Entropiebasierte Bewertung von Ontologien

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
derFakultät für Informatik
der Universität Karlsruhe (Technische Hochschule)

genehmigte

Dissertation

von

Anusch Daemi-Ahwazi

aus Teheran

Tag der mündlichen Prüfung:	07.06.2005
Erster Gutachter:	Prof. Dr. Jacques Calmet
Zweiter Gutachter:	Prof. Dr. Gabriele Kern-Isberner

Danksagung

Für den Abschluss der vorliegenden Arbeit, die von Oktober 2002 bis Mai 2005 am Institut für Algorithmen und Kognitive Systeme der Universität Karlsruhe (TH) angefertigt wurde, schulde ich vielen Menschen meinen herzlichen Dank.

An erster Stelle möchte ich meinem Doktorvater, Herrn Professor Dr. Jacques Calmet, von der Universität Karlsruhe (TH) danken, der mir viel Geduld und Vertrauen entgegengebracht und mich in schwierigen Situationen uneingeschränkt unterstützt hat. Frau Professor Dr. Gabriele Kern-Isberner von der Universität Dortmund gilt mein Dank für die Übernahme des Zweitgutachtens.

Herrn Professor Dr. Thomas A. Troge von der Musikhochschule Karlsruhe möchte ich für die Unterstützung meiner Arbeit im musikalischen Bereich danken, für die er wertvolle Hinweise und Ideen lieferte.

Bei Herrn Professor Dr. Oswald Drobnik, von der Universität Frankfurt, bedanke ich mich ganz herzlich für die Deutschkorrektur und die vielen hilfreichen und nützlichen Anmerkungen zur vorliegenden Arbeit.

Ein Dank geht auch an Frau Dr. Charlotte Kämpf, Herrn Dr. Jürgen Ihringer und die Mitarbeiter des Instituts für Wasserwirtschaft und Kulturtechnik der Universität Karlsruhe, die mich bei den auftretenden, hydrologischen Fragestellungen dieser Arbeit unterstützten.

Ein großes Dankeschön geht an meinen Kollegen, Herrn Dipl.-Inform. Thilo Mie, der mir mit seiner fachlichen Unterstützung tatkräftig zur Seite stand und bei der Korrektur dieser Arbeit wertvolle Dienste leistete.

Bei Herrn Dipl.-Inform. Stefan Kink möchte ich mich für die lebhaften Diskussionen und die Unterstützung meiner Arbeit im musikalischen Bereich bedanken.

Last, but not least habe ich mich immer auf die Unterstützung meiner Freundin, Stefanie Drobnik, verlassen können. Sie hat mir in schwierigen Situationen die nötige Kraft und Energie zum weitermachen gegeben. Ihr möchte ich diese Arbeit widmen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

In dieser Arbeit wurde eine Vereinheitlichung des Wissens über Flutkatastrophen angestrebt. Hintergrund dieser Entwicklung ist, dass eine effektive Kommunikation und Wissensvermittlung für eine gute Koordination von Hilfsmaßnahmen bei Naturkatastrophen unerlässlich ist. Dies zeigte vor allem die Flutkatastrophe an der Elbe im Sommer 2002, wie im Bericht von [Kirchbach 2003] dargestellt wird.

Die auftretenden Verständigungsprobleme resultieren unter anderem daraus, dass bei dem Risikomanagement für Umweltkatastrophen, welches sowohl natürliche als auch anthropogene Risiken umfasst, viele verschiedene Experten aus unterschiedlichen Domänen ihr Wissen austauschen. Es müssen beispielsweise Fachleute aus dem Wasserbau, dem Ingenieursbereich als auch Biologen hinzugezogen werden, die sich in kompetenter Weise um Dämme und Stauseen, Sicherung kritischer Industrieobjekte bzw. Abschätzung der Seuchengefahr durch Tierkadaver kümmern. Die Erzeugung, Kommunikation und Anwendung des von den Experten generierten Wissens wurde durch die „Internationale Dekade zur Verringerung von Naturgefahren“ von 1989 bis 1999 [Unit 1989] (International Decade for Natural Disaster Reduction - IDNDR) durch die Vereinten Nationen gefördert.

Resultate der IDNDR hinsichtlich des Risikomanagements von Flutkatastrophen sind beispielsweise die Echtzeitvorhersage des Abflusses und der Spitzenwerte der Flutwellen, GIS basierte Kartierungen und automatische Risikoeinschätzungen von überschwemmungsgefährdeten Gebieten an deutschen Flüssen [Oberle u. a. 2000]. Diese dienen, ebenso wie verschiedene Expertensysteme [Kämpf u. a. 2002], zur Unterstützung der lokalen Behörden. Einen Überblick die Ergebnisse der IDNDR in Deutschland liefert [Plate u. Merz 2001].

Zur Strukturierung und Vereinheitlichung dieses Expertenwissens wurde erstmals eine Ontologie für das Risikomanagement von Flutkatastrophen, im folgenden abgekürzt als Flutontologie, erstellt. Die Experten sind in diesem Falle Hydrologen des Institutes für Wasserwirtschaft und Kulturtechnik

(IWK) der Universität Karlsruhe (TH), die diese Doktorarbeit im Rahmen des Graduiertenkollegs Naturkatastrophen unterstützten. Einleitend werden die Kernbegriffe kurz angerissen und der Aufbau der Arbeit dargestellt.

1.1 Ontologie

Der Begriff Ontologie, vom griechischen *ontos* = das Sein, *logos* = Abhandlung abstammend, hat seinen Ursprung in der Philosophie. Dort ist es die Lehre über das *Sein* und den *Zusammenhang* der Dinge in unserer Welt [Gómez-Pérez u. a. 2004]. Aufgrund der Überlappungen zwischen Philosophie und künstlicher Intelligenz (KI) [McCarthy 1995] fand der Begriff in den letzten Jahren Einzug in die Wissensverarbeitung (knowledge engineering) als auch Wissensrepräsentation (knowledge representation). Dieser Arbeit liegt die Definition von [Studer u. a. 1998] aus der angewandten Informatik zugrunde, welcher auf [Gruber 1993b] zurückzuführen ist:

> *An ontology is a formal, explicit specification of a shared conceptualization.*

Eine Konzeptualisierung (*conceptualization*) beschreibt ein abstraktes Modell eines Phänomens in der Welt, wobei die relevanten Konzepte des Phänomens bereits identifiziert wurden. Die Bedeutung von *Formal* liegt darin, dass die Repräsentation der Ontologie maschinenlesbar sein soll. Explizit (*explicit*) bedeutet, dass die Konzepte, die Einschränkungen denen sie unterliegen und die Relationen zwischen ihnen ausdrücklich beschrieben und definiert werden. Gemeinsam (*shared*) soll verdeutlichen, dass die Konzepte bzw. das Wissen, welches durch die Ontologie repräsentiert wird, von einer hinreichend großen Gemeinde akzeptiert und als tragfähig für diesen Zwecke befunden worden ist. Nachteilig an dieser Definition ist jedoch die Verwendung des Begriffes *conceptualization*, da dieser nicht hinreichend genau erklärt wird, bzw. auf mehrdeutigen Begriffen basiert Bittner u. a. [2004].

1.2 Strukturierung von Wissen

Obwohl Ontologien einen entscheidenden Beitrag zur Vereinheitlichung und Strukturierung des durch sie repräsentierbaren Wissens darstellen, können sich bei ihrer Benutzung verschiedene Probleme ergeben. Das in der Ontologie repräsentierte Wissen wird normalerweise als vollständig für den jeweiligen Anwendungsbereich betrachtet. Ebenso sollte dieses Wissen als sicher gelten, da es von einer hinreichend großen Expertengruppe akzeptiert sein sollte. Auch sind Ontologien oder Taxonomien per Definition bereits nach einem oder mehreren Kriterien strukturiert, wie zum Beispiel anhand einer Vererbungsrelation. Dennoch muss die sich damit ergebende Struktur nicht notwendigerweise eindeutig sein. Das wohl prominenteste Beispiel für eine

nicht eindeutige Struktur ist der Nixon Diamant nach [Reiter 1980]. Ein weiteres Beispiel wäre die Frage nach dem kleinsten gemeinsamen Elternkonzept bei Mehrfachvererbung in einer durch eine Vererbungsrelation strukturierten Ontologie. Auch kann eine Strukturierung des Wissensraumes, welcher durch eine Ontologie aufgespannt wird, bestimmt werden. Der von der Ontologie dargestellte Wissensraum kann beispielsweise das subjektive Wissen von Politikern und Experten zur Minderung einer Hochwasserkatastrophe sein. Dieser kann strukturiert werden, um Ähnlichkeiten und Unterschiede zwischen dem Wissen aufzuzeigen.

Als Grundlage für eine Strukturierung der Ontologien bzw. deren Wissensbasen dienen Wahrscheinlichkeitsverteilungen, welche ein (komprimiertes) Datenmodell für das Wissen darstellen. Die konkreten Wahrscheinlichkeiten der Verteilungen werden, je nach Anwendungsfall, den Konzepten (Knoten) oder den Relationen zwischen den Konzepten (Kanten) zugewiesen. Die Wahrscheinlichkeitsverteilung muss dabei nicht notwendigerweise eine objektive, d.h. frequenzbasierte Interpretation aufweisen. Sie kann auch als *Degree-of-Belief* oder *Grad-des-Dafürhaltens* interpretiert werden. Damit dient die Wahrscheinlichkeitsverteilung zur Verkörperung von Ungewissheit unter Berücksichtigung bereits vorhandenen Wissens [Beyerer 1999]. Die Deutung des Degree-of-Belief lässt sich weiter zergliedern in einen objektiven und subjektiven Fall. Bei der *objektiven* Auffassung des Degree-of-Belief sollen Fakten, die zu einer Meinung Anlass geben, wahrscheinlichkeitstheoretisch verkörpert werden. Bei der *subjektiven* Deutung hingegen dient die Wahrscheinlichkeit zur Darstellung des Wissensstandes und der Überzeugungen eines Individuums. Hierbei spielt es keine Rolle, ob die Wahrscheinlichkeiten „wahr“ oder nur das Individuum sie für richtig hält.

Ausgehend von der natürlichen Sprache existieren bereits einige Maße [Resnik 1995; Corman u. a. 2002], um eine Strukturierung von Wissen, basierend auf Wahrscheinlichkeiten, vorzunehmen. Die Methoden aus der Computerlinguistik basieren meistens auf der Häufigkeit von Begriffen [Agirre u. Rigau 1996], die anhand mehrerer Millionen Wörter umfassenden Textkorpora [Fancis u. Kucera 1982] bestimmt wird. Techniken aus der Statistik [Yang 1999], des Information Retrieval (IR) mittels frequenz- und vektorbasierten Maßen [Dutoit u. Poibeau 2002; Hotho u. a. 2002] und Data-Minings [Celeux u. a. 1989] ermöglichen ebenfalls eine Strukturierung von Wissen. Diese Maße wurden für ihre Anwendung zur Strukturierung des durch die Ontologien repräsentierten, konkreten Wissens untersucht. Die größte Schwierigkeit bezüglich der Anwendung dieser Strukturierungsmaße lag darin, dass für deren Anwendung sehr große Datenmengen gebraucht werden, um zu konsistenten Ergebnissen zu gelangen. Solche großen Datenmengen sind für die Strukturierung von Ontologien aber nicht notwendigerweise gegeben. Desweiteren konvergiert bei frequenzbasierten Maßen der Bereich, in dem die meisten relevanten Begriffe vorkommen, nicht sonderlich schnell und die genaue Bestimmung dieses Bereiches ist ebenfalls ein nicht triviales Problem. Die sta-

tistischen und informationstheoretischen Eigenschaften der Wahrscheinlichkeitsverteilungen, auf denen viele der erwähnten Maße basieren, werden bei der Berechnung der Distanz ebenfalls nicht hinreichend berücksichtigt.

Zur Behebung dieser Problematik wurde erstmals die Anwendung entropiebasierter Distanzmaße für die Strukturierung des durch eine Ontologie repräsentierten Wissens untersucht.

1.3 Entropie

Das Konzept der Entropie existiert bereits seit Mitte des 19ten Jahrhunderts und bildet den Kern des zweiten Gesetzes der Thermodynamik [Maxwell 1871]:

1. Die Energie des Universums ist konstant.
2. Die Entropie des Universums strebt einem Maximum zu.

Nach einigen Jahren [Uffink 2001] fand die Entropie Einzug in die statistische Mechanik. In diesem Gebiet ist die Entropie, kurz gesagt, das Maß der Unordnung eine entscheidende Größe zur Bestimmung des jeweiligen Zustands eines Systems.

Davon ausgehend entstand Anfang bis Mitte des 20ten Jahrhunderts die informationstheoretische Definition der Entropie nach Shannon und Hartley [Shannon 1948; Hartley 1928]. Die Gleichungen und Herangehensweise sind denen in der Thermodynamik und statistischen Mechanik ähnlich, die zugrundeliegende Idee ist jedoch eine andere. In der Informationstheorie spielt bei der Definition der Entropie die Unsicherheit ebenfalls eine große Rolle, allerdings in einem etwas anderen Zusammenhang. Nach [Shannon 1948] ist die Entropie einer Nachricht ihr *Informationsgehalt*, oder, aus der Sicht des Empfängers, die Unsicherheit über die vom Sender produzierten Nachricht, bevor sie empfangen wurde. Die formale Definition der Entropie in der statistischen Mechanik und der Informationstheorie sind nahezu gleich, jedoch sind die Herkunft und Idee verschieden voneinander [Pierce 1980].

Die zwei wichtigsten, entropiebasierten Distanzmaße, welche die Verbindung zu den Ontologien herstellen, sind die gegenseitige Information und die relative Entropie [Cover u. Thomas 1991].

Die *gegenseitige Information* ist ein Maß für die Information, die eine Zufallsvariable Y über eine andere Zufallsvariable X enthält, d.h. wieviel Bits an Information erhält man über X, wenn zusätzlich noch Y bekannt und nicht unabhängig von X ist. Die Dualität von Information und Unsicherheit berücksichtigend, kann man auch sagen, dass die gegenseitige Information die Verringerung der Unsicherheit der einen Zufallsvariable durch Kenntnis der anderen beschreibt.

Das zweite verwendetes Maß ist die *relative Entropie*. Die relative Entropie ist ein Maß, um die Distanz zwischen zwei Wahrscheinlichkeitsverteilun-

gen **p** und **q** zu berechnen. Dabei ist zu beachten, dass die relative Entropie, auch bekannt unter dem Namen Kullback-Leibler Distanz und minimaler Kreuzentropie, kein Distanzmaß im axiomatischen Sinne ist. Sie erfüllt nur das Axiom der positiven Definitheit, die Symmetrie- und Dreiecksungleichsaxiome werden nicht erfüllt. Weiterhin gehören die relative Entropie zu den informationstheoretischen Distanzmaßen nach [Ali u. Silvey 1966], welche einige nützliche Eigenschaften besitzen.

1.4 Strukturierung von Ontologien durch Entropie

In dieser Arbeit werden zwei entropiebasierte Distanzmaße, die gegenseitige Information und relative Entropie, als Strukturierungsmöglichkeiten für Wissen, welches durch Ontologien dargestellt werden kann [Staab u. Studer 2004], untersucht. Für das durch die Ontologie repräsentierte Wissen sei dabei eine Wahrscheinlichkeitsverteilung gegeben, wie in Abschnitt 1.2 beschrieben wurde. Die Wahrscheinlichkeitsverteilung kann dabei eine frequenzbasierte Bedeutung besitzen oder auch als Degree-of-Belief interpretiert werden. Distanzmaße wurden als Strukturierungsmöglichkeit auch deshalb gewählt, da mit ihnen eine Klassifizierung und damit auch Strukturierung des Wissens in Gruppen (*Cluster*) möglich ist [MacKay 2003].

1.4.1 Theoretisches Modell - Gegenseitige Information

Zunächst wird ein Modell zur Strukturierung von Wissen mittels entropiebasierter Distanzmaße dargelegt. Wissen wird im Sinne des Karlsruher Ansatzes zur integrierten Wissensforschung [Weber u. a. 2002] als semantische Information gesehen, wobei Information als Möglichkeitsausschluss in einem Wissenraum definiert ist. Der Möglichkeitsausschluss lässt sich durch Wahrscheinlichkeitsverteilungen darstellen, auf denen aufbauend entropiebasierte Distanzmaße zur Strukturierung von Wissen definiert werden können. Insbesonders auch deshalb, weil die semantische Information den „Informationsgehalt“ einer Aussage angibt, welcher definiert ist als die Klasse der ausgeschlossenen, d.h. aussagenkonträren logischen bzw. empirischen Möglichkeiten im jeweiligen Wissensraum.

Als erstes Maß wurde die gegenseitige Information zur Strukturierung von Wissen, dass mit Ontologien dargestellt werden kann, untersucht. Die gegenseitige Information ist ein Maß für die Information, die eine Zufallsvariable X über eine andere Zufallsvariable Y enthält. Die Dualität von Information und Unsicherheit berücksichtigend, kann man auch sagen, dass sie die Verringerung der Unsicherheit der einen Zufallsvariable durch Kenntnis der anderen beschreibt. Dieses Konzept wurde für die Anwendung auf Ontologien verfeinert, indem die Zufallsvariablen mit den Begriffen in der Ontologie identifiziert werden. Dafür muss die Ontologie als gerichteter, azyklischer und hierarchische Graph darstellbar sein. Die Knoten stellen die Konzepte

dar, die Relationen zwischen den Konzepten werden durch gerichtete Kanten dargestellt.

Eine Zufallsvariable repräsentiert im einfachsten Fall die Wahrscheinlichkeitsverteilung eines Begriffes und wird dem dazugehörigen Konzept in der Ontologie zugewiesen. Anhand dieser Wahrscheinlichkeiten kann die Entropie von X und Y berechnet werden. Für die Berechnung der gegenseitige Information ist aber lediglich die bedingte Entropie zwischen X und Y wichtig. Sie berechnet sich aus den bedingten Wahrscheinlichkeiten von X und Y und wird den Kanten zwischen den Konzepten zugewiesen.

Normalerweise ist nicht nur ein Begriff vorhanden, der etwas über einen anderen aussagt, sondern mehrere, sodass $\mathbf{Y}$ als Vektor von Begriffen angesehen werden kann. Um die Genauigkeit der Strukturierung des Wissens zu erhöhen soll nicht nur der kürzeste Pfad zwischen zwei Begriffen mit dazwischenliegenden Begriffen Y betrachtet werden, sondern es sollen weitere l Pfade in Betracht gezogen werden. Diese l Pfade sollen dabei noch einen hinreichend großen Informationsgewinn liefern.

Bei der Anwendung dieser Methode stellte sich die Berechnung der bedingten Entropie als problematisch heraus, denn diese basiert auf den bedingten Wahrscheinlichkeiten der Konzepte auf den l Pfaden. Um diese zu berechnen, sind die Verbundwahrscheinlichkeiten des Vektors $\mathbf{Y}$ und X zu berechnen, was sehr aufwändig ist. Aus diesem Grund wurde auf eine konkrete Anwendung dieses Maßes verzichtet und die relative Entropie für Ontologien näher untersucht.

1.4.2 Relative Entropie für Ontologien

Neben ihrer herausragenden Bedeutung in der Statistik und anderen Gebieten, wurde die relative Entropie erstmals als Maß für die Strukturierung von Wissen, welches durch Ontologien dargestellt werden kann, vorgeschlagen. Die Bestimmung der für die Benutzung der relativen Entropie notwendigen Wahrscheinlichkeitsverteilungen erfolgt anhand der in Abschnitt 1.2 vorgestellten Vorgehensweise. Die Wahrscheinlichkeiten werden hierbei den Kanten (Relationen) zwischen den Konzepten zugewiesen, und nicht direkt den Konzepten. Die Bedeutung der Wahrscheinlichkeitsverteilungen ist dabei bewusst offen gehalten worden. Damit sind wiederum objektive, frequenzbasierte als auch subjektive Interpretationen der Wahrscheinlichkeitsverteilungen möglich. Bei einer frequenzbasierten Interpretation ist die Semantik der Wahrscheinlichkeiten eindeutig als relative Häufigkeit gegeben. Damit ist ebenfalls die Semantik der Distanzmaße, insbesondere der relativen Entropie, eindeutig gegeben. Bei der Verwendung von Wahrscheinlichkeiten, welche as Degree-of-Belief interpretiert werden, ist die Semantik der Wahrscheinlichkeiten nicht so eindeutig wie im frequenzbasierten Fall. In diesem Falle kann es beispielsweise vorkommen, dass eine Wahrscheinlichkeit von 0.5 eine bestimme, subjektive Aussage repräsentiert oder „Nichtwissen“ bzw. Ignoranz

über einen bestimmten Sachverhalt. Diese Tatsache muss bei der Berechnung und vor allem der Interpretation der relativen Entropie berücksichtigt werden.

Da die relative Entropie in der Literatur als Distanzmaß [Kotz u. Johnson 1981] angesehen wird, obwohl sie nicht symmetrisch ist und die Dreiecksungleichung nicht erfüllt, können mit ihr Cluster ähnlichen Wissens (kleine Distanz) und unterschiedliche (große Distanz) gebildet werden. Generell lässt sich sagen, dass die relative Entropie angibt, inwieweit informationstheoretisch eine Übereinstimmung zwischen dem durch die Ontologie modellierten Wissen mit der Wahrscheinlichkeitsverteilung **p** und demjenigen Wissen mit Wahrscheinlichkeitsverteilung **q** besteht. Somit gibt die relative Entropie auf Ontologien die durchschnittliche Information an, dass Wissen aus **p** nicht aus einer Verteilung **q** stammt [Kudjoi 2004].

