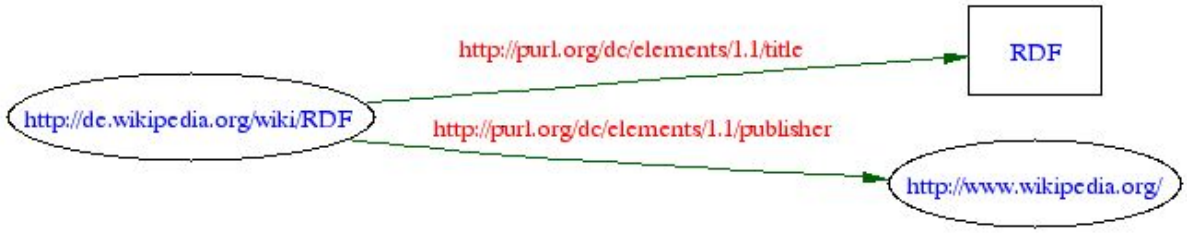
3.1.4. RDF

Das *Resource Description Framework* (RDF) ist eine Spezifikation für ein Modell zur Repräsentation von Metadaten (Informationen über Webseiten und andere Objekte, siehe 2.1.4), die erstmals 1999 vom *World Wide Web Consortium* vorgelegt wurde und üblicherweise in Form einer XML-Sprache verwendet wird. Mit der Erweiterung des Modells durch RDF Schema und Ontologien soll RDF als grundlegende Technik für den Datenaustausch im so genannten Semantic Web dienen. RDF lässt sich aber auch für weniger komplizierte Dinge einsetzen. Man muss zwischen dem *RDF-Modell*, der *RDF-Syntax* und *RDF-Schema* unterscheiden. Eine Einführung in RDF sowie unter anderem eine Übersicht von gängigen Parsern findet sich in [Powders].

16 http://www.w3c.rl.ac.uk/pasttalks/slidemaker/XML_UK_SW_Thes/Overview-3.html
17 Siehe http://www.uni-mainz.de/~wassmer/fulltext/MLCcomparison_Wassmer.pdf (2002) und http://www.uni-mainz.de/~wassmer/icon-pdf.gif (2003) von Thomas Waßmer

RDF-Modell

Informationen sind in RDF in so genannten *Statetements* abgelegt, das sind Aussagen in Form eines *N-Tripel* aus Subjekt, Prädikat und Objekt. Alle drei Bestandteile werden durch einen *Universal Resource Identifier* (URI) identifiziert und damit allgemein als *Resource* bezeichnet. Als Objekt einer Aussage sind auch Zeichenketten (*Literal*) möglich. Die Verknüpfung mehrerer Tripel lässt sich als ein gerichteter Graph mit Knoten- und Kantenbeschriftung verstehen, wobei Subjekt und Objekt eines Tripels Knoten und die Prädikate Kanten sind. Folgende Grafik wurde mit dem W3C RDF Validator[18] aus dem weiter unten angegebenen Beispiel erzeugt. Dargestellt sind zwei Tripel, mit dem selben Subjekt `http://de.wikipedia.org/wiki/RDF`.



Die Besonderheit des RDF-Modells liegt zum einen darin, dass über die als Prädikat verwendeten Resourcen (*Properties*) auch wiederum Aussagen getroffen werden können. Dadurch lassen sich Properties selbst mit RDF beschreiben und als Metadatenformat ablegen. Andere RDF-Angaben können diese Vokabulare durch Referenzierung weiterverwenden. Ein prominentes Beispiel dafür ist die Repräsentation von Dublin Core in RDF. Zum anderen bilden in RDF Statements selber Resourcen, auf die mit weiteren Statements verwiesen werden kann. Diese Technik der Aussagen über Aussagen wird als *Reification* bezeichnet.

Zusätzlich enthält RDF vordefinierte Datentypen für Listen und Mengen, um Gruppen von Resourcen zusammenzufassen. Resourcen, die keine explizite URI haben, sondern nur zur Gruppierung von anderen Objekten dienen, werden in der Regel durch so genannte „blank nodes“ modelliert. Ein Beispiel dafür ist die Zuweisung eines Namens, der aus separaten Zeichenketten für für Vor- und Nachnamen besteht.

Syntax und Speicherung

Das RDF Modell ist unabhängig von speziellen Darstellungsform. Am meisten verbreitet ist die Repräsentation in XML. Eine kürzere Syntax ist die von Tim Berners-Lee entworfene *Notation 3* (N3)[19]. Für die Speicherung von RDF in Datenbanken und Datenstrukturen gibt es verschiedene Konzepte, da eine reines Ablegen der N-Tripel in einer Tabelle nicht sehr effektiv ist. Da sich die selben RDF-Aussagen in einer Syntax mitunter auf viele verschiedene Arten ausdrücken lassen, ist es sinnvoll zur Verarbeitung von RDF-Daten einen RDF-Parser zu verwenden, der auch die Validierung gegen ein RDF Schema vornehmen kann. Vorhandene APIs können auch direkt N-Tripel zurückliefern, so dass sich um die korrekte Representation von RDF (sei es in XML oder in einer anderen Form) nicht gekümmert werden muss. Zur Suche in RDF-Daten gibt es verschiedene Anfragesprachen zum Beispiel RQL (RDF Query Language) und RDF Squish.

Beispiel

Die Aussage „`http://de.wikipedia.org/wiki/RDF` hat den Titel ‚RDF‘ und den Heraus-

18 http://www.w3.org/RDF/Validator/
19 http://www.w3.org/DesignIssues/Notation3.html

geber `http://www.wikipedia.org/`“ (wobei Titel und Herausgeber nach Dublin Core definiert sind) wird in RDF mit zwei Tripeln ausgedrückt. Die N3-Noation ist:

```
<http://de.wikipedia.org/wiki/RDF> has
   <http://purl.org/dc/elements/1.1/title> "RDF" .
<http://de.wikipedia.org/wiki/RDF> has
   <http://purl.org/dc/elements/1.1/publisher> <http://www.wikipedia.org/> .
```