Ein wichtiger Grund für die Verwendung der relativen Entropie als Strukturierungsmaß besteht darin, dass sie den informationstheoretischen Abstand zwischen Datenkompressionsmodellen, dargestellt durch Wahrscheinlichkeitsverteilungen, widergibt und damit die statistischen Eigenschaften der Wahrscheinlichkeitsverteilungen berücksichtigt. Der Abstand ist genau dann Null, wenn beide Modelle exakt gleich sind. Ansonsten ist der Abstand größer Null und wird umso größer, je unterschiedlicher die Modelle zueinander sind [Cover u. Thomas 1991]. Weiterhin ist die relative Entropie additiv, d.h. für zwei unabhängige Wahrscheinlichkeiten ist die Distanz der Verbundverteilungen gleich der Summe der jeweiligen Randverteilungen. Eine weitere nützliche Eigenschaft für die Strukturierung von Ontologien behandelt das *Data Processing Theorem (DPT)*. Es besagt, dass keine statistische Verarbeitung der Daten die relative Entropie erhöht [Kullback u. Leibler 1951]. Dies beinhaltet also die Invarianz gegenüber Operationen wie zum Beispiel Mittelung, Gruppierung oder Verdichtung. Ein letzter Grund für die Verwendung der relativen Entropie ist ihre schnelle und einfache Berechenbarkeit gegenüber anderen, informationstheoretischen Distanzmaßen.

Die soeben besprochene Vorgehensweise wurde anhand eines Beispiels mit der Flutontologie, die im nächsten Abschnitt kurz vorgestellt wird, angewandt. Eine konkrete Anwendung ergab sich bei der Strukturierung von Musikstücken, anhand der im übernächsten Abschnitt vorgestellten Ontologie für die menschliche, musikalische Wahrnehmung.

1.4.3 Flutontologie

In Kooperation mit dem Institut für Wasserwirtschaft und Kulturtechnik (IWK) wurde eine Ontologie für das Risikomanagement im Falle eines Hochwassers, kurz Flutontologie genannt, erstellt. Da für das Risikomanagement von Hochwasser natürliche als auch anthropogene Risiken betrachtet werden müssen, sind die verschiedensten Wissensdomänen an diesem Gebiet beteiligt. Dies sind beispielsweise Ingenieure, die verantwortlich sind für großan-

gelegte, bauliche Maßnahmen wie Dämme oder Polder. Weiterhin sind Chemiker und Biologen beteiligt, welche die Schäden aus eventuell austretenden, toxischen Substanzen von Industrieanlagen oder Seuchengefahren abschätzen. Die Flutontologie bildet somit eine einheitliche, semantische Grundlage für Lokalisierung und Bereitstellung von domänenspezifischem Wissen für Wissenschaftler, welche in diesem Bereich tätig sind. Damit kann die Ontologie auch als Grundlage für ein verteiltes, agentenbasiertes Informationsgewinnungssystem dienen.

In den betrachteten Wissensdomänen wurden einige zentrale Konzepte identifiziert, um eine zuverlässige Modellierung des Diskursbereiches der Ontologie zu erreichen. Das Risikomanagement von Hochwasser muss natürlich die *Umgebung* des Flusses betrachten, d.h. die geographische Struktur, in welcher der Fluss eingebettet ist. Desweiteren müssen so genannte *Felddaten*, die als Parameter in Modelle für die Berechnung des Abflusses einfließen, berücksichtigt werden, wie beispielsweise meteorologische, hydrologische als auch geophysikalische Daten. Der *Abfluss* ist ein elementares Konzept der Hydrologie, welcher es ermöglicht, Flutwellen eines Flusses vorherzusagen. Ein weiterer zentraler Bestandteil der Ontologie sind die *Schutzmaßnahmen* vor Flutkatastrophen. Das sind zum einen die langfristige Präventivmaßnahmen und die kurzfristige Hilfe im akuten Katastrophenfall. Schließlich müssen noch die *administrativen* Daten berücksichtigt werden, beispielsweise Identifikationsnummern von Pegelmessstationen, Metadaten oder Wasserschutzgebiete. Nach der Erstellung der Ontologie wurde deren Benutzbarkeit anhand einer Annotation der Webseiten des USGS überprüft.

1.4.4 Ontologien und Entropie in der Musik

Zur Demonstration der breiten Anwendungsmöglichkeiten der relativen Entropie für Ontologien wurde eine, von der Flutontologie unterschiedliche, Wissensdomäne ausgewählt. Zum einen waren für die Flutdomäne nicht hinreichend genug Daten vorhanden, die strukturiert werden konnten. Zum anderen befasst sich der ausgewählte Wissensraum mit Musik, bzw. der menschlichen, musikalischen Wahrnehmung und stellt somit eine andere Wissensart dar, als das durch die Flutontologie modellierte Wissen. Die Unterschiede bezüglich der Wissensarten manifestieren sich insbesondere in den Dimensionen *Inhalt* (inhaltlich nicht genau bestimmbar vs. informativ) und *Ausdruck* (inkorporiertes Wissen vs. expliziter Ausdruck) [Weber u. a. 2002].

Dazu schlug im Laufe dieser Arbeit schlug bekannter Komponist, Herr Boris Yoffe, ein Modell für die Kreativität in der Musik vor [Yoffe 2004]. Dieses Modell zeigte eine starke Ähnlichkeit mit den Modellen aus der Informatik, insbesondere aus dem Bereich der Agenten [Weiss 2000] und „Emotionen in Multi-Agenten Systemen“ [Petta u. Trappl 2001]. Ein wesentlicher Bestandteil dieser Modelle sind Ontologien, mit denen es möglich ist, einen Kontext und daraus eine (spezialisierte) Semantik zu entwickeln. Im

Hinblick auf den musikalischen Kontext bedeutet dies, dass Kreativität eine Art „Erweiterung“ der zugrundeliegenden Ontologie ist, und damit auch des Kontextes und der Semantik, durch hinzufügen neuer Konzepte. Im musikalischen Bereich ist die daraus neu entstehende Semantik persönlich, da Kunst im allgemeinen sehr subjektiv ist. Ein Problem, dass sich hierbei stellte, war eine bislang fehlende Datenstruktur zur Formalisierung von gehörter Musik. Zwar ist in der Musik die Verwendung von Computern heutzutage alltäglich, sei es zur Erstellung eines Notensatzes, zur Livebearbeitung des Auftrittes einer Musikgruppe oder zur Komposition von elektronischer Musik. All diesen Anwendungen ist aber gemeinsam, dass sie den Musiker bei der Erschaffung oder Bearbeitung von Musik unterstützen. Den umgekehrten Weg (Transkription) zu beschreiten, ist dagegen bedeutend schwieriger [Temperley 2002]. Hierbei soll anhand eines gehörten Musikstückes eine geeignete, formale Repräsentation, beispielsweise durch Noten, gefunden werden. Erste Ansätze zur Transkription decken lediglich einen Teilbereich eines Musikstückes ab, wie beispielsweise die Extraktion des Rhythmus [Gouyon u. Dixon 2005] und konzentrieren sich auf frequenzbasierte Analysen der Musikstücke. Eine formale Darstellung eines gehörten Musikstückes, anhand derer eine abstrakte Repräsentation im Computer möglich ist, soll im folgenden mit Hilfe einer Ontologie ermöglicht werden. Die zuverlässige Modellierung des Diskursbereichs der Ontologie gewährleistete dabei eine enge Zusammenarbeit mit der Musikhochschule Karlsruhe, insbesondere Herrn Prof. Dr. T. Troge.

Zur formalen Darstellung eines gehörten Musikstückes wurden die wichtigsten Konzepte musikalischer Wahrnehmung identifiziert und in der Ontologie dargestellt. Dazu gehört zunächst das *Hörerlebnis*. Das Hörerlebnis fasst alles hörbare zusammen, insbesondere Klänge, Geräusche und auch Pausen. Weiterhin wichtig ist das Konzept *Zeitmaß*, mit dem die Zeiteinteilung eines Stückes beschrieben werden kann. Um eine möglichst detaillierte Zeiteinteilung zu ermöglichen, wird das Zeitmaß durch spezialisiertere Konzepte weiter verfeinert. Dazu dienen Metrum, Takt und Rhythmus, welche die grundlegenden Strukturen eines Stückes bilden. Auch können linear langsamer bzw. schneller werdende Passagen eines Stückes formalisiert werden. Lautstärken bzw. *Lautstärkenverhältnisse* zwischen zwei Hörerlebnissen sind bei der menschlichen Wahrnehmung von Musik ebenfalls wichtig und wurden demzufolge auch bei der Erstellung der Ontologie berücksichtigt. Von besonderer Wichtigkeit in einem Musikstück sind die *Linien*. Diese gliedern sich unter anderem weiter auf in eine

- Melodielinie, welche die Melodie eines Stückes darstellt,
- Begleitlinie, welche die Begleitung (Klavier-, Gitarrenbegleitung) enthält,
- Harmonielinie, durch welche die Harmonien (Akkorde) eines Stückes

formalisiert werden,

Anhand der Ontologie wurde dann ersichtlich, dass die relative Entropie als Strukturierungsmaß für Ontologien und der durch sie dargestellten Musikstücken dienen kann. Dazu wurde das Thema aus der Klaviersonate in A-Dur von Wolfgang Amadeus Mozart mit Hilfe der Ontologie formalisiert. Anschließend wurden zwei Variationen des Themas der Sonate sowie als Vergleichsstück der türkischen Tanz (Rondo Alla Turca) ausgewählt und formalisiert. Die Variationen sollten hierbei eine geringe Distanz zu dem Thema aufweisen, da sie, musikalisch gesehen, das Thema aufgreifen und in verschiedenen Aspekten variieren. Das Vergleichsstück hingegen sollte eine große Distanz zu dem Thema aufweisen, da es nur sehr wenige Gemeinsamkeiten mit dem ihm aufweist. Zur Durchführung der Strukturierung der Ontologien mittels der relativen Entropie wurden frequenzbasierte Wahrscheinlichkeitsverteilungen erstellt, die sich an den musikalischen Vergleichsmöglichkeiten für Musikstücke [Frisius 1984] orientieren.

1.5 Gliederung der Arbeit

Die Arbeit untergliedert sich im folgenden in insgesamt sieben Kapitel. Wie eingangs dargelegt, ist die derzeit schlechte Verständigung bei Flutkatastrophen von offizieller Seite [Kirchbach 2003] bestätigt und insbesondere auf die unzureichende Kommunikation von Wissen zurückzuführen. Ontologien stellen eine Möglichkeit dar, Wissen zu vereinheitlichen und werden daher in Abschnitt 2.1 näher erläutert. Dabei wird die geschichtliche sowie aktuelle Bedeutung von Ontologien herausgestellt, sowie deren Formalisierungsmöglichkeiten und heutige Definition in der Informatik vorgestellt. Um Wissen zu strukturieren gibt es verschiedene Möglichkeiten, welche in Abschnitt 2.2 aufgezeigt werden. Die vorgestellten Strukturierungsmöglichkeiten sind vor allem Distanzmaße aus der Computerlinguistik und dem Data Mining. Auch dient dieses Kapitel dazu, eventuelle Verwechslungsmöglichkeiten der in dieser Arbeit benutzten entropiebasierten Distanzmaße auszuschließen. Entropiebasierte Distanzmaße werden in der Computerlinguistik bzw. dem Data Mining ebenfalls eingesetzt, haben dort allerdings eine andere Bedeutung.

In Abschnitt 2.3 wird das Konzept der Entropie ausführlich vorgestellt, um die später benutzten, entropiebasierten Distanzmaße gebührend einzuführen. Da dieses Konzept in der Domäne der Wissensrepräsentation weitgehend unbekannt ist, wird der Entropiebegriff kurz über die Thermodynamik und statistische Mechanik eingeführt. Anschließend wird die Bedeutung in der Informationstheorie aufgezeigt und die theoretischen Grundlagen der Entropie erläutert. Nach der Erläuterung der Grundlagen der Entropie wird in Kapitel 3 erstmals ein theoretisches Modell für ein entropiebasiertes Distanzmaß auf Ontologien vorgestellt, welches eine Verfeinerung der gegensei-

tigen Information darstellt. In diesem Zusammenhang wird auch ein Strukturierungsmodell für Wissen mittels entropiebasierter Distanzmaße vorgestellt.

Um die Bedeutung der relativen Entropie für die Strukturierung des durch Ontologien darstellbaren Wissens aufzeigen zu können, wird in Kapitel 4 die Ontologie über Flutkatastrophen vorgestellt. In diesem Kapitel wird die Vorgehensweise bei der Erstellung der Ontologie besprochen, und es werden die wichtigsten Konzepte und Relationen vorgestellt. Ein Annotationstest, welcher die Benutzbarkeit der Flutontologie überprüfte, wird im Anschluss daran vorgestellt. In Kapitel 5 wird schließlich die Strukturierung von Wissen, welches durch Ontologien dargestellt werden kann, mittels der relativen Entropie gezeigt und anhand eines Beispieles demonstriert.

In Kapitel 6 wird eine Ontologie für die menschliche, musikalische Wahrnehmung vorgestellt. Anschließend wird gezeigt, wie durch die relative Entropie Ähnlichkeiten bzw. Unterschiede zwischen Musikstücke, welche durch die Ontologie formalisiert wurden, festgestellt werden können. Den Abschluss der Arbeit bildet die Zusammenfassung in Kapitel 7.

Kapitel 2

Stand der Technik

In diesem Kapitel soll ein Überblick über die in dieser Arbeit benutzten Technologien gegeben werden. Zunächst wird das Konzept der Ontologie als strukturiertes Wissen vorgestellt. Anschließend werden verschiedene Strukturierungsmaße für Wissen vorgestellt. Das Ende dieses Kapitels bilden die Entropie und die mit ihr verwandten Distanzmaße, die im späteren Verlauf auf Ontologien angewandt und erweitert werden.

2.1 Ontologien

In diesem Abschnitt wird zunächst die gebräuchlichste Definition von Ontologien in der Informatik nach [Gruber 1993b] vorgestellt. Danach werden verschiedene, formale Repräsentationsmöglichkeiten für Ontologien dargestellt: Logik (*Logic*), Klassen (*Frames*) und Semantische Netze (*Semantic Nets*). Zum Schluss werden noch kurz einige bekannte Ontologien erwähnt, sowie Kategorisierungsmöglichkeiten für dieselbigen vorgestellt.

2.1.1 Ontologien in der Informatik

In den letzten zehn Jahren gewannen Ontologien in der KI, insbesondere bei der Wissensverarbeitung (knowledge engineering) als auch der Wissensrepräsentation (knowledge representation) an Bedeutung. Es gibt jedoch eine Vielzahl von Definitionen für diesen Begriff. Die dieser Arbeit zugrundeliegende Definition stammt von Gruber [Gruber 1993b] aus dem Bereich der Wissensrepräsentation und soll im folgenden vorgestellt werden. Er definiert eine Ontologie als

> *[...] explicit specification of a conceptualization.*

Einige Jahre später verfeinerten [Studer u. a. 1998] diese Definition:

> *An ontology is a formal, explicit specification of a shared conceptualization.*

Eine Konzeptualisierung (*conceptualization*) beschreibt ein abstraktes Modell eines Phänomens in der Welt, wobei die relevanten Konzepte des Phänomens bereits identifiziert wurden. Die Bedeutung von *Formal* liegt darin, dass die Repräsentation der Ontologie maschinenlesbar sein soll (siehe Abschnitt 2.1.2). Explizit (*explicit*) bedeutet, dass die Konzepte, die Einschränkungen denen sie unterliegen und die Relationen zwischen ihnen ausdrücklich beschrieben und definiert werden. Gemeinsam (*shared*) soll verdeutlichen, dass die Konzepte bzw. das Wissen, welches durch die Ontologie repräsentiert wird, von einer hinreichend großen Gemeinde akzeptiert und als tragfähig für diesen Zwecke befunden worden ist. Aus dem Bereich der formalen Ontologien wird jedoch berechtigte Kritik an dieser Definition geübt, da *conceptualization* nicht hinreichend genau spezifiziert wird [Smith 2004].

Der Begriff der Ontologie ist inzwischen, auch aufgrund des semantischen Web (*Semantic Web*) [Berners-Lee u. a. 2001] bzw. kritischer [Uschold 2003], recht weit verbreitet wobei er manchmal als Synonym für Taxonomien verwendet wird. Unter einer Taxonomie versteht man eine hierarchische Gliederung eines Diskursbereiches, meistens anhand einer Vererbungsrelation (*is-a relation*), welche die Konzepte vom allgemeinen zum speziellen hin strukturiert [Nilsson 1998]. Um hier eine Unterscheidung treffen zu können, führte man *leichtgewichtige* (lightweight) bzw. *schwergewichtige* (heavyweight) Ontologien ein. Die leichtgewichtigen ähneln Taxonomien, strukturieren also das Wissen der betrachteten Domäne anhand relevanter Konzepte, meist in Form eines hierarchischen Baumes. Beispiele sind der Yahoo-Katalog [1] oder der Produktkatalog von Google [2]. Schwergewichtige Ontologien entsprechen hingegen Ontologien wie sie bereits zuvor definiert wurden. Sie enthalten mehr Semantik als leichtgewichtige Ontologien, die durch unterschiedliche Relationen und Kardinalitäten zwischen den Konzepten modelliert wird. In einer Taxonomie findet man üblicherweise nur eine Vererbungsrelation (*is-a relation*) vor.

Zusammenfassend lässt sich sagen, dass eine Ontologie aus Konzepten sowie Relationen zwischen diesen Konzepten besteht. Es kristallisiert sich eine Struktur heraus, welche die Interpretationsmöglichkeiten der Konzepte einschränkt. Damit kann eine Ontologie als strukturiertes Wissen betrachtet werden, welches nach Möglichkeit in der untersuchten Domäne allgemein anerkannt ist und formal repräsentiert werden kann. Die Möglichkeiten, eine Ontologie formal zu repräsentieren, werden im nächsten Abschnitt behandelt.

[1] http://www.yahoo.de
[2] http://froogle.google.de

2.1.2 Formale Repräsentation

Ontologien werden heutzutage in der Künstlichen Intelligenz mit einer der drei bekannten Wissenrepräsentationsparadigmen dargestellt. Dies sind Logik, meistens Beschreibungslogik [Baader u. a. 2003], Klassen [Gruber 1993a] sowie semantische Netze [Sowa 2000]. Nähere Informationen zu diesen Paradigmen finden sich beispielsweise in [Russel u. Norvig 2003] oder [Nilsson 1998].

Zur Formalisierung dieser Techniken stehen höchst unterschiedliche Sprachen zur Verfügung, von denen die wichtigsten im folgenden kurz vorgestellt werden. Weitere Formalisierungsmöglichkeiten finden sich in [Uschold u. Grüninger 1996] und [Gómez-Pérez u. a. 2004].

Modellierung mit Klassen

In der Wissensrepräsentation und -verarbeitung sind insbesondere das *Knowledge Interchange Format* (KIF) von [Genesereth u. Fikes 1992] und damit eng verbunden der *Ontolingua* (Gruber 1996 und Farquhar u. a. 1997) hervorzuheben. KIF wurde mit dem Ziel entwickelt, die Heterogenität der verschiedenen, in der Wissensrepräsentation verwendeten Sprachen, zu beenden. Dazu sollte KIF einen einheitlichen Standard für den Austausch von Wissen zwischen verschiedenen Informationssystemen definieren. KIF ist eine Präfixnotation der Prädikatenlogik erster Ordnung mit einigen Erweiterungen und somit sehr aussagekräftig. Auf KIF aufbauend wurden verschiedene Zwischenstufen definiert, unter anderem die klassenbasierte *Open Knowledge Base Connectivity* (OKBC) [Chaudhri u. a. 1998], da es mit KIF sehr mühsam ist, eine Ontologie zu erstellen.

Mit dieser etwas einfacheren, allerdings auch weniger ausdrucksstarken Sprache, wurde *Ontolingua* entwickelt. In Ontolingua wurden inzwischen sehr viele Ontologien erstellt, welche auf dem Ontolingua Server[3] gespeichert sind. Weiterhin erlaubt OKBC die Kommunikation mit mehreren Ontologien (siehe Abb. 2.1), darunter auch die bekannte CyC Ontologie von [Lenat u. Guha 1990], da es ein Protokoll zum Wissensaustausch beinhaltet. Ein unmittelbarer Vorteil aus der Benutzung von logikbasierten Sprachen besteht darin, dass Inferenzmechanismen zur Generierung bzw. Validierung neuen Wissen direkt implementiert werden können.

2.1.3 Modellierung mit Beschreibungslogik

Im folgenden sollen die wichtigsten XML[4] [Harold 2002] basierten Sprachen vorgestellt werden, da sie im Kontext des *Semantic Web*[5] [Berners-Lee u. a. 2001] und der Rahmenforschungsprogramme der Europäischen Union

[3] http://ontolingua.stanford.edu
[4] http://www.w3.org/XML
[5] http://www.semanticweb.org

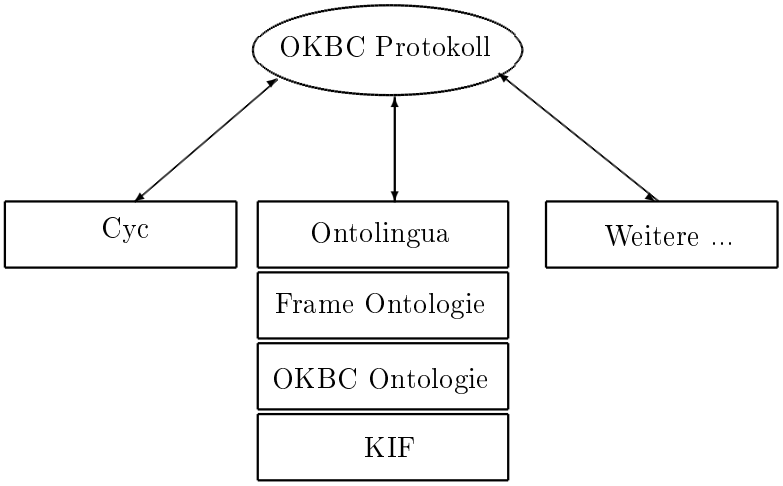


Abbildung 2.1: OKBC Protokoll

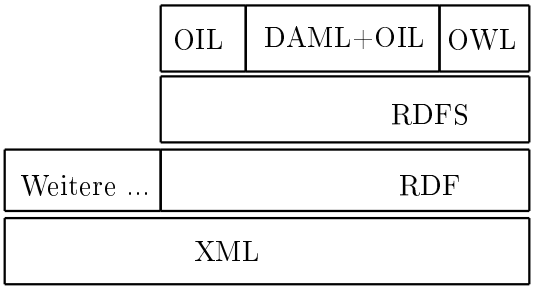


Abbildung 2.2: Aufbau der auszeichnungsbasierten Ontologiesprachen

(EU)[6] besondere Aufmerksamkeit erfahren haben. Diese Sprachen werden auch netzbasierte Ontologiesprachen (web-based ontology languages) oder auszeichnungsbasierte Ontologiesprachen (ontology markup languages) genannt.

In Abb. 2.2 ist der hierarchische Aufbau dieser Sprachen dargestellt. XML, eine Metasprache welche seit Februar 1998 eine Empfehlung des W3C ist, dient als Basis um strukturierte, standardisierte und wohldefinierte Dokumente zu erstellen. Die meisten der auszeichnungsbasierten Ontologiesprachen benutzten als Basis RDF (Resource Description Framework) bzw. RDF(S) (RDF Schema), welches in Abschnitt 2.1.4 vorgestellt wird, weil das zugrundeliegende Modell ein semantisches Netz ist.

Von den vielen, vorhandenen webbasierten Ontologiesprachen sollen lediglich DAML+OIL [van Harmelen u. Patel-Schneider 2001] und OWL [McGuinness u. van Harmelen 2004] besprochen werden, da sie in dieser Arbeit für die konkrete Implementierung von Ontologien benutzt wurden und sich

[6] http://fp6.cordis.lu/fp6/home.cfm

sehr ähnlich sind; OWL löste im Frühjahr 2004 DAML+OIL als offizielle, vom W3C unterstützte, webbasierte Ontologiesprache ab. Die Flutontologie (siehe Kapitel 4) wurde noch in DAML+OIL entwickelt, da zum damaligen Zeitpunkt OWL noch nicht vollständig spezifiziert war und auch keine Hilfsmittel zur Ontologieerstellung in dieser Sprache verfügbar waren. Hingegen wurde die Ontologie für die menschliche, musikalische Wahrnehmung (siehe Kapitel 5) in OWL implementiert.

Der Grund für die Verwendung von DAML+OIL bzw. OWL war eine breite Verfügbarkeit von Hilfsmitteln, welche die Implementierung der Ontologien erleichtern. Auch sind diese Sprachen als Standards des W3C für das Semantic Web definiert. Für zukünftige Arbeiten existieren weiterhin eine Reihe von Beweisern[7] [8] [9], die verschiedene Inferenzoperationen auf in DAML+OIL oder OWL definierten Ontologien erlauben.

DAML+OIL

DAML+OIL entstand aus einer Zusammenarbeit zwischen der Europäischen Union im Rahmen des IST Projektes OIL (Ontology Inference Layer) und den USA im Kontext des DARPA Projektes DAML (DARPA Agent Markup Language). Mit dieser Sprache wurde die Flutontologie modelliert. DAML+OIL hält folgende *Primitive* bereit, um Ontologien zu modellieren:

- Primitive zur Modellierung Klassen, Restriktionen und verschiedenen, grundlegenden Datentypen. Restriktionen beziehen sich auf Klassen, z.B. existentielle Restriktionen, Einschränkungen der Kardinalität u.s.w.
- Primitive um Attribute und Relationen zu definieren. Die Relationen können als zusätzliche Eigenschaft transitiv oder injektiv sein.
- Primitive um Container zu definieren.
- Vordefinierte Klassen. *Ding* als generellste Klasse und *Nichts* als speziellste.
- Primitive um Literale zu definieren.

Um semantisch reichhaltige Aussagen in DAML+OIL zu erstellen, existieren noch folgende Operationen:

- Konjunktion, Disjunktion und Negation von Klassen
- Sammlung von Individuen (Existenzquantor)
- Möglichkeiten um Attribute der Klassen einzuschränken (z.B. Kardinalität, Einschränkungen des Wertebereichs, etc.)

[7] http://www.sts.tu-harburg.de/~r.f.moeller/racer
[8] http://www.mindswap.org/2003/pellet/index.shtml
[9] http://jena.sourceforge.net

- Zu einer gegeben Relation kann eine inverse Relation definiert werden.
- Äquivalenzen zwischen Klassen, Relationen und Eigenschaften von Klassen können definieren werden.
- Zwei Instanzen können als voneinander verschieden deklariert werden.
- Möglichkeit, diverse Metainformationen darzustellen.

Funktionen sowie formale Axiome sind nicht Teil der Spezifikation von DAML+OIL. Binäre Funktionen können mittels eines Hilfskonstruktes erzeugt werden, höherwertige Funktionen sind nicht darstellbar. Es existieren einige Inferenzmechanismen (z.B. eine Axiomatisierung in KIF) sowie Implementierungen von Beweisern (beispielsweise FaCT[10] in Java).

Als Hilfsmittel zur Erstellung einer Ontologie in DAML+OIL wurde das vom AIFB (Institut für Angewandte Informatik und Formale Beschreibungsverfahren) der Universität Karlsruhe (TH) entwickelte Ontologiemodellierungswerkzeug *OntoEdit* [11] benutzt [Sure u. a. 2002]. Die Visualisierung der Ontologie erfolgte mit den Werkzeugen IsaViz[12] des W3C und GraphViz[13], entwickelt von AT&T Research.

OWL

OWL wurde von der W3C Web Ontology (WebOnt) Arbeitsgruppe geschaffen und im Februar 2004 zu einem offiziellen Standard erhoben. OWL ist von DAML+OIL abgeleitet und baut ebenfalls auf RDF(S) auf (siehe Abb. 2.2). Viele der in DAML+OIL vorhandene Primitive wurden schlicht umbenannt. OWL wurde in drei Sprachspezifikationen unterteilt: OWL Lite, OWL DL und OWL Full (siehe Abb. 2.3).

OWL Lite ist eine Erweiterung von RDF(S) und beinhaltet die nützlichsten Eigenschaften von OWL. Sie ist für Benutzer gedacht, die lediglich Taxonomien mit wenig bis keiner Semantik erstellen wollen. OWL DL (DescriptionLogic) beinhaltet das volle Vokabular von OWL und wird hier vorgestellt, da es zur Erstellung der Ontologie für die menschliche, musikalische Wahrnehmung (siehe Kapitel 5) benutzt wurde. OWL Full bietet mehr Flexibilität als OWL DL, allerdings ist die Komplexität, insbesondere hinsichtlich der Inferenzmechanismen, höher. Der Aufbau und Inhalt der meisten Primitive ist denen von DAML+OIL ähnlich, so dass im folgenden lediglich die Unterschiede und Ergänzungen herausgestellt werden. Aufgrund dieser Ähnlichkeiten ist eine automatische Konversion von DAML+OIL Ontologien nach OWL ohne größere Probleme möglich:

[10] http://www.cs.man.ac.uk/ horrocks/FaCT/
[11] http://www.ontoprise.de/products/ontoedit
[12] http://www.w3.org/2001/11/IsaViz/
[13] http://www.graphviz.org

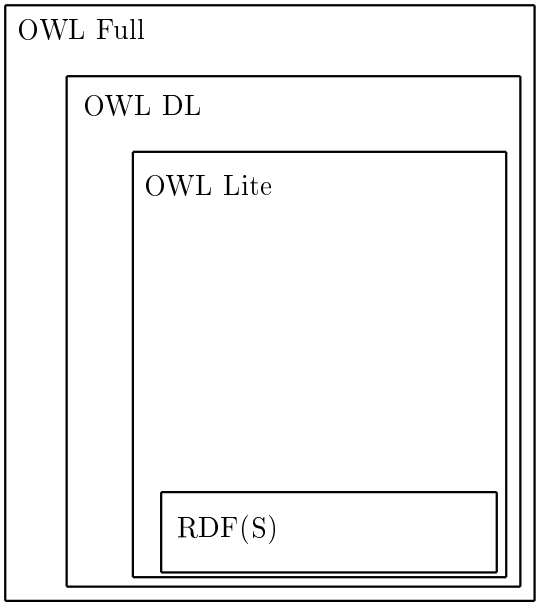


Abbildung 2.3: Aufteilung von OWL in drei Schichten

- Es wurden Primitive hinzugefügt, um symmetrische Relationen zu definieren
- Es ist möglich, Metainformationen zu Klassen, Eigenschaften und Instanzen als auch der Ontologie hinzuzufügen.
- Es existieren Primitive, um Enumerationen von Datentypen zu erzeugen.
- Unterstützung für die Versionierung von Ontologien wurde hinzugefügt.
- Primitive um anzuzeigen, dass einige Instanzen sich von anderen unterscheiden.
- Primitive um disjunktes Wissen in der Ontologie zu definieren.

Implementierung von Ontologien in OWL-DL

Die Ontologie für die menschliche, musikalische Wahrnehmung (siehe Kapitel 6) wurde in OWL-DL implementiert. Dabei wurden bestimmte Sprachkonstrukte sowie ein Designmuster verwendet, welche im folgenden vorgestellt werden.

Bei der Implementierung von Ontologien in OWL überlappen sich normalerweise einzelne Klassen der Ontologie. Demzufolge kann man nicht annehmen, dass eine konkrete Instanz einer Klasse nicht noch Mitglied weiterer

Klassen ist, nur weil die anderen Klassen nicht explizit aufgeführt sind. Zur Trennung der einzelnen Klassen bzw. ihrer Instanzen werden sie in OWL als disjunkt (*disjoint*) zueinander gesetzt. Damit wird sichergestellt, dass eine explizit definierte Instanz einer bestimmten Klasse nicht zu anderen Klassen, welche als disjunkt dazu definiert wurden, gehören kann. Wenn beispielsweise die Klassen *Zeitmaß* und *Puls* disjunkt zueinander sind, kann eine Instanz der Klasse *Zeitmaß* nicht gleichzeitig Instanz der Klasse *Puls* sein.

Für das im nächsten Abschnitt definierte Designmuster wird ein so genanntes Überdeckungsaxiom (*covering axiom*) benötigt. Ein Überdeckungsaxiom besteht aus zwei Teilen: Der Klasse, die überdeckt wird und denjenigen Klassen, welche die Überdeckung bilden. Nehmen wir an, wir hätten drei Klassen, A, B und C definiert, wobei die Klassen B und C Unterklassen von A bilden. Weiterhin wurde ein Überdeckungsaxiom für die Klasse A definiert, welches besagt, dass A von den Klassen B und C überdeckt wird. Damit muss ein Mitglied der Klasse A ein Mitglied der Klasse B und/oder C sein. Falls B und C disjunkt zueinander sind, müssen die Mitglieder von Klasse A entweder Mitglieder von Klasse B oder C sein. In OWL manifestiert sich das Überdeckungsaxiom als die Vereinigung der Klassen, welche die Überdeckung bilden (B und C). Diese formen dann eine Superklasse für A:

$$\cup \, B \, \cup \, C$$

Das Designmuster *Value Partition* wird verwendet, wenn bestimmte Klassen für eine ihrer Eigenschaften ausschließlich bestimmte Werte annehmen können. Beispielsweise könnte man die Klasse *Mahlzeiten* in eine der folgenden drei Kategorien einorden: scharf, mittelscharf oder mild. Dazu wird in OWL zunächst eine Klasse *Value Partition* erstellt, deren Unterklassen die möglichen Eigenschaften sind. In unserem Beispielfall wäre dies etwa eine Klasse mit dem Namen *SchärfeValuePartition*. Diese enthält als Unterklassen dann die möglichen Eigenschaften, also *Scharf*, *Mittelscharf* und *Mild*. Diese sind untereinander disjunkt und werden von *SchärfeValuePartition* mittels des Überdeckungsaxioms überdeckt. Anschließend wird eine Relation erstellt, welche die Klassen mit der entsprechenden Value Partition verknüpfen. In unserem Beispiel habe die Relation den Namen *hatSchärfeGrad*. Die Domäne dieser Relation ist nun die Klasse *Mahlzeit*, das Bild die *SchärfeValuePartition*. Unterklassen von Mahlzeit werden jetzt über die Relation *hatSchärfeGrad* und einer Value-Restriction mit den gewünschten Schärfegraden verknüpft.

Ein wichtiger Bestandteil von OWL sind Einschränkungen, welche auf den Eigenschaften einer Klasse definiert werden können (Die Klasse *A* beinhalte für die folgenden Beispiele entsprechende Restriktionen auf ihren Eigenschaften). Im wesentlichen gibt es drei Restriktionen:

- Quantifikatiorrestriktionen

- Kardinalitätsrestriktionen
- *hasValue* Restriktionen

Existenzrestriktionen, in OWL Terminologie „someValuesFrom“, beschreiben diejenige Menge von Individuen, die mindestens eine spezifische Relation zu Individuen aufweisen, die Mitglieder einer bestimmten Klasse sind:

$$\exists \text{ relationName } \textit{Klasse } B$$

Diese Restriktion beschreibt anschaulich diejenigen Instanzen von A, die mindestens eine Relation namens *relationName* zu anderen Instanzen besitzen, welche zur Klasse B gehören.

Die Menge der Instanzen, welche für eine gegebene Relation ausschließlich Beziehungen zu anderen Individuen besitzen die Mitglied einer bestimmten Klasse sind, werden durch die Universalrestriktion, in OWL Terminologie „allValuesFrom“, beschrieben:

$$\forall \text{ relationName } \textit{Klasse } B$$

Zu dieser Menge gehören alle Instanzen von A, welche ausschließlich Beziehungen mit dem Namen *relationName* zu Instanzen der Klasse B aufweisen, oder Instanzen, die keinerlei Beziehung zu der Relation *relationName* aufweisen. Universalrestriktionen sagen in OWL also nichts über die Existenz der betroffenen Relation aus [Horridge u. a. 2004].

Die *hasValue* Restriktion beschreibt diejenigen Instanzen, welche durch eine bestimmte Eigenschaft in Relation mit einer spezifischen Instanz einer Klasse stehen:

$$\textit{prop} \ni \textit{InstanzABC}$$

Dadurch werden diejenigen Instanzen von A beschrieben, welche anhand der Eigenschaft *prop* eine Beziehung zu der konkreten Instanz *InstanzABC* aufweisen.

Mittels Kardinalitätsrestriktionen kann man solche Instanzen der Klasse A beschreiben, die *mindestens*, *höchstens oder* exakt eine bestimmte Anzahl n an Beziehungen zu anderen Instanzen aufweisen:

$$\geq \text{relationName } \geq n$$

$$\leq \text{relationName } \leq n$$

$$= \text{relationName } = n$$

Entwicklungswerkzeug Protégé

Für die Entwicklung der Ontologien in OWL wurde Protégé[14] in der Version 3.0 eingesetzt [Gennari u. a. 2002]. Protégé wurde in Java implementiert

[14] http://protege.stanford.edu

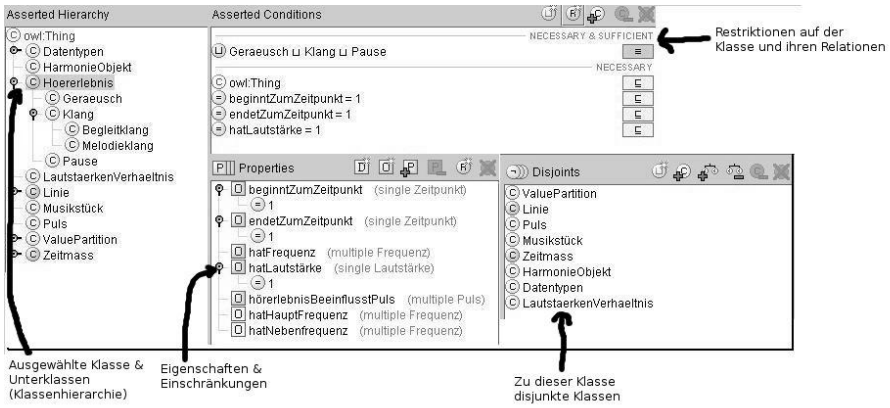


Abbildung 2.4: Protégédarstellung einer Klasse, ihrer Relationen und Einschränkungen.

und unter einer Open Source Lizenz zur Verfügung gestellt. Ein Plugin[15] erleichtert die Entwicklung von Ontologien in OWL. Die graphische Visualisierung in Form eines hierarchischen Baumes übernimmt ebenfalls ein Plugin namens GraphViz[16] von der Universität Manchester. Um die Konsistenz der Ontologie zu prüfen wurde RACER[17] eingesetzt. Racer ist ein Reasoner für Beschreibungslogik und kann mittels eines Plugins ebenfalls direkt in Protégé verwendet werden.

In Abb. 2.4 ist die in Kapitel 4 und 5 verwendete Darstellung der Klassen der jeweiligen Ontologien abgebildet. Auf der linken Seite ist in einer Baumdarstellung die ausgewählte Klasse farbig unterlegt und weiterhin werden ihre Subklassen dargestellt. Der Bereich rechts daneben (Asserted Conditions) finden sich übersichtlich die Einschränkungen auf den Relationen oder Eigenschaften der Klasse. Auch finden sich hier die direkt auf der Klasse definierten Einschränkungen, wie beispielsweise ein Covering Axiom. Darunter befindet sich die Darstellung der Relationen (Properties). Relationen zu anderen Konzepten werden durch ein „O“ (Objectproperty) gekennzeichnet, Relationen zu einfachen Datentypen wie Strings oder Zahlen durch ein „D“ (Datatypeproperty). Hinter dem Namen der Relation steht der dazugehörige Bildbereich. Die Notation der Einschränkungen folgt der weiter oben beschriebenen Vorgehensweise. Die „Disjoints“ stellen schließlich die zur ausgewählten Klasse disjunkten Konzepte dar.

[15] http://protege.stanford.edu/plugins/owl/index.html
[16] http://www.co-ode.org/downloads/owlviz/co-ode-index.php
[17] http://www.fh-wedel.de/~mo/racer

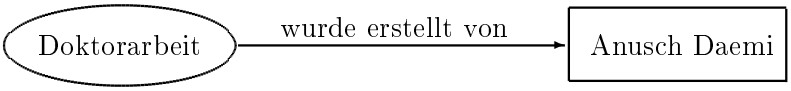


Abbildung 2.5: Aufbau einer RDF Aussage

2.1.4 Sprachen für Semantische Netze

Den Regeln von XML folgend wurde RDF (Resource Description Framework) bzw. RDF(S) (RDF Schema) 1999 vom W3C spezifiziert [Lassila u. Swick 1999]. Das in RDF benutzte Datenmodell entspricht dem eines semantischen Netzes und besteht primär aus drei Objekttypen: *Ressourcen*, *Eigenschaften* und *Aussagen*.

Ressourcen beschreiben jegliche Daten, welche durch einen *Uniform Ressource Identifier* (URI) identifiziert werden können. Damit können nicht nur Webseiten, sondern beispielsweise auch Bücher, Waren und ähnliches beschrieben werden.

Die Beschreibung einer Ressource erfolgt durch Eigenschaften. Eine Eigenschaft beschreibt somit ein Attribut oder Beziehungen zwischen Ressourcen. Eigenschaften werden durch einen eindeutigen Namen identifiziert und ein dazugehöriges Schema beschreibt den Wert oder Wertebereich, den die Eigenschaft annehmen kann.

Eine RDF-Aussage kombiniert eine Ressource mit einer Eigenschaft und ihrem Wert. Anders ausgedrückt, ist eine RDF-Aussage ein einfacher Satz mit Subjekt, Prädikat und Objekt. Die zu beschreibende Ressource ist das Subjekt. Die Eigenschaft, um die Ressource zu beschreiben, ist das Prädikat und der Wert der Eigenschaft ist das Objekt der Aussage. Natürlich kann das Objekt einer RDF Aussage auch eine andere RDF Aussage sein. *RDF Schema* (RDFS) [Brickley u. Guha 2004] beschreibt die Semantik und die zulässige Syntax der in RDF vorkommenden Sprachelemente. Beispielsweise zeigt Abb. 2.5 den Satz

Diese Doktorarbeit wurde von Anusch Daemi erstellt.

in Form eines semantischen Netzes. Die Ressource ist die *Doktorarbeit*, die Eigenschaft oder Relation ist *erstellen* und der Wert der Eigenschaft ist *Anusch Daemi*.