In RDF/XML lässt sich die Aussage so ausdrücken:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://de.wikipedia.org/wiki/RDF">
    <dc:title>RDF</dc:title>
    <dc:publisher rdf:resource="http://www.wikipedia.org/"/>
  </rdf:Description>
</rdf:RDF>
```

RDF-Schema

Ebenso wie XML im konkreten Anwendungsfall die Definition eines speziellen Dokumenttyps benötigt (z.B. als DTD oder XML Schema), legt das RDF-Modell nur eine Syntax für den gemeinsamen Datenaustausch fest. Zur Interpretation von in RDF formulierten Aussagen bedarf es eines gemeinsamen Vokabulars wie zum Beispiel Dublin Core. Ein solches Vokabular wird auch Ontologie genannt, wenn es gleichzeitig Regeln für die richtige Verwendung der in ihm definierten Resourcen enthält. *RDF Schema* (RDFS) ist ein Vokabular zur Formulierung solcher Ontologien in RDF. RDFS liegt die Idee eines Objektorientierten Klassenmodells zugrunde. Neben RDFS existieren schon länger eine Reihe weiterer Ontologie-Beschreibungssprachen die weitgehend den unter 2.2.9 genannten Regelsprachen entsprechen.

Das RDF-Modell selbst kann in RDF-Schema formal definiert werden. Unter http://www.w3.org/TR/REC-rdf-syntax sind dazu folgende Resourcen definiert:

Klassen

- Statement (Tripel einer Aussage)
- Property (Eine Property, die als Prädikat einer Aussage benutzt werden kann)
- Bag (Ungeordnete Menge von Resourcen)
- Seq (Geordnete Menge von Resourcen)
- Alt (Menge von alternativen Resourcen)

Properties

- subject (Weist einem Statement eine Resource als Subjekt zu)
- predicate (Weist einem Statement eine Propertye als Prädikat zu)
- object (Weist einem Statement ein Objekt zu)
- type (Identifiziert die Klasse einer Resource)
- value (Einfache Zuweisung einer Eigenschaft)

3.1.5. Topic Maps

Topic Maps sind ein abstraktes Modell und ein dazu gehöriges in SGML- beziehungsweise XML-basiertes Datenformat zur Formulierung von Wissensstrukturen. Topic Maps wurden 1999 als ISO-Standard ISO/IEC 13250 normiert und später als *XML Topic Maps* (XTM) in XML formuliert.

Im Gegensatz zu RDF, das eine Computerverstehbare Formalisierung zum Ziel hat, sind Topic Maps eher zur Strukturierung von Wissen aus Sicht der Menschen konzipiert. Topic Maps sollen die bessere Navigation und Suche in Internetresourcen und anderen Dokumenten ermöglichen und dem Austausch von Metadaten dienen. Sie haben ihre Wurzeln in Glossaren, Klassifikationen und Thesauri, gehen aber in ihrer Ausdrucksstärke über diese Hinaus. So lassen sich mit Topic Maps Ontologien formulieren, die unter anderem auch auf RDF abgebildet werden können.[20]

In der Praxis werden mit Topic Maps jedoch oft lediglich einfache (facettierte) Klassifikationen modelliert, so dass auch die vereinfachte Untermenge *XFML* (siehe unten) ausreicht. Topic Maps werden aufgrund ihres Praxisbezugs und der verfügbaren Software bereits mehr als RDF auch außerhalb von Forschungsprojekten eingesetzt (Beispiele siehe unter http://easytopicmaps.com/).

Von ihrer Struktur her sind die dem unter 3.2 vorgestellten Thema-Datenmodell sicherlich am nächsten. Näheres zu Topic Maps siehe unter http://www.topicmaps.org/ sowie in [Widhalm] und [Park].

Bestandteile

Topic Maps bestehen aus so genannten *Topics* (abstrakte Begriffen), *Associations* (Relationen zwischen Topics) und *Occurences* (Eine Topic zugeordnete Dokumente oder andere Topics). Alle drei können Instanz eines übergeordneten Topics sein (Instanzrelation). Einem konkreten Topic können

- Bezeichnungen in Form von Zeichenketten oder anderen Ressourcen und
- Occurences als Vorkommen des Topics

zugeordnet werden. Dabei kann jeweils bei Bedarf ein *Scope* angegeben werden, der bestimmt, in welchem Kontext die Bezeichnung oder die Occurence gültig ist. Ein Scope ist wieder ein Topic oder eine andere Ressource. Innerhalb eines Scopes müssen die Bezeichnungen eindeutig sein.

Die Relationen (*Associations*) in Topic Maps sind n-Tupel von Topics oder anderen Ressourcen, denen als *Members* so genannte *Roles* zugewiesen werden können – vergleichbar mit Einträgen in einer Datenbank wobei die Roles die einzelnen Datenbankfelder kennzeichnen.

XFML (eXchangeable Faceted Metadata Language)

Das Format XFML ist ein Austauschformat für facettierte Klassifikationen. Eine facettierte Klassifikation ist ein hierarchisches Begriffssystem, in dem mehrere (mono)hierachische Unterteilungen nach verschiedenen Aspekten (Facetten) existieren (siehe 2.1.5 und 2.2.8). XFML wurde 2002 von Peter Van Dijck zunächst für den Einsatz in Weblogs entwickelt. Die Version 1.0 (siehe http://xfml.org/) zielt allgemeiner auf den Austausch von Verschlagwortungen zwischen Webseiten. XFML ist ein eigenes Format, das aus einer Untermenge von Topic Maps besteht. Eine Besonderheit von XFML besteht darin, dass bei der Indexierung (Zuweisung von Topics/Begriffen zu Webseiten) eine so genannte *Occurrencestrength* als Maß der Sicherheit einer Zuweisung angegeben werden. Begriffe aus verschiedenen XFML-Klassifikationen können miteinander verknüpft werden, so dass fremde Indexierungen mit anderen Klassifikationen übernommen werden können.

20 Zur Beziehung zwischen RDF und Topic Maps siehe [Garshol] und http://www.ontopia.net/topicmaps/materials/rdf.html (Ten Theses on Topic Maps and RDF)

3.2. Das Thema-Datenmodell

Das folgende Datenmodell ist aus einer Untersuchung der in Kapitel 2 behandelten Begriffssysteme und ihrer Bestandteile entstanden.Die Bezeichnung „Thema“ kann zum einen als Abkürzung für „Thesaurus Markup“ verstanden werden und deutet zum anderen darauf hin, dass dieses Daten/modell dazu dient, Themengebiete in Form der in ihnen vorkommenden Begriffe und Beziehungen zu modellieren. Das Thema-Datenmodell eignet sich unter anderem für Thesauri, Klassifikationen, Schlagwortlisten, Glossare, Wörterbücher, Begriffsnetze und Ontologien. Es beinhaltet im Kern Begriffe, Bezeichnungen und Relationen. Darüber hinaus werden häufig vorkommende Phänomene und Arten von Relationen und Bezeichnungen abgebildet. Das Konzept der Vererbung und formal logische Regeln sind nicht Bestandteil des Modells. Sie lassen sich jedoch mit einer beliebigen Regelsprache über den Objekten und Relationen hinzufügen.

Im folgenden soll die graphische Übersicht des Modells in UML-Notation (siehe folgende Seite) anhand der einzelnen Bestandteile erläutert werden (3.2.1 bis 3.2.3). Die dabei angegebenen Beispiele sind in der *Thema-DTD*, einer XML-Repräsentation des Modells formuliert. Die Thema-DTD wird in Kapitel 3.3 anhand eines vollständigen Beispiels eingeführt.

Basisklassen

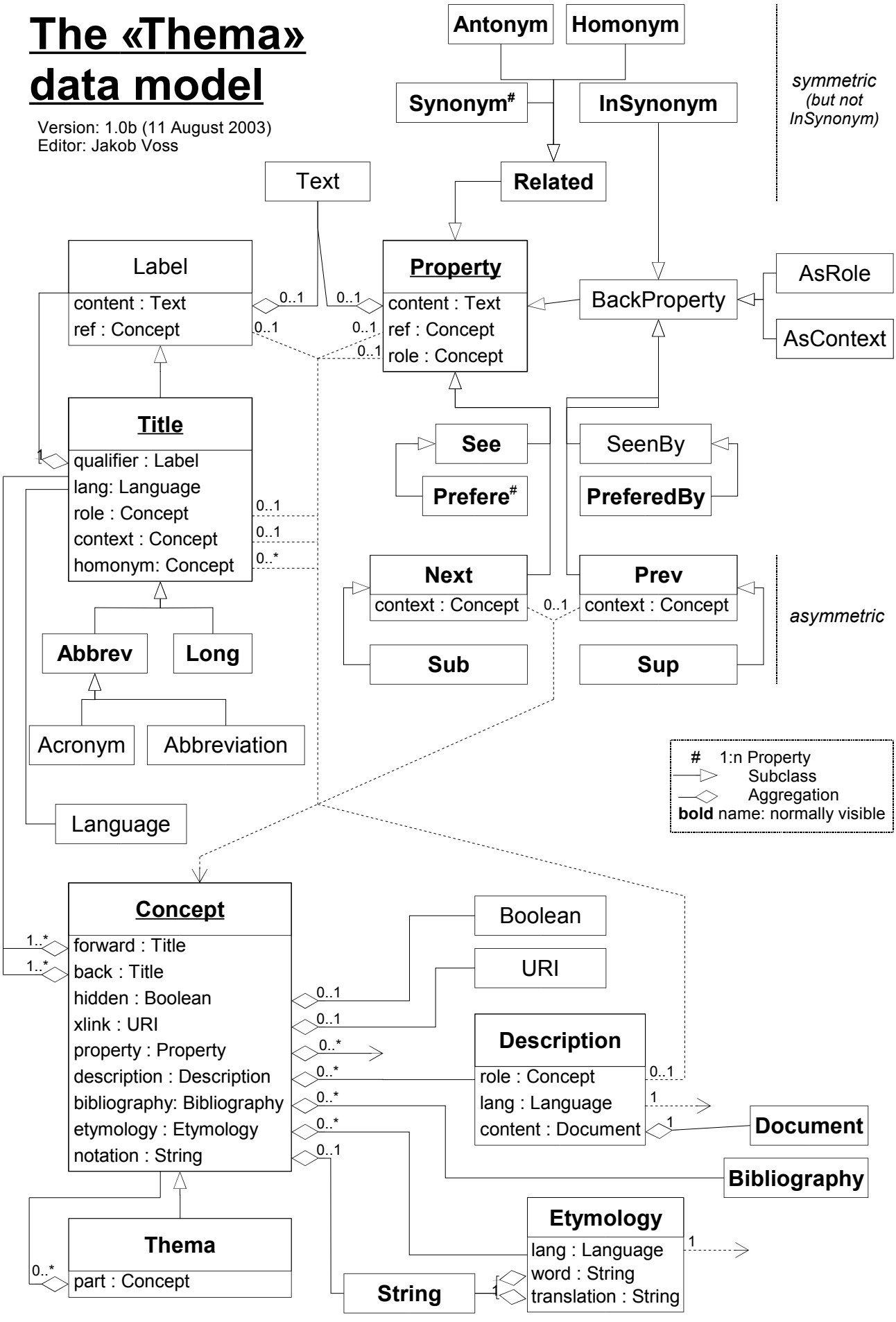
Die Grundlegenden Basisklassen stellen Datentypen für Begriffe (**Concept**, 3.2.1), Bezeichnungen (**Title**, 3.2.2) und Relationen (**Property**, 3.2.3) dar. Eine Reihe von häufigen Spezialfällen sind von diesen Klassen abgeleitet. Daneben gibt es eine Reihe von einfachen Datentypen, die von anderen Klassen benutzt werden. Dies sind:

- **Boolean**
  Eine logischer Binärwert (ja/nein)
- **Language**
  Eine Sprache einer Bezeichnungen oder eines Textes. Einzelne Sprachen werden durch ihren zwei bzw. drei Zeichen langen ISO-Code identifiziert (ISO 639-1 bzw. ISO 639-2).
- **String**
  Eine einfache Unicode-Zeichenkette.
- **Text**
  Eine Unicode-Zeichenkette, die mit einer Auswahl von XML-Elementen ausgezeichnet werden kann. Die erlaubten Elemente entstammen der *inline*-Gruppe der DiML-DTD mit den Elementen der Module *common*, *citation*, *media* und *mathematics*.[21] Entspricht im vereinfachten Thema-Format einem *String*.
- **URI**
  Ein *Uniform Resource Identifier* (RFC 2396) zur systemübergreifenden Identifikation. In der Regel handelt es sich um eine URL.
- **Document**
  Strukturierter Text, der aus einzelnen Absätzen besteht, die wie Text weitere Auszeichnungen sowie Tabellen, Listen, Grafiken etc. enthalten können.
- **Bibliography**
  Strukturiertes Verzeichnis von Verweisen auf externe Dokumente

21 Siehe http://edoc.hu-berlin.de/diml/group/inline.php

The «Thema» data model

Version: 1.0b (11 August 2003)
Editor: Jakob Voss



3.2.1. Concept

Die grundlegende Basisklasse für die einzelnen Begriffe eines Begriffssystems ist *Concept*. Über verschiedene Attribute können einem *Concept* Eigenschaften zugeordnet werden, von denen bis auf die Angabe mindestens einer Bezeichnung (*forward/back* vom Datentyp *Title*) alle optional sind. Eine spezielle Art eines *Concept* ist ein durch Gruppierung gebildetes *Thema*-Objekt.

Attribut	*Datentyp*	*Beschreibung*
forward	Title	Standardbezeichnung des Begriffes bzw. Bezeichnung in Hinrichtung
back	Title	Bezeichnung des Begriffes in Rückrichtung (entspricht meist *forward*)
xlink	URI	Angabe einer URI zur eindeutigen Identifikation des Begriffes
notation	String	Angabe einer Notation für den Begriff
property	Property	Eigenschaften und Relationen mit Werten oder anderen Begriffen
description	Description	Beschreibungen, Erklärungen, Beispiele, Definitionen etc. als Text
bibliography	Bibliography	Bibliographische Angaben zum Begriff (eigener Datentyp)
etymology	Etymology	Etymologische Angaben zum Begriff (eigener Datentyp)
hidden	Boolean	Gibt an, ob es sich um einen versteckten Begriff handelt. Versteckte Begriffe sind nicht eigentliche Bestandteile eines Begriffssystem, sondern dienen beispielsweise als Relationsarten der internen Verknüpfung

In einer konkreten Repräsentation des Thema-Datenmodelles müssen die einzelnen Begriffe zusätzlich mit einem eindeutigen Identifikator versehen sein. Dieser kann einer Notation oder einer eindeutigen Benennung entsprechen oder auch aus einer rein internen Datensatznummer bestehen.

In der Thema-DTD geschieht die Verknüpfung von Begriffen mittels XML-IDs. Jeder Begriff muss mit dem Attribut *id* eine ID besitzen, die innerhalb eines Dokumentes eindeutig ist. Verweise auf andere Begriffe werden je nach Verwendung mit dem Attribut *ref* oder einem anderem Attribut gesetzt, dessen Wert einer ID innerhalb des selben Dokumenten entsprechen muss. Bei Begriffskombinationen (möglich mit *Synonym* und *Prefere*) wird auf mehrere, durch Leerzeichen getrennte IDs verwiesen. Um auf Begriffe zu verweisen, die sich in einem anderen Dokument befinden, muss ein *xlink*-Attribut benutzt werden.

Description und Document

Für Erläuterungen, Definitionen, Beschreibungen, Beispiele etc. zu einzelnen Begriffen existiert der Datentyp *Description*. Der Inhalt einer *Description* besteht aus einem *Document* Der darin enthaltene Text kann mit einer Auswahl von XML-Auszeichnungen des DiML-Formates formatiert werden. Die Auszeichnungsmöglichkeiten beinhalten Absätze, Gliederungen, Zeichenformatierungen, Hyperlinks, Listen, Tabellen und Grafiken. Zu jeder *Description* kann eine Sprache angegeben werden. Falls dies nicht geschieht, wird die Sprache des übergeordneten *Thema* bzw. des gesamten Begriffssystems angenommen. Mit *role* kann die Art der Beschreibung genauer spezifiziert werden. Das Thema-Datenmodell beinhaltet keine vordefinierten Beschreibungsarten; es ist jedoch ratsam, zu diesem Zwekc eine begrenzte Menge von Konzepten (z. B. Definition, Beispiel, Erklärung, Anmerkung etc.) zu definieren.

Attribut	*Datentyp*	*Beschreibung*
role	Concept	Ein Konzept, dass die Art der Beschreibung angibt
lang	Language	Sprache der Beschreibung
content	Document	Beschreibung in Form eines (Teil-)Dokumentes (strukturierter Text)

Beispiel

```
1 <concept id="Bezeichnung">
2   <description role="Definition">
3     <p>Repräsentation eines <term ref="Begriff">Begriffs</term>
4        mit sprachlichen oder anderen Mitteln.</p>
5   </description>
6   <description role="Anmerkung">
7     <p>Die Trennung zwischen Begriff und Benennung ist auch
8        als <term ref="TerminologischeKontrolle"/> bekannt.</p>
9   </description>
10</concept>
```

Etymology

Die Untersuchung der Herkunft und Geschichte von Ausdrücken (Etymologie) kann oft Hinweise auf die Bedeutung der sie bezeichnenden Begriffe geben. Bei den mit *Etymology* angegebenen Worten kann es sich auch um einfache Ausdrücke handeln, die mit dem Begriff in Verbindung stehen. Zu jedem Ausdruck (*word*) wird seine Sprache (*lang*) und eine Übersetzung, Herleitung oder Erläuterung (*translation*) angegeben.

Attribut	*Datentyp*	*Beschreibung*
lang	Language	Sprache eines Ausdrucks, aus dem sich der Begriff ableitet
word	String	Ein Ausdruck, aus dem sich der Begriff ableitet
translation	String	Übersetzung und/oder Beschreibung des Ausdrucks

Beispiel

```
1 <concept id="glossar">
2   <title>Glossar</title>
3   <etymology lang="la" word="glossarium">Spruchsammlung</etymology>
4   <etymology lang="el" word="glossarion">Wörtchen</etymology>
5 </concept>
```

Bibliography

Quellenangaben und Literaturhinweise zu einzelnen Begriffen (*Bibliography*) sollten soweit wie möglich genauer strukturiert sein, um die Weiterverarbeitung und einheitliche Gestaltung zu ermöglichen. Denkbar aber nicht immer praktikabel ist die direkte Verwendung von Metadatenformaten wie Dublin Core, BibTeX, MARC, MAB, DLMeta, METS, ONIX etc. Die Thema-DTD übernimmt für bibliographische Angaben das Element *citation* der DiML-DTD.[22]

Es sei erwähnt, dass Literaturangaben auch innerhalb eines Beschreibungstextes stehen können und sich daraus eine Bibliographie teilweise automatisch erstellen lässt.

Thema

Die Klasse *Thema* repräsentiert eine Aggregation von mehreren Begriffen (*Concept*). Ein gesamtes Begriffssystem wird dementsprechend als ein *Thema* modelliert. Ein *Thema* ist gleichzeitig auch ein eigenes *Concept*. Damit können verschiedene Begriffssystem als unterschiedliche Konzepte miteinander in Beziehung gesetzt werden. Indem die Begriffe eines *Thema* selbst *Thema*-Objekte sein können, lassen sich Begriffsysteme wie unter (2.2.8/S. 32) beschrieben gruppieren und verschachteln.

Attribut	*Datentyp*	*Beschreibung*
part	Concept	Begriff, der Bestandteil des Begriffssystems ist

22 Siehe http://edoc.hu-berlin.de/diml/element/citation.php

Beispiel