Eine Transformation Semantischer Netze in eine klassenbasierte Darstellung und umgekehrt ist einfach zu bewerkstelligen und auch sinnvoll, denn beide haben ihre Vorteile haben, je nach Anwendungszweck. Weitere Formalisierungen semantischer Netze sind so genannte Topic Maps [Pepper u. Moore 2001; Rath 2002].

2.1.5 Kategorisierung von Ontologien

Top-level Ontologien (top level oder upper ontologies) beschreiben sehr allgemeine und meistens abstrakte Konzepte, welche über alle Domänen hinweg gültig sind, beispielsweise Raum, Zeit und weitere abstrakte Konzepte. Diese werden meist dazu benutzt, um Domänenontologien mit allgemeinen Konzepten zu bereichern und eine gemeinsame Basis zu anderen Domänenontologien zu schaffen. Der Aufbau solcher Top-level Ontologien ist, aufgrund unterschiedlicher philosophischer Ansichten, nicht unstrittig. Es sind inzwischen mehrere erstellt worden, unter anderem die CyC Ontologie von [Lenat u. Guha 1990], die Kategorisierung nach [Sowa 2000] und die Standard Upper Merged Ontology (SUMO[18]) von [Niles u. Pease 2003], die im Rahmen der IEEE Standard Upper Ontology Arbeitsgruppe gefördert wurde. Mittels solcher Top-level Ontologien können beispielsweise die allgemeinen Konzepte (Zeit, Raum, etc.) in der Flutontologie modelliert werden.

Auch gibt es linguistische Ontologien, welche semantische Konstrukte in Sprachen beschreiben. Sie sind meistens aus den grammatikalischen Einheiten einer Sprache, z.B. Substantive, Adjektive und Verben aufgebaut. Eine der bekannteren linguistische Ontologie ist WordNet, erstellt von [Miller u. a. 1993]. Sie konzentriert sich auf die Deutung von Wörtern, während andere die Abbildungen zwischen Konzepten verschiedener Sprachen betrachten (EuroWordNet[19] von Vossen 1998). Diese Ontologien sind natürlich von der verwendeten Sprache abhängig (z.B. WordNet für Englisch), was für Domänenontologien nur in gewissem Maße gilt. Manche Konzepte haben, je nach Übersetzung, unterschiedliche Bedeutung, oder stehen in der anderen Sprache für allgemeinere oder speziellere Dinge.

Die zahlenmäßig größte Kategorie der Ontologien sind die Domänenontologien von denen zwei im Rahmen dieser Arbeit entwickelt wurden und in Kapitel 4 und 6 näher vorgestellt werden. Durch den Boom des e-Commerce [Fensel 2000] existieren noch unzählige weitere Domänenontologien für Produktkategorisierung, Handel zwischen Geschäftspartnern etc.

2.2 Strukturierungmaße

Im vorigen Abschnitt wurden Ontologien als strukturiertes Wissen vorgestellt. In diesem Abschnitt sollen verschiedene Möglichkeiten dargelegt werden, wie unstrukturiertes Wissen strukturiert werden kann. Es sollen zunächst verschiedene Strukturierungsmöglichkeiten, insbesondere aus dem Information Retrieval (IR) und dem Data Mining vorgestellt werden. Da für die meisten der aufgezeigten Verfahren Distanzmaße benutzt werden, sollen sie ebenfalls näher erläutert werden.

[18] http://ontology.teknowledge.com
[19] http://www.hum.uva.nl/ewn

2.2.1 Distanz

Die Idee, Distanzmaße zur Strukturierung beliebiger Informationen zu benutzen, wurde bereits im antiken Griechenland von Euklid vorgebracht. Eine metrische Distanz zwischen zwei Punkten a und b ist durch drei Axiome definiert:

1. Positiv definit: $\forall\ a, b:\ d(a,b) \geq 0$ und $d(a,b) = 0$ genau dann, wenn $a = b$ gilt.
2. Symmetrie: $d(a,b) = d(b,a)$
3. Dreiecksungleichung: $\forall\ a, b, c:\ d(a,b) \leq d(a,c) + d(c,b)$

Das bekannteste Distanzmaß dürfte die (geometrische) euklidische Distanz sein, welche die kürzeste Strecke zwischen zwei Merkmalen bestimmt:

$$d_{jk} = \sqrt{\sum_{i=1}^{n} |x_{ji} - x_{ki}|^2}$$

d_{jk} ist definiert als die Summe der quadratischen Differenzen von Merkmalen x in einem kontinuierlichen Raum (z.B. $\mathbb{R}^2$) wobei j und k die zu differenzierenden Merkmale sind. Dies sind im Fall $i = 1$ Punkte, ansonsten Vektoren der Länge n.

Falls das Distanzmaß in einem mehrdimensionalem Raum benutzt werden soll, in denen die Dimensionen unterschiedliche Semantiken vorweisen (z.B. Höhe und Gewicht), müssen die unterschiedlichen Skalen noch normiert werden. Ein Beispiel hierfür ist die Mahalanobisdistanz nach [Mahalanobis 1936]. Dieses Maß berechnet die Standardabweichung und Kovarianz jedes Merkmales im Merkmalsraum und drückt die Distanz zwischen den Merkmalen als Produkt zwischen den Standardabweichungen und Kovarianzen aus [Kotz u. Johnson 1981].

Falls man diskrete Merkmale mit einem Distanzmaß versehen möchte, bietet sich als einfachstes Maß das so genannte Hamminggewicht (bzw. -distanz) an. Sie ist definiert als die Gesamtzahl der verschiedenen Zeichen zwischen den Merkmalen. Diese Distanz wird vor allem in der Kodierungstheorie verwendet, da man mit ihr Codewörter, die über einen verrauschten Kanal verschickt wurden, in gewissen Maßen korrigieren kann [Pierce 1980].

2.2.2 Vektorraummodelle

Um die Distanz zwischen diskreten Merkmalen wie Wörtern, Sätzen, Dokumenten und ähnlichem zu berechnen, wurden in der Computerlinguistik (computational linguistic) verschiedene Verfahren eingeführt. Dies sind zum einen Maße, welche die Häufigkeiten, also die Frequenz, der zu messenden Merkmale betrachten (frequency approach), und zum anderen die Position

der Merkmale, nachdem sie in einen Vektorraum überführt wurden (positioning approach).

Der Vektorraumansatz benutzt so genannte Vektorraummodelle (VSM - Vector Space Models) um Wörter, Sätze und Dokumente in einen hochdimensionalen Vektorraum zu überführen. Die Basisvektoren dieser Vektorräume bestehen aus Indextermen (index terms). Diese Indexterme sind Wörter, welche für das zu messende Problem relevant sind. Beispielsweise können dies Wörter sein, die ein Dokument charakterisieren. Die Position des Vektors im Vektorraum wird durch Gewichte bestimmt, welche die Indexterme zugewiesen bekommen. Die Bestimmung dieser Gewichte erfolgt normalerweise durch Berechnung der Termfrequenz (term frequency - *tf*) [Li u. a. 2003], d.h. die relative Anzahl der Indexterme im Dokument:

$$tf = \frac{\text{Häufigkeit des Indexterms}}{\text{Gesamtzahl der Wörter}}$$

Falls man noch die Anzahl der betrachteten Dokumente bei der Berechnung miteinfließen lässt, ergibt sich daraus die invertierte Dokumentfrequenz (inverted document frequency - *idf*), näher vorgestellt in [Witten u. a. 1999].

Da die Position eines Dokumentes, repräsentiert durch einen Vektor mit entsprechenden Indextermen und deren Gewichtung, in dem betrachteten Vektorraum damit festgelegt ist, können durch Berechnung des Kosinus zwischen den Vektoren die Dokumente miteinander vergleichen werden. Die Vektoren müssen vorher noch normalisiert werden, daher wird der Koeffizient zwischen den Vektoren auch *normalisierter Korrelationskoeffizient* genannt wird.

Ein entscheidender Schritt für die Güte der Ähnlichkeitsbestimmung ist natürlich die Auswahl der Indexterme, die sehr sorgfältig geschehen muss, damit man sinnvolle Resultate erhält. Wenn man zum Beispiel sehr häufig vorkommende Wörter, wie Artikel, Pronomen oder Zahlen als Indexterme wählt, werden sich die normalisierten Häufigkeiten kaum voneinander unterscheiden und das Ähnlichkeitsmaß verliert seine Aussagekraft. Varianten und Optimierungen des Vektorraummodells werden insbesondere von den verschiedenen Suchmaschinen im Internet eingesetzt um korrekte Webdokumente auf die an sie gestellten Anfragen zu liefern.

2.2.3 Frequenzbasierte Distanzmaße

In frequenzbasierten Ansätzen werden die Häufigkeiten von Wörtern in einem Dokument, bzw. in mehreren Dokumenten, als Klassifizierungskriterium verwendet. Damit die Trennschärfe der Klassifizierung erhalten bleibt, müssen vor der Klassifikation unter anderem so genannte Stopwörter (stop words), herausgefiltert werden. Stopwörter sind diejenigen Wörter, die sehr häufig vorkommen (Artikel, Pronomen), aber wenig Informationsgehalt besitzen. Die Bedeutung des Informationsgehaltes ergibt sich nach Shannon, d.h. vereinfacht gesagt, besitzen äußert selten oder häufig vorkommende Wörter in

der Regel weniger Aussagekraft über das Dokument, als solche, die regelmäßig erscheinen.

Anschließend wird die Häufigkeit eines Wortes j in Dokumenten d_i, mit $i = 1 \ldots n$ berechnet, was die jeweiligen Termfrequenzen tf ergibt. Um eine genauere Klassifizierung zu ermöglichen, wurde dieses Maß um eine Gewichtungsfunktion erweitert. Diese Gewichtungsfunktion verringert tf, falls ein Wort zu oft oder selten in den Dokumenten vorkommt. Dieses Maß ist das im Information Retrieval häufig genannte $tfidf$ (term frequency/inverted document frequency):

$$tfidf(i,j) = tf(i,j) \cdot \log\left(\frac{n}{df(j)}\right)$$

$df(j)$ bezeichnet die Dokumentfrequenz und gibt an, in wie vielen Dokumenten der Term j vorkommt. Dieses Maß dient als Grundlage für viele andere Informationsextraktionsmethoden, welche versuchen aus Dokumenten jeglicher Art relevante Terme zu extrahieren und diese zu gruppieren und kategorisieren. Beispiele hierfür finden sich in [Hotho u. a. 2002], ein Vergleich mit anderen Maßen wird in [Orăsan u. a. 2004] vorgestellt und eine Kombination solcher Maße findet sich in [Li u. a. 2003].

Ein weiteres grundlegendes Distanzmaß, welches auf der Entropie (siehe Kapitel 2.3) basiert, ist das von [Resnik 1995] vorgeschlagene, das ebenfalls zu den frequenzbasierten Distanzmaßen gehört. Diese Methode berechnet die Ähnlichkeit zweier Begriffe in einer Taxonomie anhand deren Informationsgehaltes. Dazu wird den, in der Taxonomie vorkommenden, Konzepten c eine Auftreffenswahrscheinlichkeit $p(c)$ zugewiesen, die sich zum Beispiel aus dem englischsprachigen Korpus von [Fancis u. Kucera 1982] ergibt. Der Informationsgehalt eines solchen Konzeptes wird durch den negativen Logarithmus charakterisiert: $-\log p(c)$. Die Idee ist nun, dass sich zwei Konzepte um so ähnlicher sind, je mehr Informationsgehalt sie sich teilen. Der gemeinsame Informationsgehalt wird durch den Informationsgehalt desjenigen Konzeptes angezeigt, welches in der Vererbungshierarchie einer Taxonomie auf der niedrigst möglichen Ebene das gemeinsame Elternkonzept beider Konzepte ist.

Formal gesehen ist die Distanz zwischen Konzepten c_1 und c_2 einer Taxonomie definiert als:

$$d(c_1, c_2) = \max_{c \in S(c_1, c_2)} (-\log p(c))$$

$S(c_1, c_2)$ ist diejenige Menge von Konzepten, die in der Vererbungshierarchie die gemeinsamen Eltern von c_1 und c_2 sind.

2.2.4 Data-Mining

Data-Mining beschäftigt sich mit der Frage, wie extrem große Datenbestände strukturiert werden können. Durch die Strukturierung sollen die Daten-

bestände in handlichere Größen gebracht werden, damit Wissen aus ihnen gewonnen werden kann. Eine zentrale Forderung an das Data-Mining ist dabei, dass bei der Strukturierung der großen Datenbestände soviel inheräntes Wissen der Daten wie möglich erhalten bleibt.

Um große Datenbestände zu kleineren, sinnvollen Datensätzen zusammenzufassen, so dass möglichst viel Information gewonnen werden kann, gibt es eine ganze Reihe statistischer Methoden [Thong u. Liu 2004]. Um Data-Mining effektiv zu betreiben, reicht es nach [Billard u. Diday 2003] nicht aus, nur statistische Methoden auf die Datenmengen anzuwenden, sondern man muss sie vielmehr in symbolische Daten umformen. Damit werden nicht mehr einzelne Werte betrachtet, sondern Intervalle, Listen oder Kategorien von Werten. Diese symbolische Datenmengen können nun durch eine Vielzahl von unterschiedlichen Distanzmaßen strukturiert werden. Klassische Metriken, wie z.B. die Minkowskimetrik [Weisstein 2005] finden ebenso Anwendung wie Distanzmaße für Wahrscheinlichkeitsverteilungen, beispielsweise die Chernoff Distanz [Chernoff 1952].

Desweiteren werden viele Methoden aus der deskriptiven Statistik angewandt. Sie dienen sie dazu, Bündel (Cluster) in den Datenmengen zu finden, in denen ähnliche Objekte zusammengefasst sind (vgl. Celeux u. a. 1989 und Aggarwal u. a. 1999). Um die Distanz zwischen solchen Clustern zu berechnen, werden ebenfalls Methoden aus der Statistik wie die Kullback-Leibler Distanz [Kullback u. Leibler 1951] angewandt.

2.3 Entropie

In den nächsten Abschnitten soll ein Überblick über den Begriff der Entropie gegeben werden. Dazu wird der Begriff zunächst kurz historisch eingeführt und anschließend näher auf die Bedeutung und Interpretation in der Mathematik und Informatik eingegangen.

2.3.1 Geschichte der Entropie

Die Entropie hat ihren Ursprung in der Thermodynamik, insbesondere in dem zweiten Hauptsatz der Thermodynamik. Wichtige Arbeiten hierfür lieferten Carnot, Clausius und Kelvin [Uffink 2001]. Clausius beobachtete, dass in einem offenen, umkehrbaren Prozess der Zustandswechsel von einem Zustand s_i in einen anderen Zustand s_f zu Folge hat, dass das Integral

$$\int_{s_i}^{s_f} \frac{dQ}{T}$$

unabhängig vom Integrationsweg ist. Es hängt lediglich von den Initial- und Endzuständen ab. T bezeichnet hierbei die Temperatur des Wärmereservoirs, mit dem das System die Wärmemenge dQ austauscht. Dies impliziert die

Existenz einer Zustandsfunktion S, so dass gilt:

$$\int_{s_i}^{s_f} \frac{dQ}{T} = S(s_f) - S(s_i) \tag{2.1}$$

Die Funktion S wird die Entropie des System genannt. Mit diesen und weiteren Vorarbeiten formulierte Clausius schließlich seine zwei Hauptsätze der Thermodynamik, so wie sie heute benutzt werden:

1. Die Energie des Universums ist konstant.
2. Die Entropie des Universums strebt einem Maximum zu.

Weitere bedeutende Arbeiten leistete Planck und Gibbs [Gibbs 1878], da er im Gegensatz zu den bisher genannten Forschern nicht primär Prozesse, sondern Equilibriumszustände für die Einführung der Entropie betrachtet. Carathéodory und in neuerer Zeit Lieb und Yngvason [Lieb u. Yngvason 1999] formalisierten die Entropie mittels eines mathematisches Axiomenfundaments [Lieb u. Yngvason 1999, 2000, 2002].

Neben den in diesem Abschnitt genannten Arbeiten gibt es natürlich noch viele weitere, die einen Beitrag zur Definition der Entropie in der Thermodynamik geleistet haben, z. B. [Maxwell 1871] in seiner *Theory of Heat*, auf die hier nicht näher eingegangen werden soll.

2.3.2 Statistische Mechanik

Die im vorigen Abschnitt vorgestellte Notation der Entropie befasste sich mit Konzepten wie Wärme und mechanischer Energie. Sie liefert jedoch keine detaillierte physikalische Beschreibung der zugrundeliegenden Prozesse, wie Geschwindigkeit und Position der beteiligten Moleküle. Diese Problematik behandelt die statistische Mechanik [Tolman 1979], in der die Entropie, ganz allgemein gesprochen, als Maß für die Ordnung bzw. das Chaos definiert ist. Dabei stellt sich sofort die Frage, nach der genauen Definition von Ordnung und Chaos.

Eine solche Definition ist laut [Pierce 1980] immer in Zusammenhang mit Wissen zu betrachten, da beispielsweise eine sehr komplexe Anordnung von Molekülen geordnet sein kann, wenn wir die Position und Geschwindigkeit jedes Moleküles kennen. Chaos bedeutet demzufolge einen unvollständigen Wissensstand über einen Prozess, so dass zum Beispiel die Position und Geschwindigkeit der Moleküle in einem Gefäß nicht gleichzeitig vorhergesagt werden kann. Wenn wir aber einen gewissen Wissensstand über einen Prozess erlangen, verringert sich damit auch unsere Unsicherheit. Falls wir beispielsweise wissen, dass sich alle Moleküle in einem Behälter auf einer Seite befinden, also in einer bestimmten Form geordnet sind, ist unsere Unsicherheit über den Aufenthaltsort geringer, als wenn wir lediglich wissen, dass sich die Moleküle irgendwo in dem Behälter befinden.

Je detaillierter unser Wissen über das betrachtete physikalische System ist, desto geringer ist auch unsere Unsicherheit darüber. Mit diesem Wissen ist es möglich eine gewisse Struktur zu erkennen, und Ordnung zu schaffen. Durch zusätzliches Wissen, welches Ordnung schafft, ist die Entropie in unserem System geringer, als wenn wir dieses Wissen nicht hätten, also eine größere Unsicherheit vorhanden wäre.

Die formale Definition der Entropie in der statistischen Mechanik nach Boltzmann lautet

$$-k \sum_i P_i \log P_i$$

wobei k die Boltzmannkonstante ist. Die P_i geben die Wahrscheinlichkeiten an, dass sich ein bestimmtes Molekül in einem bestimmten Mikrozustand befindet. Ein Mikrozustand ist beispielsweise die kinetische Energie der Moleküle in einem Gefäß. Die P_i werden für einen Makrozustand ausgewertet, wobei ein Makrozustand zum Beispiel die Temperatur eines Systems darstellt, der dann als Erwartungswert über die Mikrozustände berechnet wird.

Zusammenfassend lässt sich also sagen, dass die Entropie in der statistischen Mechanik die Unsicherheit darstellt, in welchem makroskopischen Zustand sich das betrachtete System gerade befindet.

2.3.3 Informationstheorie

Die Informationstheorie (oder Kommunikationstheorie) hat ihren Ursprung in den Arbeiten von Morse Anfang bis Mitte des 19ten Jahrhunderts. Die zentrale Frage in dieser Domäne ist die effektive Transformation einer Nachricht im klassischen Sinne in elektrische Signale.

Im Jahre 1924 stellte R. V. L. Hartley eine Gleichung für die Entropie einer Nachricht auf, die sich später in gewisser Weise als Spezialfall der Shannonschen Entropie herausstellte [Hartley 1928]. In Hartleys Modell ist der Sender mit einer endlichen Anzahl von Symbolen ausgestattet, aus der er nacheinander eines auswählt und damit eine Sequenz von Symbolen, eine Nachricht, generiert. Er beobachtete, dass die so generierte Nachricht auch durch einen Zufallsgenerator hätte erzeugt werden können. Aufgrund dieser Beobachtung definierte er den Informationsgehalt H einer Nachricht als

$$H = n \log s$$

n gibt die Anzahl der selektierten Symbole an, und s ist die Anzahl der verschiedenen Symbole, die für die Kodierung zur Verfügung stehen. Man beachte hierbei, das dies ein mengentheoretischer Ansatz ist [Kolmogorov 1965], im Gegensatz zu dem im folgenden vorgestellten Ansatz von Shannon, welcher auf Wahrscheinlichkeiten beruht.

Nach dem zweiten Weltkrieg veröffentlichte Shannon sein bekanntes Papier [Shannon 1948] über die Entropie. Darin beschäftigte er sich mit dem

Problem, eine Nachricht effizient als elektrisches Signal zu kodieren und über eine verrauschte Leitung zu senden. Die Entropie einer Nachricht spielt dabei eine entscheidende Rolle. Shannon definierte sie als den Informationsgehalt einer Nachricht. Aus Sicht des Empfängers kann die Entropie daher auch als die Unsicherheit über die nächste, vom Sender zu sendende Nachricht, aufgefasst werden. Die formale Definition lautet

$$-\sum_i p(i) \log_2 p(i),$$

wobei $p(i)$ die Wahrscheinlichkeit ist, dass Nachricht i empfangen wird. Die Ähnlichkeit der Formel zu derjenigen von Boltzmann ist dabei nicht zu übersehen.

Die Grundlagen, die er mit seiner Arbeit gelegt hat, trugen wesentlich zum Erfolg der Informationstheorie bei. Die mathematischen Grundlagen der Shannon Entropie und ihre Eigenschaften werden in Abschnitt 2.3.5 ausführlich dargestellt.

2.3.4 Weitere Entwicklungen der Entropie

Eine grundlegende Analyse und Weiterentwicklung der von Shannon definierten Entropie wurde vor allem von Alfred Rényi durchgeführt. Rényi generalisiert in seinen Arbeiten die Entropie, in dem er sie einerseits auf eine axiomatisch definierte Basis stellt und andererseits die Voraussetzungen zu deren Definition abschwächt. Dazu führt er eine Entropie der Ordnung α ein:

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \sum_{x \in \chi} P^\alpha(x)$$

wobei P als eine Verteilung auf einer endlichen Menge χ definiert ist und $0 < \alpha < 1$ gilt. Die Shannon Entropie ergibt sich, falls $\alpha \longrightarrow 1$.

Für nähere Informationen hierzu ziehe man die Arbeiten [Rényi 1970], [Rényi 1976a], [Rényi 1976b] und [Rényi 1982] zu Rate. Ergänzende Erläuterungen finden sich in [Ciszár 1995].

Eine weitere Anwendung findet die Entropie nach der Definition von Kolmogorov in der Komplexitätstheorie [Li u. Vitányi 1993]. Er definiert die Entropie als die kürzest mögliche Beschreibung eines Objektes, also der absolut minimalen Information, mit der ein Objekt beschrieben werden kann (Beschreibungskomplexität). Im Gegensatz zu der von Shannon definierten Entropie werden hier konkrete Objekte betrachtet. Die kürzeste Beschreibungslänge eines Objektes liefert demzufolge Aussagen über dessen Komplexität, denn ein Objekt das eine kurze Beschreibung hat, ist weniger komplex als solch eines mit einer längeren Beschreibung. Beispielsweise kann eine sich wiederholende, längere Zeichenfolge wie $x = 10101010\ldots$ recht einfach durch eine Turingmaschine dargestellt werden. Bei einem komplexeren Objekt, wie

einer Primzahl, ist die inhärente Beschreibung dieser Zahl die kürzeste. Ein weiteres Beispiel: Die kleinste mit zwanzig Worten beschreibbare Primzahl. Falls es eine solche gäbe, hätten wir sie soeben mit sieben Wörtern beschrieben.

In der modernen Physik wird die Entropie nicht nur für Equilibriumszustände betrachtet, sondern auch für Zustände, die sich nicht im Equilibrium befinden. Die von [Tsallis u. a. 1988] definierte Entropie fand dabei besondere Anerkennung und hat zusätzlich auch außerhalb der statistischen Mechanik Anwendung gefunden [Plastino u. Plastino 1999]. Eine Übersicht über die Verwendung der Entropie und deren Relation zu anderen Definitionen gibt [Bashkirov 2003].

2.3.5 Theoretische Grundlagen

Im diesem Abschnitt sollen die mathematischen Grundlagen für die Definition der Entropie nach Shannon [Shannon 1948] gelegt werden.

Definition

Die Entropie in der Informationstheorie [Cover u. Thomas 1991] ist ein Maß für die Unsicherheit einer Zufallsvariable oder äquivalent deren Informationsgehalt. Sei nun X eine diskrete Zufallsvariable mit entsprechendem Alphabet Υ und $p(x) = Pr\{X = x\}, x \in \Upsilon$ sei die dazugehörige Wahrscheinlichkeitsfunktion. Damit ist die Entropie einer Zufallsvariable X definiert als:

$$H(X) = -\sum_{x \in \Upsilon} p(x) \log p(x)$$

Die üblicherweise verwendete Basis des Logarithmus ist zwei, da Shannon Information als Zustände binärer Schalter auffasste. Die Einheit der Entropie ist daher *bit*, welches übrigens eine von Shannons Kollegen Tukey geprägte Abkürzung für binary digit ist. Wie man anhand der Gleichung sieht, ist die Entropie *nicht* von den tatsächlichen Werten der Ereignisse abhängig, sondern nur von deren Wahrscheinlichkeiten.

Die Entropie kann auch als Erwartungswert von $\log \frac{1}{p(X)}$ dargestellt werden, wobei X durch die Wahrscheinlichkeitsfunktion $p(x)$ bestimmt ist:

$$H(X) = E_p \left[\log \frac{1}{p(X)} \right]$$

Weiterhin gilt, dass die Entropie immer größer gleich 0 ist:

$$H(X) \geq 0$$

Ein einfaches Beispiel soll im folgenden die Dualität zwischen Unsicherheit und Information aufzeigen. Sei X gegeben mit:

$$X = \begin{cases} 1 & \text{mit Wahrscheinlichkeit } p \\ 0 & \text{mit Wahrscheinlichkeit } 1 - p \end{cases}$$

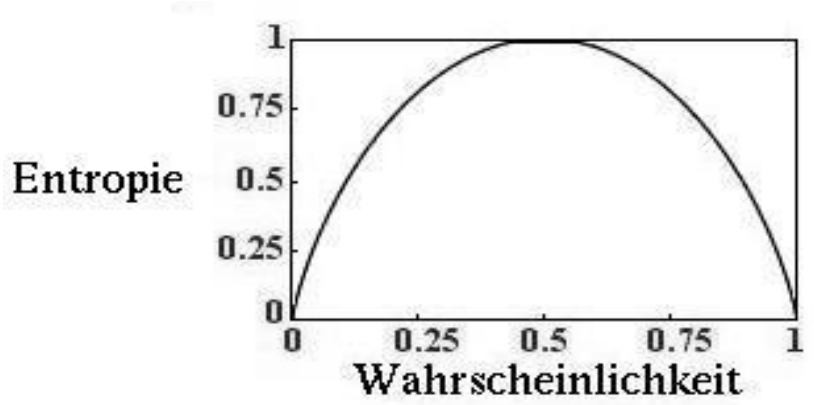


Abbildung 2.6: Entropie $H(p)$ aufgetragen über Wahrscheinlichkeit p

Damit ist

$$H(X) = -p \log p - (1-p) \log(1-p) =: H(p)$$

In Abb. 2.6 sieht man den Verlauf von $H(X)$. Die Entropie erreicht bei $p = \frac{1}{2}$ ihr Maximum von eins. Das ist intuitiv verständlich, denn die größte Unsicherheit besteht genau dann, wenn sämtliche Werte gleichverteilt sind. Damit besitzt man keinerlei Informationen über die Struktur der Wahrscheinlichkeitsverteilung.

Dagegen ist die Entropie für $p = 0$ bzw. $p = 1$ gleich Null, d.h. es ist keine Unsicherheit vorhanden, da die Variable X ja nicht mehr zufällig ist, sondern im Voraus schon bekannt ist, dass das modellierte Ereignis nie oder immer eintritt. Demzufolge ist durch Beobachten des Ereignisses auch kein Informationsgewinn vorhanden.

Verbundentropie und bedingte Entropie

Die Verbundentropie (joint entropy) $H(X,Y)$ zweier diskreter Zufallsvariablen X und Y mit der gemeinsamen Verteilung $p(x,y)$ ist definiert als:

$$H(X,Y) = -\sum_{x \in \Upsilon} \sum_{y \in \Phi} p(x,y) \log p(x,y)$$

Die bedingte Entropie (conditional entropy) einer Zufallsvariable ist definiert als der Erwartungswert der Entropien der bedingten Verteilungen, gemittelt über die bedingende Zufallsvariable. Sei (X,Y) Zufallsvariablen

mit dazugehöriger Verteilung $p(x, y)$:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \Upsilon} p(x) H(Y|X = x) && (2.2) \\ &= -\sum_{x \in \Upsilon} \sum_{y \in \Phi} p(x, y) \log p(y|x) && (2.3) \\ &= -E_{p(x,y)} \log p(Y|X) && (2.4) \end{aligned}$$

Die Verbundentropie besagt, dass die Entropie zweier Zufallszahlen der Entropie einer Zufallszahl plus der bedingten Entropie der anderen Zufallszahl entspricht.

$$H(X, Y) = H(X) + H(Y|X)$$

Daraus ergibt sich dann direkt folgende Erweiterung der Kettenregel:

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$$

Es ist noch anzumerken, dass die bedingte Entropie nicht symmetrisch ist, also $H(X|Y) \neq H(Y|X)$.

Relative Entropie

Die Entropie in der Informationstheorie bedeutet den bisherigen Definitionen zufolge also die durchschnittlich benötigte Information, gemessen in bits, um eine Zufallsvariable zu beschreiben. In diesem und nächsten Abschnitt sollen nun die verwandten Konzepte der *relativen Entropie* und der *gegenseitigen Information* vorgestellt werden.

Vor allem die relative Entropie, auch bekannt als Kullback-Leibler Distanz nach [Kullback u. Leibler 1951] oder minimale Kreuzentropie im Bereich des nichtmonotonen Schließens [Jaynes 1979, 1981; Goldszmidt u. a. 1993; Benferhat u. a. 1997; Bourne 2003], spielt im späteren Verlauf dieser Arbeit eine entscheidende Rolle. Hier soll lediglich die mathematische Definition und deren Interpretation in der Wahrscheinlichkeitstheorie gegeben werden.

Die relative Entropie $D(p||q)$ wird in der Wahrscheinlichkeitstheorie als Abstandsmaß zwischen Wahrscheinlichkeitsverteilungen p und q genutzt. Dabei muss allerdings beachtet werden, dass die relative Entropie kein Distanzmaß im metrischen Sinne ist, denn sie erfüllt nicht die Axiome der Symmetrie und die Dreiecksungleichung.

Die relative Entropie $D(p||q)$ ist auch ein Maß für Ineffizienz. Dazu wird beispielsweise bei einem Experiment angenommen, die Verteilung der Daten sei q, die wirkliche, unbekannte Verteilung ist aber p. Nun nehmen wir an, wir wüssten die wahre Verteilung p der Zufallsvariablen. Damit wäre die optimale Kodierung $H(p)$ bits. Wenn wir diese Verteilung aber nicht kennen, sondern mit q versuchen zu approximieren, brauchen wir im Schnitt $H(p) + D(p||q)$

bits um diese Zufallsvariable zu beschreiben. Formal ist die relative Entropie zwischen zwei Wahrscheinlichkeitsverteilungen $p(x)$ und $q(x)$ definiert als:

$$D(p||q) = \sum_{x \in \Upsilon} p(x) \log \frac{p(x)}{q(x)} \tag{2.5}$$

Ein Nachteil dieses informationstheoretischen Distanzmaßes, welches zur Klasse der Ali-Silvey Maße gehört [Ali u. Silvey 1966], ist die Konvergenz gegen unendlich, falls die Wahrscheinlichkeit von $q(x)$ Null ist: $p \log \frac{p}{0} = \infty$. Diesen Nachteil besitzt zum Beispiel die Chernoff-Distanz nach [Chernoff 1952] nicht. Jedoch ist sie schwieriger zu berechnen, da hier Optimierungsprobleme gelöst werden müssen.

Gegenseitige Information

Die gegenseitige Information, welche in Abschnitt 3 auf Ontologien erweitert wird, ist ein Maß für die Information die eine Zufallsvariable über eine andere enthält. Die Dualität von Information und Unsicherheit berücksichtigend lässt sich auch sagen, dass die Verringerung der Unsicherheit der einen Zufallsvariable durch das Wissen der anderen beschrieben wird.

Gegeben seien zwei diskrete Zufallsvariablen X und Y mit einer gemeinsamen Wahrscheinlichkeitsverteilung $p(x, y)$ und den marginalen Wahrscheinlichkeitsverteilungen $p(x)$ und $p(y)$. Dann ist die gegenseitige Information $I(X;Y)$ definiert als die relative Entropie zwischen der Verbund- und der Produktverteilung $p(x)p(y)$:

$$I(X;Y) = \sum_{x \in \Upsilon} \sum_{y \in \Phi} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.6}$$

$$= D(p(x,y)||p(x)p(y)) \tag{2.7}$$

Die gegenseitige Information $I(X;Y)$ kann auch als eine Summe von Entropie und bedingter Entropie definiert werden:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.8}$$

$$= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y)) \right) \tag{2.9}$$

$$= H(X) - H(X|Y) \tag{2.10}$$

Da die Gleichung

$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

wegen

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

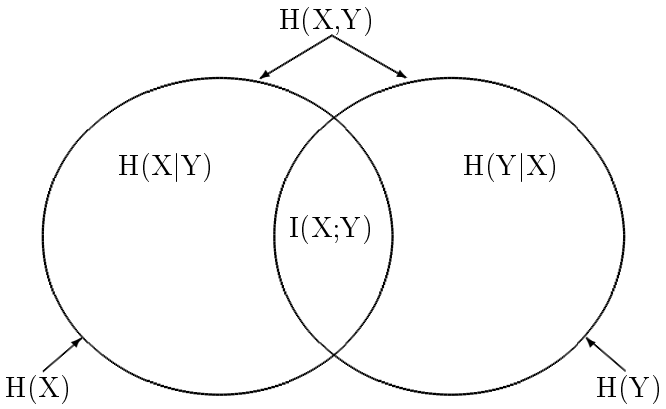


Abbildung 2.7: Beziehung zwischen Entropie und gegenseitiger Information

symmetrisch ist, folgt für die gegenseitige Information $I(X;Y)$ ebenfalls Symmetrie:

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = I(Y;X)$$

Das heißt, X enthält ebensoviel Information über Y, wie Y über X. Die Entropie wird manchmal auch Selbstinformation über eine Zufallsvariable X genannt. Der Grund hierfür wird bei der Betrachtung von $I(X;X)$ sichtbar:

$$I(X;X) = H(X) - H(X|X) = H(X)$$

$H(X|X)$ ist dabei gleich Null, da X über sich selbst keine zusätzliche Information liefert.

Der Zusammenhang zwischen Entropie und gegenseitiger Information wird nochmals anhand des Venn-Diagramms in Abb. 2.7 verdeutlicht.

Weitere Eigenschaften

In diesem Abschnitt sollen noch kurz einige weitere Eigenschaften der Entropie, relativen Entropie und gegenseitigen Information [Wehrl 1978] gegeben werden, da sie in den folgenden Kapiteln benötigt werden.

Seien $p(x), q(x)$ zwei Wahrscheinlichkeitsfunktionen mit $x \in \Upsilon$. Dann gilt für die relative Entropie:

$$D(p||q) \geq 0$$

mit Gleichheit, genau dann wenn

$$p(x) = q(x) \quad \forall x.$$

Die gegenseitige Information für zwei Zufallsvariablen X, Y ist Null, wenn sie unabhängig voneinander sind und ansonsten größer Null:

$$I(X;Y) \geq 0.$$

Die bedingte, gegenseitige Information einer Zufallsvariable X, unter der Bedingung Y und Z ist definiert als:

$$I(X;Y|Z) \quad = \quad H(X|Z) - H(X|Y,Z)$$

Für sie gilt, ähnlich der gegenseitigen Information, ebenfalls:

$$I(X;Y|Z) \geq 0,$$

mit Gleichheit genau dann wenn X sowie Y, gegeben Z unabhängig voneinander sind.

Weiterhin gilt,dass die Entropie maximal ist, wenn die Elemente in Υ gleichverteilt sind:

$H(X) \leq \log|\Upsilon|$, mit $|\Upsilon|$ Anzahl der Elemente von Υ, und Gleichheit genau dann wenn X gleichverteilt über Υ ist.

Folgendes Theorem besagt, dass im Schnitt die Kenntnis einer zusätzlichen Zufallsvariable Y die Unsicherheit der Zufallsvariable X reduziert. Dies gilt nur im Durchschnitt, denn es kann durchaus sein, dass $H(X|Y = y)$ größer oder gleich $H(X)$ ist.

$$H(X|Y) = \sum_y p(y)H(X|Y = y) \leq H(X)$$

Ein kleines Beispiel wäre ein Fall vor einem Gericht, wo ein zusätzlicher Beweis die Unsicherheit im Moment zwar erhöhen kann, aber über das ganze Verfahren betrachtet doch die Unsicherheit senkt.

Kapitel 3

Entropie für Ontologien

Im folgenden Kapitel wird eine Vorgehensweise, Wissen mittels entropiebasierter Distanzmaße zu strukturieren, nach [Daemi u. Calmet 2004b] vorgestellt. Dazu wird zunächst zum Vergleich eine klassische Vorgehensweise zur Strukturierung von Daten vorgestellt. Danach wird die Strukturierung von Wissen, welches durch Ontologien repräsentiert werden kann, anhand von entropiebasierten Distanzmaßen näher erläutert. Anschließend erfolgt die Vorstellung eines entropiebasiertes Distanzmaß, die gegenseitige Information, welches erstmals auf Ontologien angewandt und verfeinert wurde [Calmet u. Daemi 2004].

3.1 Klassische Strukturierung

In Abb. 3.1 ist eine klassische Vorgehensweise zur Strukturierung von Daten, wie zum Beispiel einer Messreihe, dargestellt. Daten sind hierbei als Einzelfallbezogene Informationen über einen konkreten Sachverhalt (Experiment) zu verstehen. Information wird als inhaltliches Wissen angesehen, ohne weitere Bestimmungen und Zusätze. Informationen stellen also einen kriterienfreien Wissensinhalt oder Wissensbeitrag dar [Weber u. a. 2002]. Die erhaltenen Daten müssen nun interpretiert werden, damit die gewonnenen Informationen in qualifizierte Informationen, d.h. Erkenntnisse, transformiert werden können. Die Interpretation der Daten erlaubt es somit, Einsichten in den die Daten erzeugenden Prozess zu erhalten. Ein Weg, zu einer erkenntnisreichen Interpretation der Daten zu gelangen, ist diese nach bestimmten Strukturen zu durchsuchen oder verborgene Strukturen zu erkennen. Es gibt hierbei viele verschiedene Möglichkeiten, um Strukturen auf Daten zu erkennen. Eine oft beschrittene Möglichkeit zur Strukturierung ist die Daten zu clustern, d.h. Mengen ähnlicher Merkmale, zusammenzufassen. Diese werden anschließend mit verschiedenen Klassifikationsalgorithmen [Celeux u. a. 1989] weiter zusammengefasst und anhand unterschiedlichster Kriterien strukturiert.

In dieser Arbeit wird eine weitere Möglichkeit betrachtet, Daten zu struk-

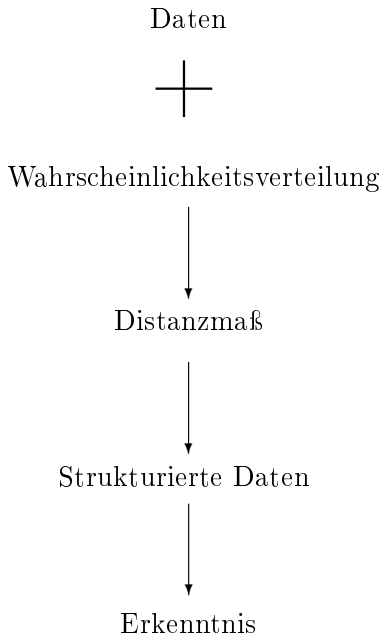


Abbildung 3.1: Klassisches Modell zur Strukturierung von Daten.

turieren, welche auf Wahrscheinlichkeiten und den dazugehörigen Distanzmaßen basiert. Zur Strukturierung der Daten wird dabei zunächst versucht eine möglichst gut zu den Daten passende Wahrscheinlichkeitsverteilung zu finden. Das Finden einer solchen Wahrscheinlichkeitsverteilung ist nicht trivial und wird ausführlich im dem Gebiet der Statistik [Gaul 2000] behandelt. Eine besondere Herausforderung besteht bei dieser Vorgehensweise darin, die Eindeutigkeit einer solchen Wahrscheinlichkeitsverteilung zu zeigen. Es kann durchaus vorkommen, dass die ausgewählte Wahrscheinlichkeitsverteilung nicht die einzige ist, die optimal zu den Daten passt.

In den weiteren Betrachtungen nehmen wir aber an, dass eine geeignete Wahrscheinlichkeitsverteilung gefunden wurde. Damit kann mit einer ganzen Reihe von Distanz- oder Dissimilaritätsmaßen [Kotz u. Johnson 1981], oder den in Abschnitt 2.2 vorgestellten Distanzmaßen, eine Strukturierung der Daten vorgenommen werden. Mit der gewonnenen Struktur der Daten ist es möglich, Erkenntnisse anhand von Theorien, Modellen oder Vergleichen mit bereits bekannten Strukturen zu gewinnen.

Die vorgestellte Vorgehensweise stellt also eine Möglichkeit dar, mittels Wahrscheinlichkeitsverteilungen und den dazugehörigen Distanzmaßen Daten zu strukturieren um dann mit Hilfe der gewonnenen Strukturen zu Erkenntnissen zu gelangen. Im nächsten Abschnitt soll, ebenfalls mit Hilfe von Wahrscheinlichkeiten und Distanzmaßen, eine Vorgehensweise zur Struktu-

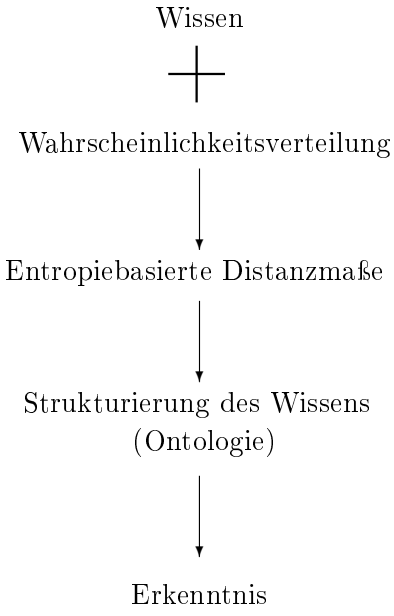


Abbildung 3.2: Modell zur Strukturierung von Wissen (Ontologie)

rierung von Wissen vorgestellt werden, wobei insbesondere die entropiebasierten Distanzmaße näher betrachtet werden.