```
1 <thema id="gesamtThema">
2   <thema id="relationen" hidden="yes">
3     <concept
4     ...
5   </thema>
6   <thema id="begriffe">
7     <concept
8     ...
9   </thema>
10</thema>
```

3.2.2. Title

Zur Darstellung sollte jedem Begriff mindestens eine Bezeichnung zugeordnet sein. Die einfachste Form einer Bezeichnung ist ein (ggf. formatierter) String vom Datentyp *Text*. Im Thema-Datenmodell basieren Bezeichnungen (*Title*) auf der Klasse *Label*, die aus einem Text (*content*) und/oder einem Verweis auf einen definierten Begriff (*ref*) besteht. Letzteres ermöglicht es, Bezeichnungen als eigene Begriffe zu verwalten, auf die jeweils verwiesen wird. Diese Praxis ist jedoch nur in Ausnahmefällen empfohlen, um die Unterscheidung zwischen Begriffen und Benennungen eindeutig zu halten. Im folgenden Beispiel ist die Bezeichnung „Birne“ für Zwei Begriffe über einen Verweis angegeben, während die Bezeichnung „Apfel“ direkt bei dem so benannten Begriff steht.

Label

Attribut	*Datentyp*	*Beschreibung*
content	Text	Inhalt in Form eines einfachen formatierten Strings (Text)
ref	Concept	Inhalt in Form eines Verweises auf einen anderen Begriff

Beispiel

```
1 <concept id="BirneString" hidden="yes">
2   <title>Birne</title>
3 </concept>
4 <concept id="Birnenfrucht">
5   <title ref="BirneString"/>
6 </concept>
7 <concept id="Gluehbirne">
8   <title ref="BirneString"/>
9 </concept>
10<concept id="Apfel">
11  <title>Apfel</title>
12</concept>
```

Title

Zusätzlich zu *content* und *ref* können Bezeichnungen (*Title*) mit weiteren Attribute versehen sein.

Attribut	*Datentyp*	*Beschreibung*
qualifier	Label	Ein Qualifier (siehe 2.2.3) zur Unterscheidung gleicher Bezeichnungen
lang	Language	Sprache der Bezeichnung
role	Concept	Ein Konzeptes, dass die Art der Bezeichnung angibt
context	Concept	Ein Kontextes, in dem diese Bezeichnung sinnvoll ist
homonym	Concept	Verweis auf einen Begriff, der eine homonyme Bezeichnung besitzt

Beispiel

```
1 <concept id="Bibel">
2   <title lang="de">Bibel</title>
3   <title lang="en">bible</title>
4   <title context="Kirche" lang="de">Heilige Schrift</title>
5   <abbrev role="Zitierungkuerzel">BIBL</abbrev>
6 </concept>
```

Im Thema-Datenmodell werden jedem Begriff mindestens eine Bezeichnung je Richtung (*forward/back*) zugewiesen. In der Praxis besteht diese oft aus einer einzigen Bezeichnung, die für beide Richtungen benutzt wird, sofern es nicht anders angegeben ist. Wenn – beispielsweise für die Definition einer Relation – verschiedene Bezeichnungen für die einzelnen Richtungen benutzt werden sollen, muss im Thema-Format das XML-Element *role* statt *title* benutzt werden.

Beispiel

```
1 <concept id="Kausalitaet">
2   <title>Kausalität</title>  <!-- Vorzugsbenennung falls nicht Relation -->
3   <role direction="forward">Ursache</role>
4   <role direction="back">Wirkung</role>
5 </concept>
```

Abbrev, Long, Acronym und Abbreviation

Abbrev und *Long* stehen für Bezeichnungen, die jeweils eine Kurzform oder Langform sind; Ansonsten unterscheiden sie sich nicht von anderen Bezeichnungen. Bei den Kurzformen lässt sich noch einmal zwischen Abkürzungen und Akronymen unterscheiden (siehe Beispiel).

Beispiel

```
1 <concept id="Vorlesung">
2   <title qualifier="Lehrveranstaltung">Vorlesung</title>
3   <abbrev type="abbreviation">Vorl.</abbrev>
4   <abbrev type="acronym">VL</abbrev>
5   <long>Universitätsvorlesung</long>
6 </concept>
```

3.2.3. Property

Eigenschaften von und Verknüpfungen zwischen Begriffen werden mit einer *Property* bei einem Begriff angegeben. Der Inhalt einer *Property* kann aus einem *Text*-Objekt (*content*) und/oder einem Verweis auf einen anderen Begriff (*ref*) bestehen – bei einem Verweis wird so eine Verknüpfung zwischen zwei Begriffen definiert. Die Art der Verknüpfung oder Eigenschaft kann durch einen eigenen Begriff, auf den mit *role* verwiesen wird, genauer spezifiziert werden. Einigen häufig vorkommenden Arten sind im Thema-Datenmodell als eigene Klassen definiert.

Attribut	*Datentyp*	*Beschreibung*
content	Text	Inhalt in Form eines einfachen formatierten Strings (Text)
ref	Concept	Inhalt in Form eines Verweises auf einen anderen Begriff
role	Concept	Eine Konzeptes, dass die Art der Property angibt

Beispiel

```
1 <concept id="EineBestimmtePerson">
2   <property role="Alter">88</property>
3   <property role="Nationalitat" ref="Deutsch"/>
4 </concept>
```

BackProperty

Relationen verknüpfen je zwei Begriffe und lassen sich sich somit von zwei Seiten betrachten. Die Klasse *BackProperty* ist eine Eigenschaft, die die Rückrichtung einer *Property* angibt (gleichzeitig ist sie ein Spezialfall einer *Property*). Eine *Property* des Begriffes A mit dem Wert B impliziert, dass es eine *BackProperty* des Begriffes B mit dem Wert A gibt. Die Rückrichtung ist nur bei einigen Relationen von Interesse. Bei Symmetrischen Relationen entsprechen sich Hin- und Rückrichtung. Ebenso wie bei *Property* lässt sich mit *role* bei Bedarf die genauere Art der Relation angeben.

Zwei spezielle Arten von *BackProperty* sind *AsRole* und *AsContext*. Sie verweisen auf Begriffe, die einen anderen Begriff als *role* einer *Property* oder als *context* eines *Title*-benutzen. Die in Zeile 5 und 8 des folgenden Beispiels angegebenen Eigenschaften lassen sich automatisch aus Zeile 2 ermitteln.

Beispiele

```
1 <concept id="EineBestimmtePerson">
2   <property role="Nationalitaet" ref="Deutsch"/>
3 </concept>
4 <concept id="Deutsch">
5   <backproperty role="Nationalitaet" ref="EineBestimmtePerson"/>
6 </concept>
7 <concept id="Nationalitaet">
8   <asRole ref="EineBestimmtePerson"/>
9 </concept>
```

Die grundlegenden Relationsarten lassen sich in symmetrische und asymmetrische Relationen unterteilen. Die prototypische symmetrische Relation ist *Related* und die asymmetrische Relation *See*.

Related

Related drückt eine einfache Beziehung zwischen zwei Begriffen aus. Die Beziehung gilt immer in beide Richtungen. Im Datenformat muss nur eine Richtung angegeben werden, da die andere Richtung automatisch hinzugefügt wird. Die genauere Art der Relation kann mit *role* angegeben werden.

Beispiele

```
<concept id="Kuh">
  <related ref="Milch"/>
</concept>
```

Antonym und Honomyn

Zwei spezielle Arten der *Related*-Relation sind die Antonymie (*Antonym*) und Homonymie (*Homonym*). Beide sind zwar zwischen je zwei Begriffen definiert, können aber auch indirekt an anderen Stellen angegeben werden (siehe Beispiel). Antonyme Begriffe sind häufig Unterbegriffe eines gemeinsamen Oberbegriffes. Homonymbeziehungen zwischen zwei Begriffen existieren in der Regeln aufgrund einer gemeinsamen Bezeichnungen und können somit automatisch erkannt werden.

Beispiele

```
<concept id="Schwarz">
  <antonym ref="Weiss"/>     <!-- explizit. Rückrichtung wird hinzugefügt -->
</concept>
<concept id="Grundfarbe">
  <subcons antonymity="yes"> <!-- jeder Unterbegriff ist antonym zu jedem -->
    <term ref="Rot"/>
    <term ref="Gruen"/>
    <term ref="Blau"/>
  </subcons>
</concept>
<concept id="Sitzbank">
  <title>Bank</title>                          <!-- implizite Homonymie -->
  <homonym ref="Bankinstitut"/>                <!-- explizite Homonymie -->
</concept>
<concept id="Bankinstitut">
  <title>Bank</title>                           <!-- implizit Homonymie -->
  <homonym ref="Sitzbank"/>                    <!-- explizite Homonymie -->
</concept>
<concept id="riverbank" lang="en">
  <title homonym="Sitzbank Bankinstitut">bank</title>       <!-- explizit -->
</concept>
```

Synonym und InSynonym

Oft es es ratsam, synonyme Begriffe zu einem Begriff zusammenzufassen, anstatt eine Synonymiebeziehung anzugeben. Wo dies doch erwünscht ist oder Kombinationen von mehreren Begriffen mit einem anderen Begriff in Beziehung gesetzt werden sollen (siehe Seite 24), kann die Property *Synonym* benutzt werden. Bei 1-zu-1-Beziehungen ist *Synonym* seine eigene Rückrichtung. Im Falle einer Begriffskombination (1-zu-n-Beziehung) gibt *InSynonym* an, dass ein Begriff in Kombination mit anderen synonym zu einem dritten ist (siehe folgendes Beispiel).

Beispiel