3.2 Entropiebasierte Strukturierung

Das in Abb. 3.2 gezeigte Modell beschreibt, wie mittels entropiebasierter Distanzmaße Wissen, welches insbesondere durch Ontologien[1] darstellbar ist, strukturiert werden kann. Wissen ist nach [Weber u. a. 2002] der grundlegende, integrierende, wissenstheoretische Oberbegriff der gesamten Wissensterminologie. Als Wissen wird *semantische Information* bezeichnet, wobei das Wissen unabhängig von Richtigkeit und Wichtigkeit zu sehen ist. Wissen ist also inhaltliche Information über angenommene, dargestellte, behauptete, etc. Sachlagen, ohne Rücksicht auf deren Wahrheitswert oder sonstige Zusatzqualifizierungen. Der Informationsgehalt der semantischen Information kann dabei durchaus nicht vorhanden, also gleich Null sein, wie beispielsweise bei Tautologien.

Ontologien oder Taxonomien sind zwar per Definition bereits nach einem oder mehreren Kriterien strukturiert, beispielsweise anhand einer Vererbungsrelation. Die so erhaltene Struktur muss aber nicht notwendigerweise eindeutig sein. Das wohl prominenteste Beispiel für eine nicht eindeutige

[1] Das in der Ontologie repräsentierte Wissen wird normalerweise als vollständig für den jeweiligen Anwendungsbereich betrachtet. Ebenso sollte dieses Wissen sicher sein, da es von einer hinreichend großen Expertengruppe akzeptiert sein sollte.

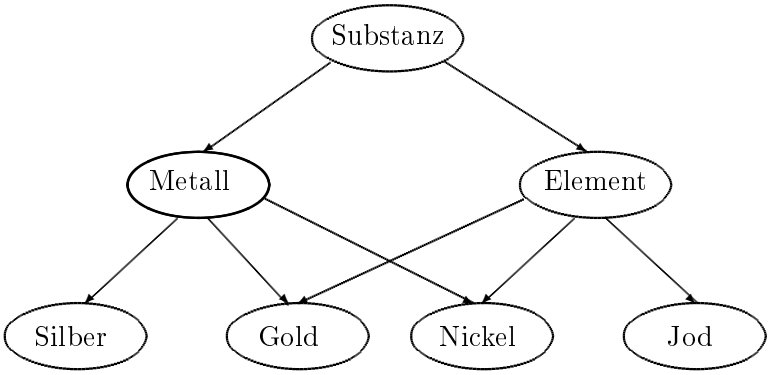


Abbildung 3.3: Mehrfachvererbung

Struktur ist der Nixon Diamant nach [Reiter 1980]. Ein weiteres Beispiel wäre die Frage nach dem kleinsten gemeinsamen Elternkonzept bei Mehrfachvererbung in einer durch eine Vererbungsrelation strukturierten Ontologie. In Abb. 3.3, die einen Ausschnitt aus der Wordnet Taxonomie [Miller u. a. 1993] zeigt, kann man beispielsweise nicht entscheiden, ob das kleinste gemeinsame Elternkonzept von Nickel nun Metall oder Element ist. In solchen Fällen ist eine genaue Aussage nicht ohne zusätzliche Information möglich.

Die ergänzende Information sei, wie bei der klassischen Vorgehensweise, durch eine Wahrscheinlichkeitsverteilung gegeben, welche somit die Grundlage für eine weitergehende Strukturierung bildet. Dieser Schritt ist für den weiteren Verlauf der Strukturierung ebenso entscheidend wie in der klassischen Vorgehensweise. Jedoch ist es im allgemeinen deutlich schwieriger eine Wahrscheinlichkeitsverteilung für Wissen zu finden als zum Beispiel für Daten. Eine Wahrscheinlichkeitsverteilung muss dabei nicht notwendigerweise eine objektive, das heißt frequenzbasierte Interpretation aufweisen, wie dies üblicherweise bei der im vorherigen Abschnitt beschriebenen Vorgehensweise der Fall ist.

Eine Wahrscheinlichkeitsverteilung kann auch als *Degree-of-Belief* oder *Grad-des-Dafürhaltens* interpretiert werden. Damit dient die Wahrscheinlichkeitsverteilung zur Verkörperung von Ungewissheit unter Berücksichtigung bereits vorhandenen Wissens [Beyerer 1999]. Die Deutung des Degree-of-Belief lässt sich weiter zergliedern in einen objektiven und subjektiven Fall, deren Anwendung für die beschriebene Vorgehensweise zur Strukturierung ebenfalls zulässig ist. Bei der *objektiven* Auffassung des Degree-of-Belief sollen Fakten, die zu einer Meinung Anlass geben, wahrscheinlichkeitstheoretisch verkörpert werden. Es wird auf Grundlage vorhandener Fakten eine Wahrscheinlichkeitsverteilung generiert, aus der eben diese Fakten wieder rekonstruiert werden können, die darüber hinaus aber möglichst keine Aussagen macht über Dinge, die nicht zwingend aus diesen Fakten folgen. Bei-

spielhaft für solch eine Modellierung von *Nichtwissen* ist das Prinzip der maximalen Entropie [Jaynes 1982; Shore u. Johnson 1981; Uffink 1995; Paris u. Vencovská 1990, 1997]. Anwendung fand dies Prinzip unter anderem im Bereich des nichtmonotonen Schließens [McCarthy 1980; Reiter 1980; Doyle 1988; Touretzky 1986; Kraus u. a. 1990; Pearl 1991; Makinson 2003].

Bei der *subjektiven* Deutung hingegen dient die Wahrscheinlichkeit zur Darstellung des Wissensstandes und der Überzeugungen eines Individuums. Hierbei spielt es keine Rolle, ob die Wahrscheinlichkeiten „wahr“ oder ob sie nur das Individuum für richtig hält. Dabei ist zu beachten, dass die ausgewählte Wahrscheinlichkeitsverteilung den Daten und damit auch dem Distanzmaß bereits eine semantische Deutung zuordnet.

Falls die Ontologie in Form eines Graphen vorliegt oder in einen solchen überführt werden kann, werden die Wahrscheinlichkeitsverteilungen durch entsprechende Werte an den Kanten zwischen den Konzepten dargestellt. Die so zugewiesene Wahrscheinlichkeitsverteilung erlaubt die Einführung entropiebasierter Distanzmaße, von denen in dieser Arbeit die im nächsten Abschnitt vorgestellte gegenseitige Information und die relative Entropie (siehe Kapitel 5) für ihre Anwendung auf Ontologien untersucht wurden. Insbesondere auch deshalb, da in [Weber u. a. 2002] jede Art von Information (syntaktisch, semantisch, pragmatisch) als „Möglichkeitsausschluss im Wissensraum“ angesehen wird. Die semantische Information gibt hierbei den Informationsgehalt einer Aussage an, welcher definiert ist als die Klasse der ausgeschlossenen, d.h. aussagenkonträren logischen bzw. empirischen Möglichkeiten im jeweiligen Wissensraum. Die aus [Weber u. a. 2002] übernommene Begriffsbildung für Wissen als „semantische Information“ bietet somit die Benutzung informationstheoretischer Maße und der Entropie im Sinne von „Informationsgehalt“ an.

Damit kann, wie im klassischen Fall, Wissen strukturiert werden, um zu weiteren Erkenntnissen zu gelangen. Falls das Wissen schon vorher eine Struktur besessen hatte, wird eine zusätzliche Strukturierung vorgenommen oder es werden Mehrdeutigkeiten in der Struktur aufgelöst.

Es ist anzumerken, dass sich eine Implementierung der Strukturierung von Daten mit Hilfe der Entropie in [Kern-Isberner 2001] findet. Als Wissensbasis dienen hier allerdings probabilistische Regeln inklusive ihrer bedingten Wahrscheinlichkeiten und keine Ontologien. Der Mechanismus zur Strukturierung basiert auf dem Prinzip der minimalen Kreuzentropie (minimum cross-entropy) [Kern-Isberner 1998; Rödder u. Kern-Isberner 1997].

3.3 Gegenseitige Information für Ontologien

Mit Hilfe der gegenseitigen Information soll Wissen, welches durch Ontologien darstellbar und im Sinne von [Weber u. a. 2002] zu sehen ist, strukturiert werden. Der Diskursbereich des zu strukturierenden Wissens wird dabei zu-

nächst durch eine Ontologie festgelegt. Anschließend werden den Konzepten in der Ontologie konkrete Instanzen des zu strukturierenden Wissens zugewiesen. Beispielsweise kann dem Konzept *Rhythmus* in der Musikontologie (siehe Kapitel 6) ein konkreter Rhythmus eines bestimmten Taktes des zugrundeliegenden Musikstückes zugewiesen werden. Zur Strukturierung von Ontologien, also beispielsweise der Musikstücke anhand eines bestimmten Rhythmuses, wird im folgenden die gegenseitige Information für Ontologien eingeführt.

3.3.1 Voraussetzungen

Dazu müssen zunächst Wahrscheinlichkeitsverteilungen, mit zugehörigen Zufallsvariablen, für eine Ontologie und das durch sie modellierte Wissen bestimmt werden. Die Wahrscheinlichkeitsverteilungen können dabei sowohl eine kognitive Bedeutung aufweisen, also als Degree-of-Belief interpretiert werden, als auch eine frequenzbasierte Interpretation besitzen. In obigem Musikbeispiel können die Wahrscheinlichkeiten die relative Häufigkeit des Rhythmus, Anzahl gleicher Harmonien u.s.w. in dem Musikstück darstellen. Bei einer kognitive Bedeutung der Wahrscheinlichkeitsverteilungen würden sie die Interessantheit des Rhythmus, Spannung der Melodie und ähnliches repräsentieren.

Bisherige auf Ontologien oder Taxonomien definierte Distanzmaße gehen grundsätzlich von einer objektiven, frequenzbasierten Wahrscheinlichkeitsverteilung aus, die in geeigneter Weise auf der Ontologie definiert ist. In [Resnik 1995] werden beispielsweise die relativen Häufigkeiten von Wörtern in Dokumenten den dazugehörigen Konzepten der Ontologie zugeordnet. Ein Problem bei der Benutzung dieser frequenzbasierter Distanzmaße ist die langsame Konvergenz innerhalb des Teils des Vokabulares, der die meisten relevanten Begriffe enthält. Zudem ist eine genaue Bestimmung dieses Bereiches schwierig.

Bevor die Anwendung der gegenseitige Information für Ontologien vorgestellt wird, soll die Bedeutung einer Zufallsvariablen X auf Ontologien erläutert werden. Eine Zufallsvariable repräsentiert im einfachsten Fall die Wahrscheinlichkeitsverteilung $\mathbf{p} = p_1, \ldots, p_n$ eines Begriffes der Ontologie:

$$X = (p_1, \ldots, p_n)$$

Die konkreten Wahrscheinlichkeiten ergeben sich dabei aus dem der Ontologie bzw. dem Begriff zugrundeliegenden Wissen. Dies kann, im Falle der frequenzbasierten Interpretation, beispielsweise die relative Häufigkeit $\mathbf{p}$ des Begriffes in einem Dokumentenkorpus sein oder die eines bestimmten Rhythmuses in einem Musikstück. Falls subjektive Wahrscheinlichkeiten verwendet werden, ist eine beliebige kognitive Bedeutung möglich. Die Wahrscheinlichkeiten können dann zum Beispiel für die Wichtigkeit oder Interessantheit

bezüglich eines durch die Ontologie beschriebenen Sachverhaltes stehen. Allerdings können auch komplexere Wahrscheinlichkeitsverteilungen gegeben sein. Eine komplexere Wahrscheinlichkeitsverteilung könnte sich beispielsweise aus den Eigenschaften der Konzepte, also ihrer Attribute und Relationen zu anderen Konzepten, berechnen wie in dem probabilistischen System von [Koller u. Pfeffer 1998]. Die exakte Bedeutung einer Zufallsvariable hängt also primär von der Interpretation der Wahrscheinlichkeitsverteilungen ab, die auf dem Wissen und damit der Ontologie definiert werden.

3.3.2 Anwendung auf Ontologien

Da die Konzepte nun mit einem Wahrscheinlichkeitsmaß versehen wurden, ist es möglich die Entropie und mit ihr verbundene Distanzmaße, insbesondere die gegenseitige Information, auf einer Ontologie zu definieren. Eine Ontologie sei für die folgenden Betrachtungen in Form eines gerichteten, hierarchischen, azyklischen Graphen modelliert. Hierarchisch bedeutet in diesem Zusammenhang, dass alle Knoten vom Wurzelknoten aus erreichbar sind. Im folgenden sei die Entropie $H(X)$ mit einem Konzept der Ontologie assoziiert. Die Entropie lässt sich daher direkt aus der dem Konzept zugeordneten Zufallsvariable X berechnen. Für die Berechnung der gegenseitigen Information müssen zusätzlich die bedingten Wahrscheinlichkeiten $P(X|Z)$ zwischen Konzepten X und Z gegeben sein. Die Berechnung der bedingten Wahrscheinlichkeiten erfolgt dabei nur zwischen Konzepten, welche durch eine Relation miteinander verbunden sind. Diese bedingten Wahrscheinlichkeiten werden mit den Kanten (Relationen) zwischen den Konzepten assoziiert, aus welchen sich schließlich die gegenseitige Information berechnen lässt.

Damit ist die gegenseitige Information für Ontologien als die Reduktion der Unsicherheit in einem Konzept X durch das Heranziehen eines weiteren Konzeptes Z definiert. Zwischen den Konzepten X und Z existiere dazu eine direkte Verbindung in der Ontologie mit dazugehöriger, bedingter Wahrscheinlichkeit. Aus dieser lässt sich dann die bedingte Entropie $H(X|Z)$ berechnen. Die mathematische Formulierung dieses Sachverhaltes wird ähnlich der gegenseitigen Information dargestellt:

$$d(X;Z) := I(X;Z) = H(X) - H(X|Z)$$

$H(X)$ bezeichne die Entropie oder den Informationsgehalt von X. Von dieser wird die bedingte Entropie $H(X|Z)$, also der Informationsgehalt von X, falls Z bekannt ist, abgezogen.

Für die Anwendung auf Ontologien wird die gegenseitige, bedingte Information verwendet, da in einer Ontologie zwischen nicht direkt benachbarten Konzepten X und Z mehrere Konzepte liegen können. Genauer gesagt, wird die Reduktion der Unsicherheit von Konzept X durch das Wissen über Z und eine weiterer Informationsgewinn durch ein zusätzliches Konzept Y, welches

auf dem Pfad zwischen X und Z liegt, betrachtet:

$$d(X;Y,Z) = H(X|Z) - H(X|Y,Z)$$

Da mehrere Konzepte auf dem Pfad zwischen X und Z liegen können, wird $\mathbf{Y}$ als Vektor $Y_1, Y_2, \ldots, Y_n$ der dazwischenliegenden Konzepte aufgefasst:

$$\begin{aligned} d(X;\mathbf{Y},Z) &= d(X;Y_1,Y_2,\ldots,Y_n,Z) && (3.1)\\ &= H(X|Z) - H(X|\mathbf{Y},Z) && (3.2)\\ &= H(X|Z) - H(X|Y_1,Y_2,\ldots,Y_n,Z) && (3.3) \end{aligned}$$

Für die Y_i gilt dabei die Einschränkung, dass mindestens eines Informationgewinn gegenüber Z liefern muss, oder anders ausgedrückt: Mindestens ein Y_i darf nicht komplett abhängig von Z sein:

$$\exists\, i \;: H(Y_i|Z) \neq 0$$

Wenn $X = Z$ gilt, haben wir selbstverständlich keinen Informationsgewinn, sondern es steht uns nur die Selbstinformation von X zur Verfügung:

$$d(X;X,0) = H(X)$$

Weiterhin gilt, dass die zusätzliche Information Z und $\mathbf{Y}$ niemals die Entropie oder Unsicherheit von X erhöhen können, denn diese Information kann einfach ignoriert werden. Damit gilt die Ungleichung

$$H(X) \geq H(X|Z) \geq H(X|\mathbf{Y},Z) \tag{3.4}$$

Somit folgt für $d(X;Z,\mathbf{Y})$ positive Definitheit:

$$\begin{aligned} d(X;\mathbf{Y},Z) &= d(X;Y_1,\ldots,Y_n,Z) && (3.5)\\ &= \underbrace{H(X|Z)}_{\geq H(X|\mathbf{Y},Z)} - H(X|Y_1,\ldots,Y_n,Z) && (3.6)\\ \Rightarrow H(X|Z) - H(X|\mathbf{Y},Z) &\geq 0 && (3.7)\\ \Rightarrow d(X;\mathbf{Y},Z) &\geq 0 && (3.8) \end{aligned}$$

Es gilt auch die Symmetrieeigenschaft der gegenseitigen Information, welche besagt, dass der Begriff X über die Y_i unter der Bedingung Z, genausoviel Information liefert wie die Y_i unter der Bedingung Z über X:

$$\begin{aligned} d(X;\mathbf{Y},Z) &= H(X|Z) - H(X|\mathbf{Y},Z) && (3.9)\\ &= I(X;\mathbf{Y}|Z) && (3.10)\\ &= I(X;Y_1,\ldots,Y_n|Z) && (3.11)\\ &= I(Y_1,\ldots,Y_n|Z;X) && (3.12)\\ &= I(\mathbf{Y}|Z;X) && (3.13)\\ &= d(\mathbf{Y},Z;X) && (3.14) \end{aligned}$$

3.3.3 Verfeinerung

Die Verfeinerung der gegenseitigen Information für Ontologien soll eine detaillierte Strukturierung der Ontologien bzw. ihrer Wissensbasen ermöglichen. Dazu werden zusätzlich zu dem kürzesten Weg zwischen den Konzepten weitere Pfade betrachtet, welche einen hinreichend großen Informationsgewinn liefern.

Sei eine Ontologie als hierarchischer, azyklischer Graphen modelliert oder in eine äquivalente Repräsentationen überführbar. Die Y_i seinen dann diejenigen Konzepte, welche auf einem Pfad zwischen Konzepten X und Z liegen, bezüglich derer eine Strukturierung vorgenommen werden soll.

Die gegenseitige Information soll dazu nicht nur auf einem Pfad, normalerweise dem kürzesten, zwischen zwei Begriffen X und Z mit dazwischenliegenden Begriffen Y_i betrachtet werden, sondern es sollen weitere l Pfade in Betracht gezogen werden. Diese l Pfade sollen dabei noch einen hinreichend großen Informationsgewinn liefern. Hierbei seien $\mathbf{Y} = Y_1, \ldots, Y_n$ diejenigen Konzepte die auf einem Weg liegen. $\mathbf{Y}^l = Y_1^l, \ldots, Y_n^l$ sind dann diejenigen Konzepte, welche auf einem der $l = 1, \ldots, j$ Pfade zwischen X und Z liegen:

$$\overline{d}_{X \to Z}(X; \mathbf{Y}^l, Z) \quad = \quad \sum_{j=1}^{l} H(X|Z) - H(X|Y_1^j, \ldots, Y_n^j, Z)$$

Hier betrachtet man also den Informationsgewinn, welchen die l Pfade mit Konzepten $\mathbf{Y}^l = Y_1^l, \ldots, Y_n^l$ liefern, jeweils zusätzlich zu dem Informationsgewinn den Z über X liefert. Die Berechnung der einzelnen Summen in endlicher Zeit kann sichergestellt werden, da der Graph zyklenfrei ist. Desweiteren gelten alle Voraussetzung und Einschränkungen aus dem vorherigen Abschnitt. Die genaue Anzahl l der zu betrachtenden Pfade ist von der jeweiligen Anwendung abhängig. Sie sollte so gewählt sein, dass der Informationsgewinn durch die betrachteten Konzepte noch hinreichend groß ist. Eine geschickte Strategie bei der Wahl der Pfade zwischen X und Z wäre, zunächst die kürzesten Pfade der Länge k zu betrachten, da diese den meisten Informationsgewinn versprechen. Die auf dem kürzesten Pfad liegenden Konzepte müssen mindestens betrachtet werden, um einen Zusammenhang zwischen X und Z herstellen zu können. Anschließend kann man den Informationsgewinn auf Pfaden der Länge $k+1, k+2, \ldots$ untersuchen. Falls dieser, in Bezug auf den kürzesten Pfad, unter einen gewissen Schwellwert fällt, kann von der Betrachtung weiterer Pfade abgesehen werden.