```
<concept id="Freiluft"/>
  <synonym ref="OpenAir"/>             <!-- Freiluft=OpenAir -->
</concept>
<concept id="Freibad">
  <synonym ref="Schwimmbad OpenAir"/>  <!-- Freibad= Schwimmbad & OpenAir -->
</concept>
<concept id="OpenAir">
  <synonym ref="Freiluft"/>            <!-- Rückrichtung -->
  <inSynonym ref="Freibad"/>           <!-- Rückrichtung -->
</concept>
```

See und SeenBy

Die Property *See* ist der Prototyp einer gerichteten (asymmetrischen) Relation. Die Rückrichtung *SeenBy* muss nicht explizit angegeben werden. Bei gerichteten Relationen ist es sinnvoll, für Hin- und Rückrichtung verschiedene Bezeichnungen anzugeben.

Beispiele

```
<concept id="Kuh">
  <see ref="Milch" role="ErzeugungsVerhaeltnis"/>
</concept>
<concept id="Milch"> <!-- Die Rückrichtung wird automatisch erzeugt: -->
  <seeenBy ref="Kuh" role="ErzeugungsVerhaeltnis"/>
</concept>
<concept id="ErzeugungsVerhaeltnis">
  <role direction="forward">liefert</role>
  <role direction="back">wird geliefert von</role>
</concept>
```

Prefere und Prefered

Die Property *Prefere* verweist von einem Begriff auf einen anderen, der in einem gewissen Kontext vorzuziehen ist. In der Praxis ist solch ein Verweis zum Beispiel zwischen Nicht-Vorzugsbenennung und Deskriptor in einem Thesaurus üblich. Im Thema-Datenmodell ist die einfachere Möglichkeit, einem Begriff mehrere Bezeichnungen zuzuweisen, wobei die jeweils erste die Vorzugsbenennung ist. Ebenso wie *Synonym* kann *Prefere* auch von einem Begriff auf eine Kombination mehrerer Begriffe verweisen. Die Rückrichtung von *Prefere* ist *Prefered*.

Beispiel

```
<concept id="Rechenanlage">
  <prefere ref="Computer"/>
</concept>
<concept id="Orangensaft">
  <prefere ref="Orange Fruchtsaft"/>
</concept>
```

Next und Prev

Ebenso wie *See* steht die Klasse *Next* für eine gerichtete Relation. Das hinter *Next* stehende Konzept ist allerdings das einer Ordnungsrelation, die sich in beide Richtungen fortsetzen lässt. Die Rückrichtung zu *Next* ist *Prev*. Auch hier kann zusätzlich ein weiterer Begriff als Kontext angegeben werden (*context*), innerhalb dessen die Relation definiert ist bzw. der für die Gültigkeit oder Interpretation der Relation von Bedeutung ist.

Attribut	*Datentyp*	*Beschreibung*
context	Concept	Kontext, innerhalb dessen die Relation zu definiert ist

Beispiel

```
<concept id="StarWarsEpisode1">
  <next ref="StarWarsEpisode2"/>
</concept>
<concept id="Explosion">
  <next role="Verursachung" ref="Zerstoerung"/>
</concept>
```

Sub und Sup

Eine der häufigsten Relationen, die Hierarchische Relation, ist im Thema-Datenmodell ein Spezialfall der Ordnungsrelation (*Next*). Die Hinrichtung (*Sub*) verweist auf einen untergeordneten und die Rückrichtung (*Sup*) auf übergeordneten Begriff. Eine Besonderheit der hierarchischen Relation ist, dass Begriffe, die durch sie geordnet werden, gleichzeitig häufig in Nebenordnung zueinander stehen. Diese Beziehung kann beispielsweise in einer Antonymie oder Reihenfolge bestehen.

Beispiele

```
1 <concept id="Adam">
2   <subcon ref="Kain" role="Nachkomme"/>
3   <subcon ref="Abel" role="Nachkomme"/>
4 </concept>
5 <concept id="Tag">
6   <subcons sorted="circle" classification="yes">
7     <term ref="Montag"/>
8     <term ref="Dienstag"/>
9     <term ref="Mittwoch"/>
10    <term ref="Donnerstag"/>
11    <term ref="Freitag"/>
12    <term ref="Samstag"/>
13    <term ref="Sonntag"/>
14  </subcons>
15  <subcons aspect="Freizeit">
16    <term ref="Wochentag"/>
17    <term ref="Feiertag"/>
18  </subcons>
19</concept>
20<concept id="Montag">
21  <next ref="Dienstag"/>    <!-- wird automatisch erzeugt -->
22</concept>
23...
24<concept id="Sonntag">
25  <next ref="Montag"/>      <!-- wird automatisch erzeugt (sorted=circle) -->
26</concept>
```

In in der XML-Repräsentation gibt es zwei Möglichkeiten, Hierarchische Beziehungen anzugeben. Zum einen als einzelne Relation mit dem Element *subcon* (Zeile 2/3) und zum anderen als Gruppe mit dem Element *subcons* (Zeile 6-14 und 16-19). Mit *subcons* können der Gruppe und der sich in ihr befindlichen Elemente auch weitere Eigenschaften zugewiesen werden, z. B. eine wechselseitige Antynomiebeziehung oder eine Ordnung zwischen den einzelnen Begriffen (siehe Beispiel).

Mit dem Attribut classification (Zeile 5) kann zusätzlich angegeben werden, ob die Einteilung in Unterbegriffe vollständig ist – dies ist jedoch bisher nicht Bestandteil des Thema-Datenmodells.

3.3. Das Thema-Dateiformat in XML

Das Thema-Dateiformat ist eine Implementation des Thema-Datenmodells in Form einer DTD bzw. eines XML Schemas. Die DTD soll im folgenden anhand einiges vollständigen Beispiels kurz vorgestellt werden (siehe auch die vorangegangenen Quelltext-Beispiele). Die DTD selbst findet sich in einer leicht gekürzten Version im Anhang (ab Seite 61).

Einführungsbeispiel

Die Thema-DTD ist für verschiedene Arten von Begriffsystemen von Lexika über Thesauri bis zu Ontologien konzipiert. Zur Einführung dient als Beispiel ein kleines Glossar mit zwei Begriffen:[23]

Ein Kleines Glossar

URI A Uniform Resource Identifier (URI) is a compact string of characters for identifying an abstract or physical resource.

URL A Uniform Resource Locator (URL) is a compact string representation for a resource available via the Internet.

In Thema-Syntax ließe sich das ganze folgendermaßen ausdrücken:

Beispiel 1 (Begriffe)

```
1 <?xml version="1.0" encoding="ISO-8859-1"?>
2 <thema id="glossar">
3   <title>Ein kleines Glossar</title>
4   <concept id="URI">
5     <description>
6       <p>
7         A Uniform Resource Identifier (URI) is a compact string of
8         characters for identifying an abstract or physical resource.
9       </p>
10    </description>
11  </concept>
12  <concept id="URL">
13    <description>
14      <p>
15         A Uniform Resource Locator (URL) is a compact string
16         representation for a resource available via the Internet.
17      </p>
18    </description>
19  </concept>
20</thema>
```

Nach dem obligatorischen XML-Header in der ersten Zeile folgt das Wurzelelement *thema*, das mit einer eindeutigen ID versehen sein muss (hier die Zeichenkette „glossar“). Innerhalb von *thema* ist mit dem Element *title* der Titel des Glossares angegeben (Zeile 3). Es folgen mit dem Element *concept* jeweils zwei Begriffe (Zeile 4-11 bzw. 12-19), die ebenso wie das Wurzelelement *thema* eine ID haben müssen. Innerhalb eines Begriffes wird dieser mit *description* beschrieben. In diesem Fall besteht jede Beschreibung aus einem Absatz (*p*).

Die in diesem Glossar enthaltenen Angaben lassen sich Schritt für Schritt weiter formalisieren. Die wesentlichen Änderungen gegenüber dem vorhergehenden Beispiel sind in den folgenden Beispielen fett hervorgehoben und werden im Text erklärt.

23 Die Definitionen sind RFC2396 (URI) und RFC1738 (URL) entnommen.

Beispiel 2 (Bezeichnungen)