Im nächsten Schritt wird noch eine Normalisierung des Maßes mit der Anzahl l der betrachteten Pfade vorgenommen, damit die Vergleichbarkeit der Distanzen gewährleistet bleibt, wenn jeweils unterschiedlich viele Pfade

betrachtet werden.

$$d_{X\to Z}(X;\mathbf{Y}^l,Z) = \frac{1}{l}\left(\sum_{j=1}^{l} H(X|Z) - H(X|Y_1^j,\ldots,Y_n^j,Z)\right)$$

Auch hier gilt, dass die zusätzliche Information Z und Y_i^l niemals die Entropie oder Unsicherheit von X erhöhen können, denn diese Information kann einfach ignoriert werden. Damit gilt hier ebenfalls die Ungleichung 3.4 und für $d_{X\to Z}(X;\mathbf{Y}^l,Z)$ folgt:

$$\begin{aligned} d_{X\to Z}(X;\mathbf{Y}^l,Z) &= \underbrace{l\cdot H(X|Z)}_{\geq H(X|\mathbf{Y},Z)\geq 0} - \sum_{j=1}^{l} H(X|Y_1^j,\ldots,Y_n^j,Z) \\ &\geq 0 \end{aligned}$$

Somit ist die *positive Definitheit* gewährleistet:

$$\forall X,\mathbf{Y}^l,Z: \quad d_{X\to Z}(X;\mathbf{Y}^l,Z)\geq 0$$

Die Symmetrieeigenschaft ist für $d_{X\to Z}(X;\mathbf{Y}^l,Z)$ erfüllt:

$$\begin{aligned} d_{X\to Z}(X;\mathbf{Y}^l,Z) &= \frac{1}{l}\left(\sum_{j=1}^{l} H(X|Z) - H(X|Y_1^j,\ldots,Y_n^j,Z)\right) \\ &= \frac{1}{l}\left(\sum_{j=1}^{l} H(Z|X) - H(Y_1^j,\ldots,Y_n^j,Z|X)\right) \\ &= d_{X\to Z}(\mathbf{Y}^l,Z;X) \end{aligned}$$

Ein Problem bei der Anwendung der gegenseitigen Information zur Strukturierung von Ontologien ist die konkrete Berechnung von $d_{X\to Z}(X;\mathbf{Y}^l,Z)$, insbesondere der dafür benötigten bedingten Wahrscheinlichkeiten. Vor allem bei größeren Ontologien mit sehr vielen Wahrscheinlichkeiten sind diese nur aufwändig zu berechnen. Aus diesem Grund wird in Kapitel 5 die Anwendung der relativen Entropie auf Ontologien vorgeschlagen, da sie effektiv berechnet werden kann. Anhand eines Beispieles mit der Flutontologie und einer konkreten Anwendung zur Strukturierung der Wissensbasis der Musikontologie wird die Anwendung der relativen Entropie auf Ontologien verdeutlicht.

Es ist anzumerken, dass die gezeigte Vorgehensweise in Analogie zu der Entwicklung einer Funktion mit einer Taylorreihe gesehen werden kann. Die bedingte Wahrscheinlichkeit $H(X|Z)$ kann als Taylorentwicklung ersten Grades von $H(X)$ um X, unter der Bedingung Z, angesehen werden. Die einzelnen Summanden von $\sum_{j=1}^{l} H(X|Y_1^j,\ldots,Y_n^j,Z)$ sind entsprechend die j-ten Annäherungen bzw. Ableitungen an den Informationsgehalt von X, unter der Bedingung Z.

3.3.4 Weitere Anwendungsmöglichkeiten

Im folgenden sollen zwei weitere Anwendungsmöglichkeiten der gegenseitigen Information für Ontologien vorgestellt werden. Bei der ersten Anwendung wird oben beschriebenes Verfahren auf Pfade gleicher Länge zwischen X und Z angewandt. Damit ist es möglich, für jede Länge den jeweils informationstheoretisch optimalen Pfad zu bestimmen, d. h. denjenigen, der die meiste Information über X liefert. Anschließend kann der Graph so restrukturiert werden, dass der oder die informationstheoretisch optimalen Pfade zwischen X und Z auch die kürzesten sind. Die Y_i, welche auf dem Pfad zwischen X und Z liegen, repräsentieren somit die ausdrucksstärksten Knoten bezüglich X, weil sie über diesen die meiste Information liefern. Damit stellen diese Konzepte als die kürzeste Beschreibung von X unter den Bedingungen Y_i und Z dar, stellen also deren Kolmogorovkomplexität dar.

Eine zweite mögliche Anwendung wäre die Ähnlichkeit zwischen X und Z festzustellen, und damit auch das kleinste, gemeinsame Elternkonzept. Dazu wird obiges Verfahren insofern modifiziert, dass zwar weiterhin eine Menge $P_{X\to Z}$ aller Pfade zwischen X und Z betrachtet wird, aber aus dieser Menge derjenige Pfad mit dem höchsten Informationsgewinn ausgewählt wird:

$$d(X;Z,\mathbf{Y}) = H(X|Z) - \max_{i\in P_{X\to Z}} \{H_i(X|Y_1,\ldots,Y_n,Z)\}$$

Eines dieser $\mathbf{Y} = Y_1,\ldots,Y_n$ muss das kleinste gemeinsame Elternkonzept von X und Z sein, denn die $Y_1,\ldots,Y_n$ auf diesem Pfad liefern den höchsten Informationsgewinn für X unter der Bedingung Z.

Kapitel 4

Flutontologie

Im Internet ist heutzutage formales als auch informales Wissen in vielfältiger Art und Weise gespeichert. Darunter fallen beispielsweise statische und dynamische Webseiten, E-Mail, elektronische Nachrichtenbretter und Chatforen. Das so gespeicherte, auch wissenschaftliche, Wissen muss in geeigneter Weise strukturiert werden, um Erkenntnisse daraus gewinnen zu können. Erste Ansätze zur Wissensstrukturierung sind beispielsweise verschiedene Online Kataloge wie die Virtuelle Bibliothek[1] wissenschaftlicher Artikel oder ein hochspezialisiertes Portal für Umweltinformationen (GEIN) [Niedersächsisches 2003]. Eine weitere Herausforderung besteht darin, aus diesem verteilten und oft heterogenen Wissen, dasjenige zu finden, extrahieren und integrieren, welches die meiste Erkenntnis für die jeweilige Problemstellung liefert.

Zur Verbesserung der bisherigen, meist auf Schlüsselwörtern basierenden Suche nach Informationen, schlug Berners Lee, wohl einer der bekanntesten Gründer des WWW, das *Semantic Web* [Berners-Lee u. a. 2001] vor. In diesem soll der Kontext und, wenn möglich, die Semantik der heterogenen Informationsquellen berücksichtigt werden. Für die Definition eines Kontext und der sich daraus ableitenden Semantik werden Ontologien verwendet. Damit bilden Ontologien eine Grundlage für die Strukturierung von Wissen und dessen weitere Verarbeitung, beispielsweise eine Suche unter Berücksichtigung eines vorgegebenen Kontextes. Eine automatische Verarbeitung des Wissen durch autonome, intelligente Agenten, welche im Auftrag menschlicher Benutzer agieren, kann ebenfalls durch Verwendung von Ontologien erreicht werden. Die Ontologien sind hierbei meist auf eine spezifische Wissensdomäne festgelegt.

[1] http://vlib.org

4.1 Anwendungsgebiete

Im Rahmen des Graduiertenkollegs Naturkatastrophen wurde in Kooperation mit dem Institut für Wasserwirtschaft und Kulturtechnik (IWK) eine Ontologie für die Wissensdomäne des Risikomanagement im Falle eines Hochwassers erstellt [Kämpf u. a. 2003]. An diesem Gebiet sind die verschiedensten Wissensdomänen beteiligt, denn für das Risikomanagement von Hochwasser müssen natürliche als auch anthropogene Risiken betrachtet werden. Beteiligte Wissensdomänen sind beispielsweise Ingenieure, die verantwortlich für großangelegte, bauliche Maßnahmen wie Dämme oder Polder sind. Auch Chemiker und Biologen spielen eine wichtige Rolle, denn sie schätzen beispielsweise die Schäden aus eventuell austretenden, toxische Substanzen von Industrieanlagen oder Seuchengefahren ab.

Die Flutontologie bildet somit eine eindeutige Betrachtung der verschiedenen, beteiligten Wissensdomänen auf das Gebiet des Risikomanagements für Hochwässer. Damit ist eine einheitliche, semantische Grundlage für die Lokalisierung und Bereitstellung von domänenspezifischem Wissen für Wissenschaftler, welche in diesem Bereich tätig sind, geschaffen worden. Darauf aufbauend kann Wissen, welches durch die Ontologie dargestellt werden kann, strukturiert und weiter verarbeitet werden, um Erkenntnisse aus dem Wissen zu gewinnen. Einschränkend muss erwähnt werden, dass die Ontologie in ihrer jetzigen Form das Wissen von Experten aus dem Gebiet der Hydrologie und Meteorologie abdeckt. Durch die Ontologie werden beispielsweise Versicherungsapsekte sowie rechtliche Problematiken nicht modelliert. Weitere Einsatzmöglichkeiten der Flutontologie ergeben sich in verteilten, agentenbasierten Informationsgewinnungssystem als grundlegende Wissensrepräsentation für die zu gewinnenden Informationen, oder in Entscheidungsfindungssystemen als Wissensbasis für die Entscheidungsprozesse.

4.2 Entwicklungsprozess

Der Entwicklungsprozess der Flutontologie orientiert sich an [Uschold u. King 1995] (siehe Abb. 4.1). Es wurden jedoch, im Gegensatz zum vorgegebenen Verfahren, mehrere Iterationen dieses Prozesses durchgeführt, insbesondere bei der Erfassung der Konzepte und Relationen.

Zunächst wurden die Anforderungen an die Ontologie festgelegt. Wichtig hierbei war vor allem die Erweiterbarkeit der Ontologie. Indem Begriffe, die als Attribute von Konzepten hätten modelliert werden können, als separate Konzepte dargestellt wurden, konnte eine Erweiterbarkeit der Ontologie in dieser Hinsicht gewährleistet werden. Dies ermöglicht die Verwendung von *Upper Ontologies* für allgemeine Konzepte wie Zeit, Maßeinheiten und Raum. Desweiteren wurde die Verwendung von abstrakten philosophischen Konzepten vermieden, um eine Einordnung in verschiedene Upper Ontologies

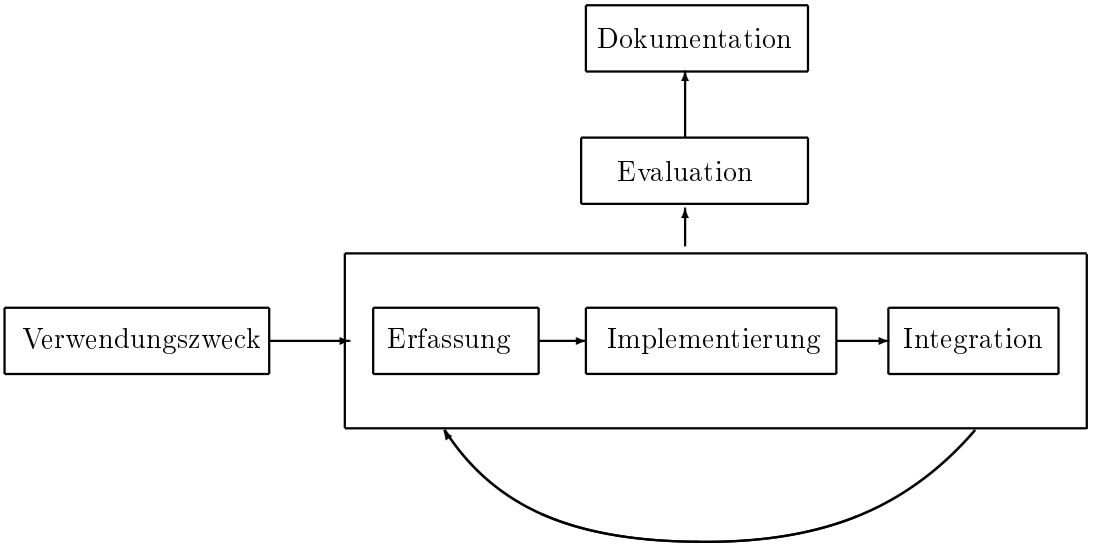


Abbildung 4.1: Entwicklungsprozess nach Uschold und King.

zu erleichtern.

Eine weitere Anforderung war die Aufteilbarkeit der Ontologie. Sie sollte so modelliert werden, dass es möglichst wenig Aufwand bereitet, einen Teil der Ontologie zu extrahieren. Diese Anforderung der Hydrologen wurde durch eine geschickte Gruppierung der Konzepte und mit der mit ihnen verbundenen Relationen erreicht, so dass im Falle einer Aufteilung wenig Nachbearbeitungsbedarf besteht.

Die eigentliche Entwicklung der Ontologie erfolgte dann iterativ durch Kommunikation mit den Experten. Dabei erwies sich in den ersten Gesprächen eine klassen- oder tabellenorientierte Modellierung der Ontologie als hilfreich, da diese Repräsentation auch für Laien auf diesem Gebiet leicht verständlich ist. Dabei wurde der *middle-out* Ansatz nach [Uschold u. King 1995] verfolgt. Es wurde bei der Modellierung der Ontologie nicht mit den generellsten oder spezifischsten Konzepten begonnen, sondern zunächst wurden wichtige Konzepte identifiziert, die etwa in der Mitte einer hierarchischen Gliederung liegen. Diese wurden dann in einem iterativen Prozess weiter spezialisiert. Anschließend konnten die gemeinsamen Elternkonzepte definiert werden, um somit eine konsistente Darstellung des Diskursbereiches der Ontologie zu erhalten.

Nachdem die Ontologie auf konzeptioneller Ebene, mittels graphischer und tabellarischer Darstellung festgelegt war, erfolgte die Implementierung in der Sprache DAML+OIL, welche auf Beschreibungslogik basiert. Die Wahl fiel auf diese Sprache, da sie auf XML basiert, und bereits über einige Werkzeuge zur Erstellung der Ontologie verfügt. Auch war abzusehen, dass der de-

signierte Nachfolger für XML-basierte Ontologiesprachen im *Semantic Web*, OWL, größtenteils die Konzepte von DAML+OIL übernimmt. Weiterhin existierten verschiedene Inferenzmaschinen, so dass die Ontologie als Wissensbasis für Inferenzsysteme dienen kann.

Nach der iterativen Erstellung der Ontologie war eine erste Anwendung die semantische Annotation der Webseiten des USGS Hydro-Climatic Data Networks (siehe Abschnitt 4.4). Die Ontologie erwies sich dafür mehr als ausreichend, so dass einige nichtbenötigte Teile abgespalten wurden. Dies gestaltete sich aufgrund der gegeben Anforderung der Aufteilbarkeit als nicht sehr schwierig.

4.3 Struktur der Flutontologie

Einen Überblick über die wichtigsten Konzepte der Flutontologie gibt Abb. 4.2 und Abb. 4.3. Die Domäne dieser Ontologie behandelt das Risikomanagement einer Flutkatastrophe, deshalb ist, abgekürzt *Flutmanagement (Floodmanagement)*, das zentrale Konzept, welches alle anderen Konzepte zueinander in Beziehung setzt. Das Flutmanagement muss die *Umgebung (Environment)* des Flusses betrachten, das heißt die geographische Struktur, in welcher der Fluss eingebettet ist. Desweiteren müssen so genannte *Felddaten (Fielddata)*, die als Parameter in Modelle für die Berechnung des Abflusses einfließen, berücksichtigt werden, wie beispielsweise meteorologische Daten als auch hydrologische und geophysikalische Daten. Der Abfluss ist ein elementares Konzept der Hydrologie, der es ermöglicht, Flutwellen eines Flusses vorherzusagen. Ein weiterer zentraler Bestandteil der Ontologie sind die *Schutzmaßnahmen (Protection)* vor Flutkatastrophen. Diese sind zum einen die langfristigen Präventivmaßnahmen und die kurzfristige Hilfe im akuten Katastrophenfall. Schließlich müssen noch die *administrativen (Administration)* Daten berücksichtigt werden, beispielsweise Identifikationsnummern von Pegelstationen. Ebenso werden verschiedene *Datentypen (Datatypes)* wie Zeitdauern, Längen, Einheiten und ähnliches in der Ontologie modelliert. Insgesamt besteht die Ontologie aus 213 Konzepte und 179 Relationen und Attributen. Die graphische Darstellung der in den folgenden Abschnitten näher beschriebenen Konzepte der Ontologie orientiert sich an Protégé (siehe Abschnitt 2.1.3).

4.3.1 Umgebung eines Flusses

Das Konzept der Umgebung (siehe Abb. Abb. 4.4) beinhaltet für diese Domäne natürlich den *Fluss (River channel)* als solchen, sein Einzugsgebiet und dessen Charakteristika sowie als Attribut den Namen des Flusses zur Identifizierung. Wichtige Relationen dieses Konzeptes sind:

- Tributary: verweist auf das Konzept der Flüsse, und stellt die Zuflüsse

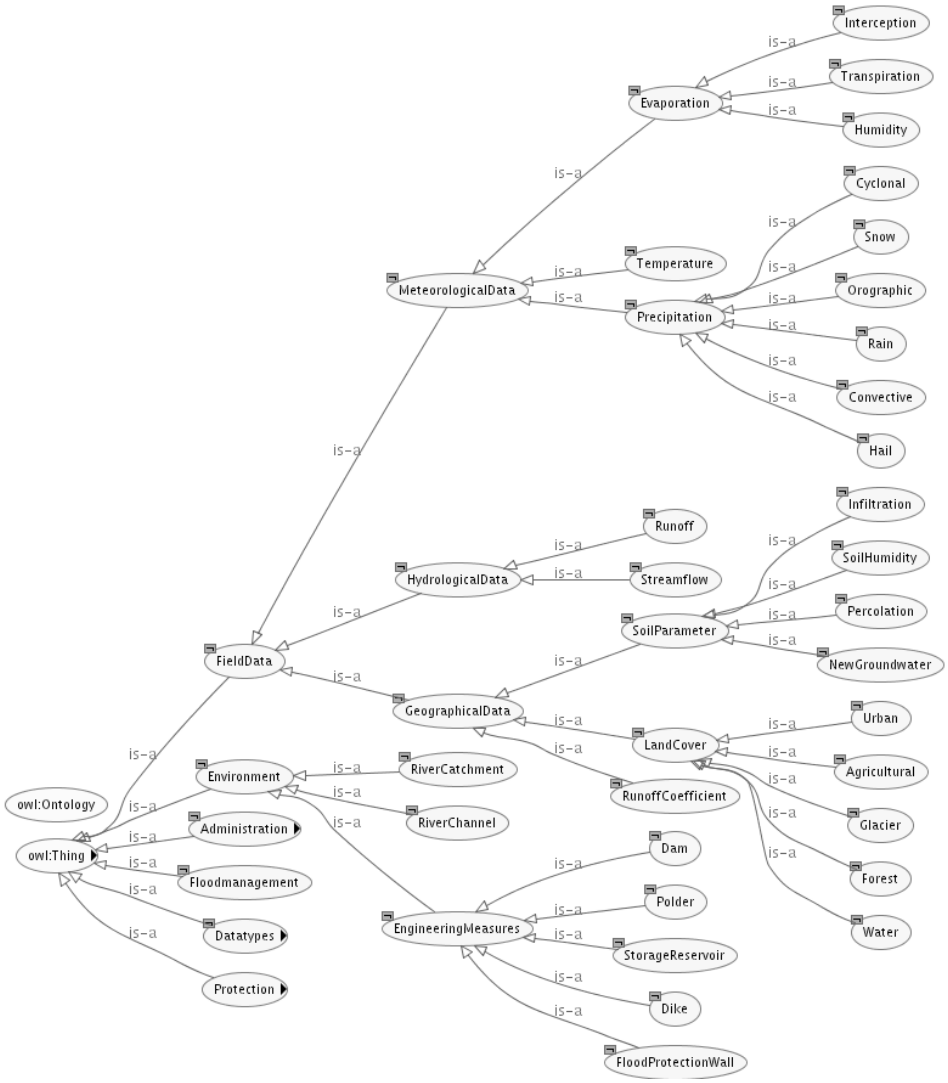


Abbildung 4.2: Visualisierung der Konzepte *Fielddata* und *Environment* der Flutontologie.

Abbildung 4.3: Visualisierung der Konzepte *Datatypes*, *Protection*, *Administration* und *Floodmanagement* der Flutontologie.