```
1 <thema id="glossar">
2   <title>Ein kleines Glossar</title>
3   <concept id="URI">
4     <abbrev>URI</abbrev>
5     <long>Uniform Resource Identifier</long>
6     <description label="Definition">
7       <p>
8         A compact string of characters for identifying an abstract
9         or physical resource.
10      </p>
11    </description>
12  </concept>
13  <concept id="URL">
14    <abbrev>URL</abbrev>
15    <long>Uniform Resource Locator</long>
16    <description label="Definition">
17      <p>
18         A compact string representation for a resource available
19         via the Internet.
20      </p>
21    </description>
22  </concept>
23</thema>
```

Die bei jedem *concept* angebene ID dient der eindeutigen Identifikation eines Begriffes und muss nicht eine sinnvolle Bezeichnung enthaltenen. Falls aber keine andere Bezeichnung für einen Begriff angegeben ist, werden die Zeichen der ID verwendet. Ein Begriff kann auch mehrere Bezeichnungen enthalten. In Beispiel 2 besitzt jeder Begriff eine Abkürzung (*abbrev*) und eine Langform (*long*). Das jeweils als erstes angegebene Element zur Bezeichnung wird als die Vorzugsbenennung eines Begriffes verwendet.

Da es sich bei den Beschreibungen der Begriffe um Definitionen handelt und nicht zum Beispiel um Beispiele oder Anmerkungen, ist es sinnvoll, dies auch anzugeben. Dies kann mit dem Attribut *label* am Element *description* geschehen. Allerdings handelt es sich bei der mit *label* angegebenen Bezeichnung um eine reine Zeichenkette, der außer der menschlichen Lesart keine Bedeutung beigelegt ist. Deshalb ist es ratsam stattdessen mit *role* auf einen bestimmten Begriff zu verweisen.

Beispiel 3 (Einfache Relationen)

```
1 <thema id="glossar">
2   <title>Ein kleines Glossar</title>
3   <concept id="Definition" hidden="yes">
4     <abbrev>DEF</abbrev>
5   </concept>
6   <concept id="resource"/>
7   <concept id="URI">
8     <abbrev>URI</abbrev>
9     <long>Uniform Resource Identifier</long>
10    <description role="Definition">
11      <p>
12        A compact string of characters for identifying an abstract
13        or physical <term ref="resource"/>.
14      </p>
15    </description>
16    <related ref="URL"/>
17  </concept>
18  <concept id="URL">
19    <abbrev>URL</abbrev>
```

```
20    <long>Uniform Resource Locator</long>
21    <description role="Definition">
22      <p>
23         A compact string representation for a <term ref="resource"/>
24         available via the Internet.
25      </p>
26    </description>
27  </concept>
28</thema>
```

In Beispiel 3 ist „Definition“ ein eigener Begriff (Zeile 3). Er wird benutzt, indem das Attribut *role* des Elementes *description* auf den Begriff mit der ID „Definition“ verwiesen wird (Zeile 10 und 21). Die Vorzugsbenennung, die dabei ausgegeben wird, besteht aus der Abkürzung „DEF“. Da der Begriff einer Definition nicht selbst im Glossar auftauchen soll, ist er mit dem Attribut *hidden* als versteckter Begriff gekennzeichnet.

Auch an anderen Stellen können zwischen Begriffen durch Verweise Relationen angegeben werden. In Zeile 13 und 23 enthalten die Definitionstexte je einen Hyperlink auf den hier Begriff mit der ID „resource“, der hier nicht weiter erläutert wird. Dies geschieht mit dem Element *term* und dessen Attribut *ref*. Für semantisch konkretere Relationen gibt es eine Reihe von speziellen Elementen. Dazu gehört beispielsweise das Element *related* für eine einfache Assoziationsbeziehung (Zeile 16).

Da die Assoziationsbeziehung mit *related* symmetrisch ist, wird die Rückrichtung automatisch ergänzt (`<related ref="URI"/>` innerhalb des Begriffes „URL“).

Neben der symmetrischen Assoziationsbeziehung mit related gibt es eine Reihe weiterer Relationsarten, die jeweils mit einem eigenen Element angegeben werden. Dies sind *synonym* (Synonomie), *antonym* (Antinomie), *hononym* (Homonymie), sowie *see* und *prefere* (gerichtete Assoziationen), *next* und *prev* (Ordnungsrelation), *subcon* und *supcon* (hierarchische Relationen) sowie *property* (Eigenschaftsrelation).

Implementation

Das beschriebene Datenmodell wurde in Form der hier beschriebenen DTD implementiert. Zusätzlich wurden eine Reihe von XSLT-Skripten zur Darstellung von Begriffssystemen im Thema-Format in HTML sowie zum Import und Export von und nach anderen Formaten entwickelt. Da es sich lediglich um Prototypen handelt und es den Rahmen dieser Arbeit sprengen würde, werden diese nicht weiter vorgestellt. Beispiele für die Anwendung des Systems sind der *Semiotische Thesaurus* unter http://www.ib.hu-berlin.de/~wumsta/infopub/semiothes/lexicon/default/index.html und die Homepage des Thema-Projektes unter http://thema.sourceforge.net/. Für den weiteren Einsatz ist jedoch – wie für eine breitere Anwendung des Semantik Web allgemein – die Entwicklung eines einfachen Editors für Begriffsnetze notwendig, da die direkte Bearbeitung in XML nicht zumutbar ist.

4. Zusammenfassung und Fazit

Wie im zweite Kapitel gezeigt wurde, verstecken sich unter den Bezeichnungen wie Thesaurus, Klassifikation, Nachschlagewerk, Begriffsnetz, Ontologie etc. ähnliche Systeme zur Repräsentation von Wissen. Obwohl sie sich unter anderem in Einsatzgebiet und Ausdrucksstärke unterschieden, lassen sich gemeinsame Grundbestandteile ausmachen, die im wesentlichen aus Begriffen, Bezeichnungen und Relationen bestehen.

Arten und Aufbau von Begriffsystemen

Grundlegend für sämtliche Begriffssysteme ist die Unterscheidung von Begriffen und ihren Bezeichnungen. Dabei kann es unter anderem vorkommen, dass eine Bezeichnung mehreren Begriffen zugeordnet wird. Diese Homonymie lässt sich unter anderem durch den Einsatz von Qualifikatoren bzw. Namensräumen lösen. Um umgekehrten Fall besitzt ein Begriff mehrere Bezeichnungen, von denen in der Regel eine als Vorzugsbenennung ausgewählt ist. Die Vorzugsbenennung kann auch je nach Kontext unterschiedlich sein. Es lassen sich verschiedenen Arten von Bezeichnungen unterscheiden. Neben natürlich sprachlichen Ausdrücken (Namen, Abkürzung, Übersetzung…) können auch künstlichen Zeichen und Zeichenfolgen benutzt werden, die ggf. bedeutungshinweisende Bestandteilen beinhalten (Notationen, Identifikatoren…). Zur übersichtlichen Auflistung von Bezeichnungen sollten gewisse Ordnungs- und ggf. Permutationsregeln beachten werden.

Mit Relationen werden einzelne Begriffe in einem Begriffssystem miteinander verknüpft, wodurch ein komplexes Begriffsnetz gebildet werden kann. Relationen sind in der Regel gerichtete Verbindungen mit der Angabe einer Relationsart, die die Bedeutung der Relation bestimmt.

Relationsarten sind ebenso wie Begriffe festgelegte Konzepte, deren Umfang je nach Begriffssystem variieren kann. In den meisten Fällen werden die möglichen Relationsarten von vorne herein festgelegt. In diesem Fall werden sie oft als Teil des Metamodells bezeichnet. Die wesentlichen Relationsarten sind Hierarchische Relationen, Ordnungsrelationen, Eigenschaftsrelationen und Koordinative Relationen. Je nach Anwendung wird feiner unterschieden. Die hierarchischen Relation zerfällt beispielsweise in Instanziierung, Vererbung und partitive Relation. Die Einteilung von Relationsarten ist jedoch kein allgültiges Schema, sondern selbst Teil eines Begriffssystems.

Falls bestimmte Relationen formale Eigenschaften besitzen, so dass sie nur in bestimmter Art und Weise verwendet werden dürfen oder weitere Relationen implizieren, so muss dies als Teil des Begriffssystems festgelegt werden. Außerdem müssen Inferenzmechanismen und Testverfahren bereitgestellt werden, um die Konsistenz des Begriffssystems gewährleisten zu können. Für Hierarchische Relationen wird beispielsweise häufig gefordert, dass diese keine Kreise bilden dürfen. Ein weiteres Beispiel ist die Einführung von Datentypen, mit denen bestimmte Anforderungen an die Verwendung der Eigenschaftsrelation gestellt werden. Allgemeine Relationen, die mehr als zwei Begriffe miteinander verbinden sind i.d.R. nicht ohne weiteres möglich, sondern müssen durch Relationen über Relationen simuliert werden, was nur von komplexeren Systemen (Ontologien) unterstützt wird. In einfacheren Systemen sind Angaben über Relationen (z.B. wann und von wem wurde welche Verknüpfung erstellt) höchstens Bestandteil eines übergeordneten Modelles.

Eine spezielle Methode, mehr als zwei Begriffe miteinander zu verbinden ist die Begriffskombination. Dabei werden mehrere Begriffe durch Kombination zu einem neuen Begriff zusammengefasst, der wiederum in Relation mit anderen Begriffen gesetzt werden kann. Dies spielt vor allem bei Anwendung der Äquivalenzbeziehung zwischen verschiedenen Begriffsystemen (Mapping) eine Rolle, da sich nicht immer direkte 1-zu-1-Beziehungen finden lassen.

Neben einfachen Anforderungen an einige Relationen können in Ontologien und Expertensystemen auch komplexere Schlussfolgerungsregeln mit aufgenommen werden. Von dieser Möglichkeit wird jedoch nur in begrenztem Umfang Gebrauch gemacht, da sich Regeln schnell komplex werden und sich schwer warten lassen. Entgegen einiger Verheissungen des Semantic Web ist nicht zu erwarten, dass Menschen im grossen Masstab mit formmallogischen Regeln umgehen werden können.

Soll ein Begriffsystem auch von bzw. für Menschen erstell- und anwendbar sein, so spielen natürlich sprachliche Anteile eine viel wesentlichere Rolle. In Hypertexten kann mittels (ungetypter) Verweise auch auf weitere Begriffe verwiesen werden. Ebenfalls von Bedeutung sind Quellen- und Literaturverweise, sowohl als Beleg als auch als Hinweis auf weiterführende Informationen.

Datenformate und Anwendung

Im dritten Kapitel wurden eine Reihe von existierenden Datenformaten für verschiedene Arten von Begriffssystemen vorgestellt. Die Sammlung umfasst XML-basierte Textformate, Formate für Thesauri und Terminologien sowie RDF und Topic Maps. Unter den Textformaten (XHTML, TEI, DocBook) sind einige bereits seit Jahren etabliert und werden je nach Anwendungsfall durch weitere Formate (OpenOffice, NITF, ISO 12083, DiML…) ergänzt. Obwohl Thesauri im Vergleich zu Texten eine ziemlich einheitliche und einfachere Struktur haben, hat sich trotz verschiedener Vorschläge bislang kein einfaches Austauschformat durchsetzen können. Mögliche Gründe dafür liegen unter anderem darin, dass Thesauri in der Regel über einen längeren Zeitraum entwickelt und benutzt werden und somit nicht so schnell auf neue Entwicklungen eingehen können. Außerdem liegt der freie Datenaustausch nicht immer im Interesse sowohl der Hersteller von Thesaurus-Software als auch der Thesauri entwickelnden Institutionen. Dieses Problem, dass auch bei der Verwendung anderer Begriffssysteme wie Klassifikationen und Ontologien im Semantic Web eine Rolle spielt, lässt sich nur durch den konsequenten Einsatz von Open Content – Lizenzen überwinden.

Die aus der Informatik stammende Datenformate RDF und TopicMaps sind von ihrer Ausdrucksstärke am weitesten entwickelt. Sie haben jedoch den Nachteil, dass sie Stellenweise zu komplex sind und ein formmallogisches Modell zugrundelegen, dass bei der Modellierung unscharfer und widersprüchlicher Informationen Probleme bereiten kann. Es ist zu beachten dass sich speziell hier noch vieles in der Erforschung und Entwicklung befindet und die praktikablen Einsatzmöglichkeiten von RDF und TopicMaps erst ausgelotet werden müssen.

Wesentlicher als das konkrete Format sind in der Praxis eher die Werkzeuge, mit denen sich die in verschiedenen Formaten abgelegten Begriffssysteme erstellen, warten und anwenden lassen. Hier fehlt es (außer für Textformate) praktisch in allen Bereichen an einfachen, frei verfügbaren Editoren.[24] Solange bspw. zur fachgerechten Annotierung von Dokumenten Spezialsoftware notwendig ist und die grossen Schlagwortkataloge und Klassifikationen nicht frei verfügbar sind, kann nicht erwartet werden, dass Autoren von (elektronischen) Publikationen ihre Dokumente selbst mit aussagekräftigen Metadaten annotieren. Zur Verwendung von Formaten muss außerdem sichergestellt werden, dass diese auch konsistent eingehalten werden. Auf der syntaktischen Ebene bietet XML mit DTD und XML Schema dazu bereits einen Mechanismus an. Falls das Datenformat weitere formal überprüfbare Semantik enthällt, müssen Mittel bereitgestellt werden, um diese zu validieren. Wichtiger als theoretische Aspekte und ausgeklügelte formale Systeme ist in der Praxis auch die Wartung und Vermittlung von Begriffssystemen. Auf dabei möglicherweise relevante Ergebnisse der Kognitions- und Lerntheorie sowie der praktischen Didaktik konnte zumindest in der vorliegenden Arbeit ebenso wie auf Methoden der Computerlinguistik und auf philosophische Aspekte nicht näher eingegangen werden.

24 Ein positives Gegenbeispiel ist der freie Ontologieeditor Protégé (http://protege.stanford.edu/).

Literatur und Quellen

[**Kuhn**] Thomas S. Kuhn: *Die Struktur wissenschaftlicher Revolutionen.* Suhrkamp, 2002

[**Neveling**] Ulrich Neveling, Gernot Wersig: *Terminologie der Information und Dokumentation.* Verlag Dokumentation, 1975

[**Umstätter**] Umstätter: *Digitales Lehrbuch der Bibliothekswissenschaft.* http://www.ib.hu-berlin.de/~wumsta/infopub/bookindex.html

[**Bowker**] Geoffrey C. Bowker, Susan Leigh Star: *SortingOut: Classification and its consequences*. MIT Press, 1999

[**Wikipedia**] Wikimedia-Stiftung (Hrsg.): *Wikipedia*. seit 2001. http://www.wikipedia.org

[**Engelbert**] Stefan Engelbert, Lothar Lemnitzer: *Lexikographie und Wörterbuchbenutzung.* Stauffenburg, 2001

[**Eversberg**] Bernhard Eversberg: *Was sind und was sollen Bibliothekarische Datenformate.* . http://www.allegro-c.de/allegro/formate/formate.htm

[**Kokkelink**] Stefan Kokkelink, Roland Schwänzl: *Expressing Qualified Dublin Core in RDF / XML*. 15.5.2002. http://dublincore.org/documents/dcq-rdf-xml/

[**Wersig**] Gernot Wersig, Ulrich Neveling: *Thesaurus-Leitfaden*. Saur, 1985

[**Wolters**] Christof Wolters: *GOS Thesaurus-Handbuch*. Konrad-Zuse-Institut Berlin, 1997 (Technical Report TR 97-19)

[**Sowa**] John F. Sowa: *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing, 2000

[**TBL_b**] Tim Berners-Lee: *Conceptual Graphs and the Semantic Web.* 2001. http://www.w3.org/DesignIssues/CG.html

[**TBL_a**] Tim Berners-Lee, Lassila Hendler: *The Semantic Web.* In: Scientific American, May/2001

[**Wikipedia**] Wikimedia-Stiftung (Hrsg.): *Wikipedia*. . http://www.wikipedia.org

[**Eco**] Umberto Eco: *Die Suche nach der Vollkommenen Sprache.* dtv, 1997

[**RSWK**] *Regeln für den Schlagwortkatalog (RSWK).* Deutsches Bibliotheksinstitut, 1998

[**Kokkelink**] Stefan Kokkelink, Roland Schwänzl: *Expressing Qualified Dublin Core in RDF / XML*. 15.5.2002

[**DIN1463-2**] *Erstellung und Weiterentwicklung von Thesauri - MehrsprachigeThesauri.*

[**ISO5964**] *Documentation - Guidelines for the establishment and developmentof multilingual thesauri*.

[**Doerr**] Martin Doerr, Irini Fundulaki: *A proposal on extended interthesaurus links semantics*. FORTH Institute of Computer Science, 1998 (Technical report FORTH-ICS/TR-215)

[**MACS**] http://infolab.kub.nl/prj/macs/

[**Hoppen**] Jeroen Hoppenbrouwers: *Architecture of the MACS System*. 2001. http://infolab.kub.nl/prj/macs/pub/architecture.pdf

[**Doerr_a**] Martin Doerr: *Semantic Problems of Thesaurus Mapping.* In: Journal of Digital Information, 8/1, 2000

[**Noy**] Natalya F.Noy, Mark A. Musen: *SMART: Automated Support for Ontology Merging and Alignment*. In: Stanford Medical Informatics Reports, SMI-1999-0813/, 1999

[**Doan**] AnHaie Doan et. al.: *Learning to Map between Ontologies on the Semantic Web.* In: Proceedings of the Eleventh International World Wide Web Conference, /, 2002

[**Volz**] Raphael Volz, Stefan Decker, Daniel Oberle: *Bubo - Implementing OWL in rule-based systems*. (2003) Unveröffentlichtes Paper für die WWW2003http://www.daml.org/listarchive/joint-committee/att-1254/01-bubo.pdf

[**Wilson**] B.M.Matthews, K. Miller, M.D.Wilson: *A Thesaurus Interchange Format in RDF*. . http://www.limber.rl.ac.uk/External/SW_conf_thes_paper.htm

[**Lee**] Maria Lee, Stewart Baillie, Jon Dell'Oro: *TML: A Thesaural Markup Language.* In: Proceedings of the 4th Australasian Document Computing Symposium, /, 1999

[**Powders**] Shelley Powers: *Practical RDF*. O'Reilly, 2003

[**Garshol**] Lars Marius Garshol: *Living with topic maps and RDF*. 2003. http://www.ontopia.net/topicmaps/materials/tmrdf.html

[**Widhalm**] Richard Widhalm, Thomas Mück: *Topic Maps: Semantische Suche im Internet.* Springer, 2002

[**Park**] Jack Park , Sam Hunting , Douglas Engelbart : *XML Topic Maps: Creating and Using Topic Maps for the Web*. Addison-Wesley, 2002

Anhang I: Das vereinfachte Dateiformat (DTD)

themaSimple.dtd

```
<!--.....................................................................
This is a selection of the DTD
  "SimpleThema"
using the following modules:
 - thema 0.9
 - themaSimple 3.9.16
Time of generation: 16-9-2003
.....................................................................-->
<!--- PARTS OF MODULE thema-->
<!ELEMENT thema (title | role | abbrev | long |
                 description | etymology | url | bibliography |
                 property | related | see | seenBy | next | prev |
                 homonym | antonym | synonym | prefere | preferedBy |
                 subcon | supcon | subcons | supcons |
                 thema | concept)*>
<!ATTLIST thema
  id ID #REQUIRED
  lang NMTOKEN #IMPLIED
  notation CDATA #IMPLIED
  hidden (yes|no) 'no'
>
<!ELEMENT concept (title | role | abbrev | long |
                   description | etymology | url | bibliography |
                   property | related | see | seenBy | next | prev |
                   homonym | antonym | synonym | prefere | preferedBy |
                   subcon | supcon | subcons | supcons)*>
<!ATTLIST concept
  id ID #REQUIRED
  lang NMTOKEN #IMPLIED
  notation CDATA #IMPLIED
  topterm (yes|no) #IMPLIED
  hidden (yes|no) 'no'
>
<!ELEMENT description (p)+>
<!ATTLIST description
  lang NMTOKEN #IMPLIED
  role IDREF #IMPLIED
  label CDATA #IMPLIED
>
<!ELEMENT etymology (#PCDATA)*>
<!ATTLIST etymology
  lang CDATA #REQUIRED
  word CDATA #REQUIRED
>
<!ELEMENT property (#PCDATA)*>
<!ATTLIST property
  role IDREF #IMPLIED
  ref IDREF #IMPLIED
  label CDATA #IMPLIED
>
<!ELEMENT related  EMPTY>
<!ATTLIST related
  role IDREF #IMPLIED
  ref IDREF #IMPLIED
  label CDATA #IMPLIED
>
<!ELEMENT see  EMPTY>
<!ATTLIST see
  ref IDREF #IMPLIED
```

```
60  role IDREF #IMPLIED
61>
62<!ELEMENT seenBy  EMPTY>
63<!ATTLIST seenBy
64  ref IDREF #IMPLIED
65  role IDREF #IMPLIED
66>
67<!ELEMENT next  EMPTY>
68<!ATTLIST next
69  role IDREF #IMPLIED
70  ref IDREF #IMPLIED
71  label CDATA #IMPLIED
72  context IDREF #IMPLIED
73>
74<!ELEMENT prev  EMPTY>
75<!ATTLIST prev
76  role IDREF #IMPLIED
77  ref IDREF #IMPLIED
78  label CDATA #IMPLIED
79  context IDREF #IMPLIED
80>
81<!ELEMENT homonym  EMPTY>
82<!ATTLIST homonym
83  ref IDREF #REQUIRED
84  role IDREF #IMPLIED
85>
86<!ELEMENT antonym  EMPTY>
87<!ATTLIST antonym
88  ref IDREF #REQUIRED
89  role IDREF #IMPLIED
90  label CDATA #IMPLIED
91>
92<!ELEMENT synonym  EMPTY>
93<!ATTLIST synonym
94  ref IDREFS #REQUIRED
95  role IDREF #IMPLIED
96  label CDATA #IMPLIED
97>
98<!ELEMENT prefere  EMPTY>
99<!ATTLIST prefere
100  ref IDREFS #REQUIRED
101  role IDREF #IMPLIED
102  label CDATA #IMPLIED
103>
104<!ELEMENT preferedBy  EMPTY>
105<!ATTLIST preferedBy
106  ref IDREF #REQUIRED
107  role IDREF #IMPLIED
108  label CDATA #IMPLIED
109>
110<!ELEMENT subcon  EMPTY>
111<!ATTLIST subcon
112  ref IDREF #REQUIRED
113  role IDREF #IMPLIED
114  label CDATA #IMPLIED
115  context IDREF #IMPLIED
116>
117<!ELEMENT supcon  EMPTY>
118<!ATTLIST supcon
119  ref IDREF #REQUIRED
120  role IDREF #IMPLIED
121  label CDATA #IMPLIED
122  context IDREF #IMPLIED
```

```
123>
124<!ELEMENT subcons (term)+>
125<!ATTLIST subcons
126  role IDREF #IMPLIED
127  label CDATA #IMPLIED
128  context IDREF #IMPLIED
129  order IDREF #IMPLIED
130  classification (yes|no) 'no'
131  antonymity (yes|no) 'no'
132  antonymRol IDREF #IMPLIED
133  sorted (yes|no|circle) 'no'
134>
135<!ELEMENT supcons (term)+>
136<!ATTLIST supcons
137  role IDREF #IMPLIED
138  label CDATA #IMPLIED
139  context IDREF #IMPLIED
140  order IDREF #IMPLIED
141  classification (yes|no) 'no'
142  antonymity (yes|no) 'no'
143  antonymRol IDREF #IMPLIED
144  sorted (yes|no|circle) 'no'
145>
146<!ELEMENT title (#PCDATA)*>
147<!ATTLIST title
148  ref IDREF #IMPLIED
149  qualifier CDATA #IMPLIED
150  qref IDREF #IMPLIED
151  role IDREF #IMPLIED
152  context IDREF #IMPLIED
153  homonym IDREFS #IMPLIED
154  lang NMTOKEN #IMPLIED
155>
156<!ELEMENT role (#PCDATA)*>
157<!ATTLIST role
158  ref IDREF #IMPLIED
159  qualifier CDATA #IMPLIED
160  qref IDREF #IMPLIED
161  role IDREF #IMPLIED
162  context IDREF #IMPLIED
163  homonym IDREFS #IMPLIED
164  lang NMTOKEN #IMPLIED
165  direction (forward|back|bidirectional) 'forward'
166>
167<!ELEMENT abbrev (#PCDATA)*>
168<!ATTLIST abbrev
169  ref IDREF #IMPLIED
170  qualifier CDATA #IMPLIED
171  qref IDREF #IMPLIED
172  role IDREF #IMPLIED
173  context IDREF #IMPLIED
174  homonym IDREFS #IMPLIED
175  lang NMTOKEN #IMPLIED
176  type (abbreviation|acronym) #IMPLIED
177>
178<!ELEMENT long (#PCDATA)*>
179<!ATTLIST long
180  ref IDREF #IMPLIED
181  qualifier CDATA #IMPLIED
182  qref IDREF #IMPLIED
183  role IDREF #IMPLIED
184  context IDREF #IMPLIED
185  homonym IDREFS #IMPLIED
```

```
186  lang NMTOKEN #IMPLIED
187>
188
189<!--- PARTS OF MODULE themaSimple-->
190<!ELEMENT bibliography (p)+>
191<!ELEMENT p (#PCDATA)*>
192<!ELEMENT term (#PCDATA)*>
193<!ATTLIST term
194  ref IDREF #IMPLIED
195>
196<!ELEMENT url (#PCDATA)*>
197<!ATTLIST url
198  href CDATA #REQUIRED
199  type (URL | DOI | URN | ISBN | ISSN | ISO | RFC) 'URL'
200>
```

Bei der hier abgedruckten DTD handelt es sich um die vereinfachte Version des Thema-Datenformates („SimpleThema“) mit 27 Elementen. Gegenüber der vollständigen Thema-DTD, die ebenso wie ein XML Schema unter http://thema.sourceforge.net verfügbar ist, wurde auf alle Elemente zur Textauszeichnung verzichtet. Das vollständige Thema-Datenformat enthält weitere, aus der DiML-DTD entnommene Elemente. Dies betrifft die Datentypen *Text*, *Document* und *Bibliography* im Thema-Datenmodell (3.2). Alle zum vereinfachten Thema-Dateiformat validen Dokumente sind auch im vollständigen Thema-Dateiformat valide.

Anhang II: Urheberrechtsvermerk und Copyleft