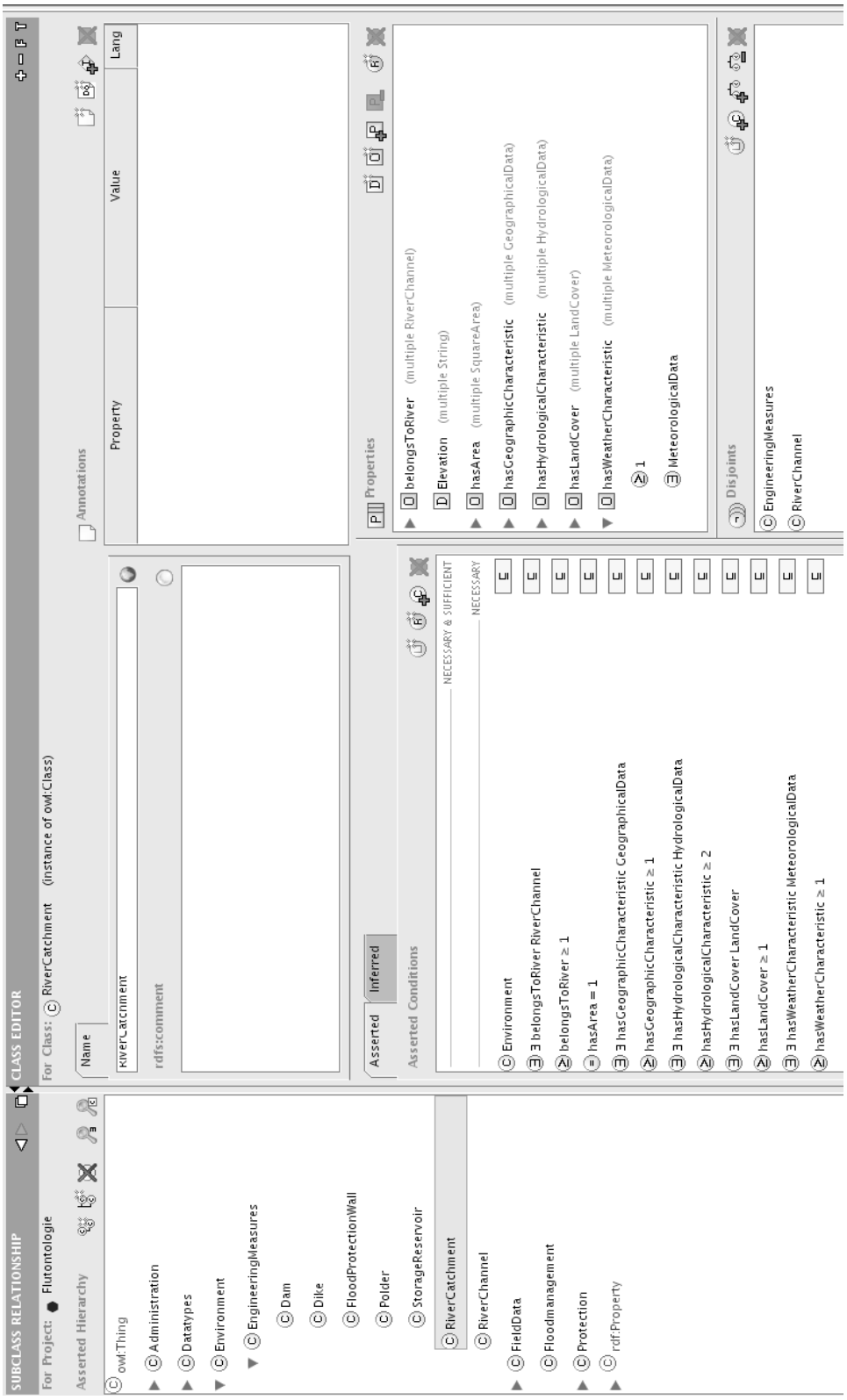


Abbildung 4.4: Modellierung der Umgebung. Gezeigt ist beispielhaft die Klasse *RiverCatchment*.

des betrachteten Flusses dar.

- Topological Characteristic: verweist auf die topologischen Daten des Flusses, wie Höhe über dem Meeresspiegel und Gefälle bzw. Steigung.

- Hydrological Characteristic: gibt die durchschnittliche Menge an Wasser an, welche pro Zeiteinheit durch den Fluss fließt. Dieser Wert wurde nicht als Attribut, sondern als Relation modelliert um die Erweiterbarkeit der Ontologie zu gewährleisten. Sowohl die Zeit- als auch die Maßeinheiten können unterschiedlich sein (z.B. englische und metrische Maße).

- Tributary_of: Falls der betrachtete Fluss Zufluss zu einem größeren Gewässer ist, zeigt das diese Relation an.

Als weiteres wichtiges Konzept wurde das *Einzugsgebiet (River catchment)* des Flusses modelliert, denn die Beschaffenheit eines Einzugsgebietes und der darauf abgehende Niederschlag beeinflussen den Wasserstand eines Flusses erheblich. Sollte zum Beispiel im Einzugsgebiet eine hohe Bodenversiegelung gegeben sein (Stadtgegend), wird der Niederschlag nahezu vollständig in den Fluss gelangen, während in einem Waldgebiet einiges durch den Boden als auch die Bäume zurückgehalten wird (Retention). Folgende Attribute und Relationen wurden für das Einzugsgebiet definiert:

- Area (Attribut): bestimmt die Größe des Einzugsgebietes.

- Elevation (Attribut): gibt die durchschnittliche Höhe des Einzugsgebietes über dem Meeresspiegel an

- Weather Characteristic (Relation): repräsentiert sämtliche meteorologischen Daten, die das Flussgebiet betreffen. Darunter fallen unter anderem durchschnittlicher Niederschlag pro Zeiteinheit, Art des Niederschlages (Schnee, Regen, Hagel) und Verdunstung (d.h. ein Teil wird nicht in den Fluss transportiert).

- Geographic Characteristic (Relation): verweist auf die Beschaffenheit des Bodens und seine Nutzung. Der Eintrag des Niederschlages in den Fluss hängt zum einen von der Beschaffenheit des Bodens ab. Je nach Durchlässigkeit wird ein Teil des Niederschlages direkt dem Grundwasser zugeführt oder eine Zeit lang im Boden zurückgehalten, so dass sich der Eintrag des Niederschlages in den Fluss über einen längeren Zeitraum hinzieht. Zum anderen muss der relative Anteil der jeweiligen Flächenarten (Waldboden, Versiegelte Fläche, Ackerland) berücksichtigt werden, um letztendlich die korrekte Menge des dem Fluss zugeführten Niederschlages berechnen zu können.

- Hydrological Characteristic (Relation): hängt stark mit den vorgehenden zusammen, denn sie beschreibt die tatsächliche Menge an Niederschlag, die in den Fluss eingebracht wird. Hier sind noch einige zusätzliche Faktoren, wie die Zirkulation des Wassers im Boden, sowie die Fließgeschwindigkeit des Grundwassers, berücksichtigt.

Als letztes Konzept, welches die Umgebung des Flusses beeinflußt, sind *wasserbauliche Maßnahmen (engineering measures)* aufgeführt. Unter diesem Konzept vereinen sich beispielsweise Dämme, Polder, Stauseen, Hochwasserschutzmauern und Rückhaltebecken. Diese Konzepte stehen in enger Beziehung zu den im Management aufgeführten langfristigen Präventionsmaßnahmen.

4.3.2 Felddaten

Unter dem Konzept der Felddaten sind die wichtigsten zu erfassenden Kenngrößen natürlicher Ereignisse zusammengefasst (siehe Abb. 4.5). Diese dienen unter anderem dazu, ein geeignetes, wasserbauliches oder meteorologisches Modell mit Daten zu versorgen und entsprechende Vorhersagen über den Wasserstand des Flusses bzw. der Wettersituation im Einzugsgebiet zu geben. Die drei zentralen Konzepte sind die meteorologischen, hydrologischen und geographischen Daten.

Meteorologische Kenndaten

Das Konzept der meteorologischen Kenndaten vereint die wichtigsten, für Hochwässer relevanten Kenndaten, aus dem Gebiet der Meteorologie. Dies sind die Niederschlagsmenge und -art (Regen, Schnee, Hagel) sowie meteorologische Art des Niederschlages (Orographisch, Zyklonal, Konvektiv). Die Niederschlagsmenge kann dabei in verschiedenen Maßeinheiten (Mittlerer Jahresniederschlag, sowie maximaler Tagesniederschlag) dargestellt werden kann. Eine weitere Kennzahl ist die Temperatur, die ebenfalls unterschiedlich repräsentiert werden kann, und zwar als mittlere Jahrestemperatur oder als Mittel der Temperaturen im Januar und August angegeben werden. Als letztes Konzept der meteorologischen Kenndaten wurde die Verdunstung modelliert. Sie setzt sich zusammen aus der Luftfeuchtigkeit, der Interzeption (Zurückhalten des Niederschlages auf Pflanzenoberflächen) und der Transpiration (Verdunstung von Wasser auf Blattoberflächen). Um die Erweiterbarkeit der Ontologie zu gewährleisten, wurden diese Begriffe nicht als Attribute sondern als Konzepte modelliert, da sie, wie gezeigt, in unterschiedlicher Art und Weise repräsentiert werden können.

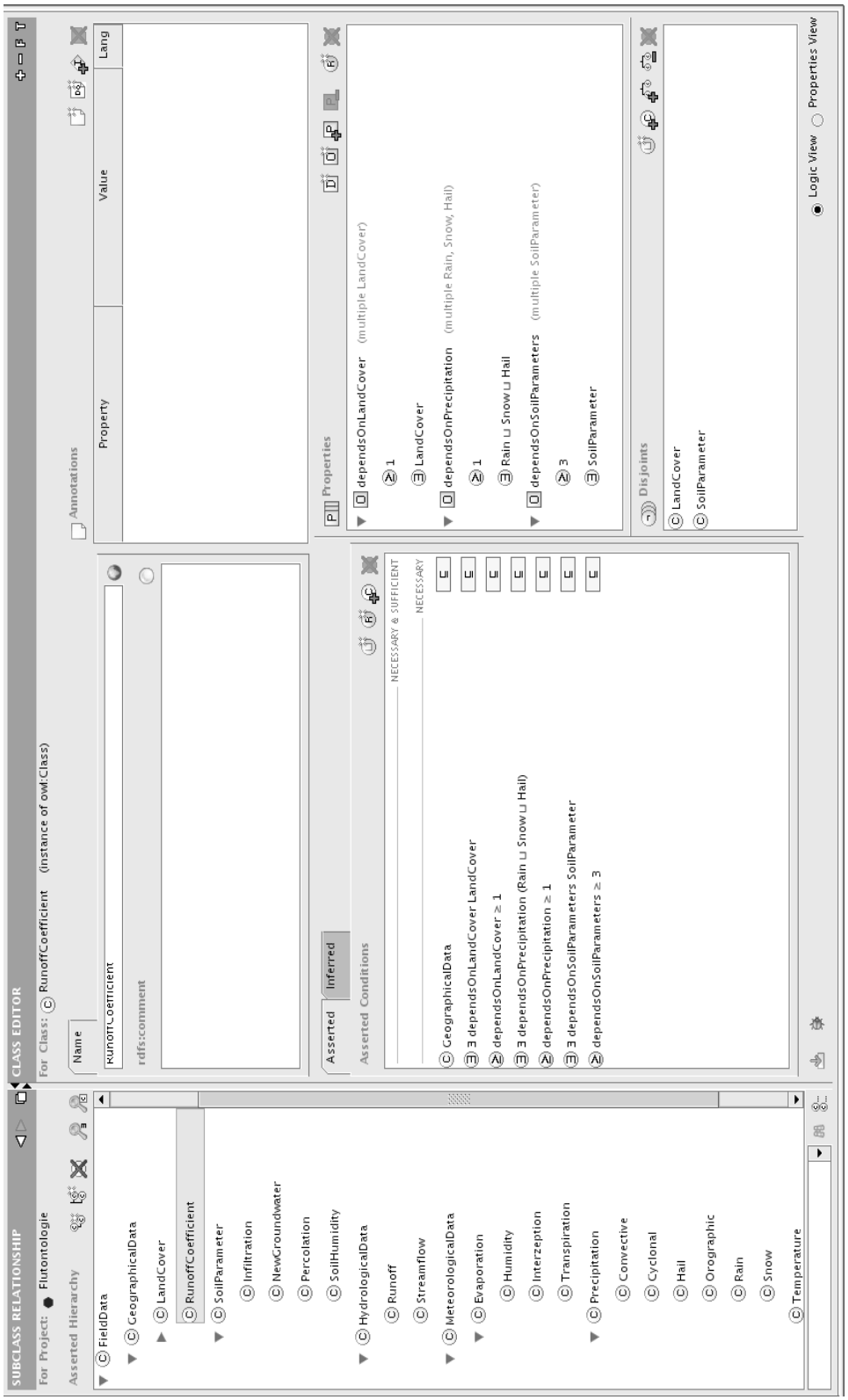


Abbildung 4.5: Modellierung der Felddaten. Gezeigt ist beispielhaft die Klasse *RunoffCoefficient*.

Geographische Kenndaten

Die geographischen Kenndaten repräsentieren den Einfluss des Bodens auf die in den Fluss eingebrachte Niederschlagsmenge. Ein wichtiges Konzept sind die Bodenparameter, welche sich weiter aufteilen in den Infiltrationsindex, Bodenfeuchtigkeit, Perkolation und neu hinzugekommenes Grundwasser.

Der Infiltrationsindex gibt die Durchlässigkeit des Bodens für Wasser an und berücksichtigt dabei die Bewegung des Wassers im Boden als auch den Sättigungsgrad, der je nach Boden und Bewuchs unterschiedlich ist. Die Bodenfeuchte spiegelt hier nicht den absoluten Wassergehalt des Bodens wieder, sondern gibt den Anteil des Wassers in den verschiedenen Ebenen des Bodens an, wie zum Beispiel den für Pflanzen verfügbaren Anteil. Die Perkolation gibt den nach unten gerichteten Bewegungsvorgang des Sickerwassers wieder. Für nähere Informationen zu diesem Thema siehe [Scheffer u. Schachtschnabel 2002]. Alle Begriffe wurden als Konzept modelliert, da sie in unterschiedlichen Einheiten repräsentiert werden können.

Weiterhin wurde die Landnutzung als Konzept aufgenommen, um die anteilige Fläche der einzelnen Nutzungsarten (Wald, Seen, Gletscher, Flüsse, Landwirtschaftliche Fläche, urban genutzte Flache und Verkehrswege) an der Gesamtfläche des Einzugsgebietes darstellen zu können. Die Nutzung der Flächen steht in direktem Zusammenhang mit den oben beschriebenen Parametern Infiltrationsindex und Bodenfeuchte.

Der Abflusskoeffizient ist schließlich der Quotient aus nicht versickertem Niederschlag, also der direkt dem Fluss zugeführte Anteil des Niederschlags, und dem insgesamt gefallenem Niederschlag. In die Berechnung des Abflusskoeffizienten fließen alle bisher besprochenen geographischen Kenndaten ein.

Hydrologische Kenndaten

Die hydrologischen Daten stellen das Konzept des *Abflusses (runoff)* in den Mittelpunkt, da anhand dessen die wesentlichen Merkmale einer (drohenden) Flutwelle ermittelt werden können. Der Abfluss setzt sich aus den folgenden Konzepten zusammen:

- Infiltration: Bezeichnet den Zugang von Wasser durch enge Hohlräume in die Lithosphäre. Dieser Begriff steht in direkter Relation zu den Bodenparametern, welche unter dem Konzept der geographischen Daten zusammengefasst sind.

- Perkolation: Bezeichnet den Durchfluss von Grundwasser durch ein festes Substrat.

- Retention: Bezeichnet die Fähigkeit, Niederschlag verzögert abzugeben. Beispielsweise fließt versickerter Niederschlag in der Humusschicht

schneller als Wasser, welches in tiefere, dichtere Bodenschichten vordringt. Je mehr Wasser so indirekt zurückgehalten wird, desto geringer und vor allem langsamer steigen die Pegel der im Einzugsgebiet liegenden Flüsse.

- Oberflächenabfluss: Bezeichnet den Teil des Niederschlages, der nicht versickert, sondern direkt in den Fluss geleitet wird. Da der Eintrag in den Fluss meistens unmittelbar nach dem Niederschlagsereignis stattfindet, steigen bei einem hohen Oberflächenabfluss die Pegel der Flüsse recht schnell an.

Desweiteren existiert noch das Konzept des Durchflusses, welches die mittlere Wassermenge pro Zeiteinheit pro Volumen an einer Pegelstation angibt. Dazu steht dieses Konzept in Relation zu der Messstation an sich sowie zu dem Einzugsgebiet des Flusses, da der Durchfluss direkt von der Landnutzung in eben diesem Einzugsgebiet abhängt.

4.3.3 Schutzmaßnahmen

Unter dem Konzept Schutzmaßnahmen (siehe Abb. 4.6) sind die verschiedenen Managementmaßnahmen hinsichtlich des Risikomanagements für Flutkatastrophen zusammengefasst. Dabei wird angenommen, dass die im folgenden beschriebenen Maßnahmen unter dem Aspekt der Nachhaltigkeit geplant und durchgeführt werden. Im Katastrophenmanagement unterscheidet man nach [Plate u. Merz 2001] zwischen den *langfristigen* Maßnahmen im Rahmen des Umweltmanagements, und den *kurzfristigen* Maßnahmen, welche in einer akuten Notlage angewandt werden.

Umweltmanagement

Unter dem Konzept des Umweltmanagements sind die langfristigen Schutzmaßnahmen zusammengefasst, welche zur Minderung der Folgen einer Flut dienen. Dabei wird zwischen strukturellen und nichtstrukturellen Maßnahmen unterschieden.

Unter den strukturellen Maßnahmen versteht man größtenteils hydrologische Baumaßnahmen und deren Planung. Aus diesem Grund steht dieses Konzept auch in enger Beziehung zu den technischen Baumaßnahmen, welche als *Baumaßnahmen* in Abschnitt 4.3.1 aufgeführt sind.

Mit den nichtstrukturellen Maßnahmen versucht man durch organisatorische und administrative Maßnahmen die Folgen einer Flutkatastrophe zu mindern. In erster Linie seien hier Flutpläne genannt. Flutpläne zeigen an, welche Gebiete entlang eines Flusslaufes im Falle eines Hochwassers überflutet werden. Dabei werden die risikogefährdeten Gebiete meistens nach der statistischen Häufigkeit eines Hochwassers eingeteilt, also wie hoch die Wahrscheinlichkeit einer Überflutung bei einem 50-, 100-, und 500-jährigem

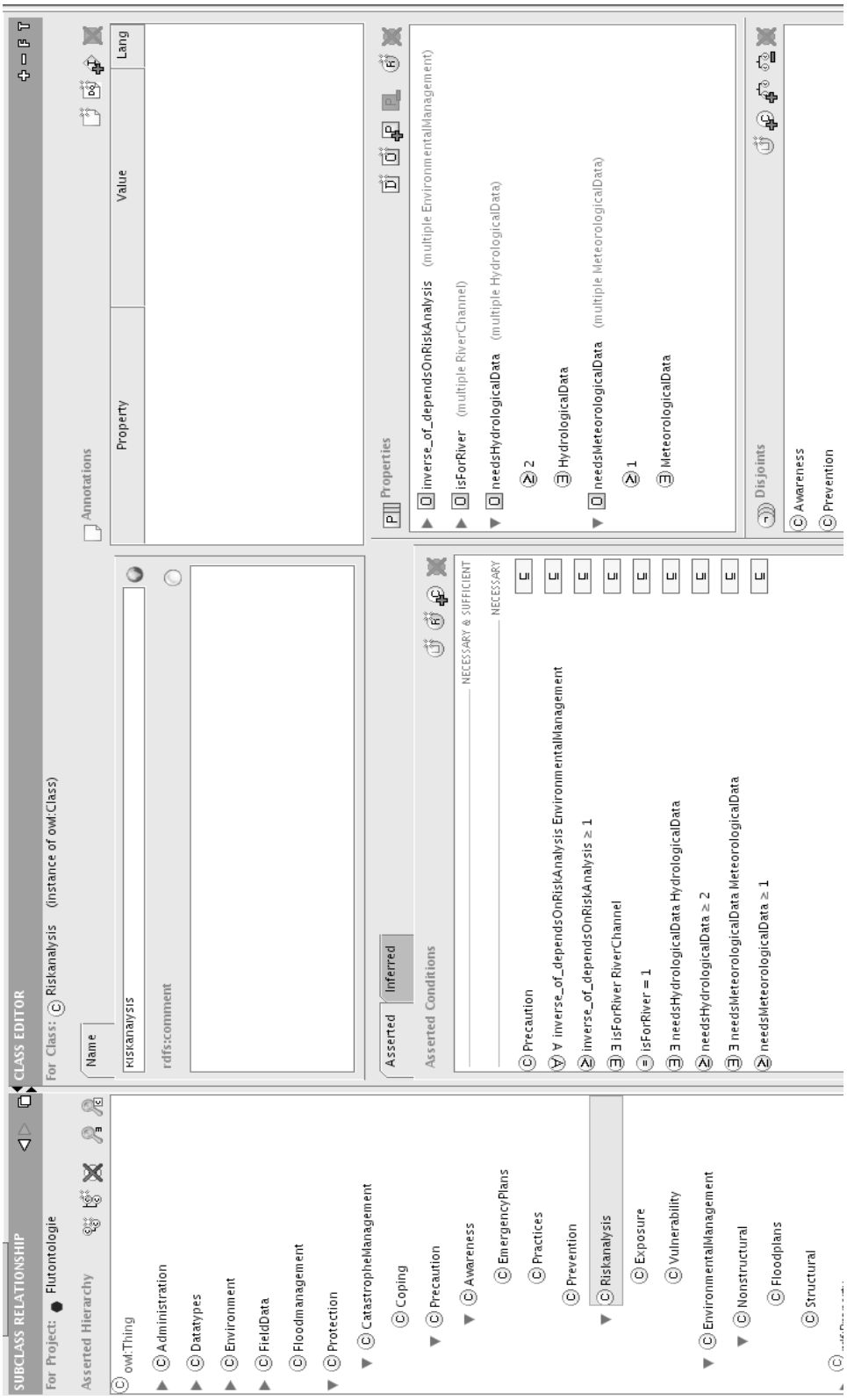


Abbildung 4.6: Modellierung der Schutzmaßnahmen. Gezeigt ist beispielhaft die Klasse *Riskanalysis*.

Hochwasser ist. Auch dieses Konzept hat enge Beziehungen zu dem Konzept der Flussumgebung als auch zu den Felddaten, da diese für die korrekte Berechnung der Flutpläne nötig sind.

Katastrophenmanagement

Das kurzfristige Katastrophenmanagement unterteilt man in Vorsorge und Bewältigung. Unter dem Konzept der Bewältigung sind schließlich alle Maßnahmen aufgeführt, die während eines akuten Hochwassers und kurze Zeit danach relevant sind. Dies umfasst zum Beispiel die Koordination der Rettungsdienste, wobei hier auf die Beziehung Bereitschaftserhöhung zurückgegriffen wird, sowie die Handhabung der unmittelbare Situation nach der Flut. Dies wären beispielsweise Räumungsaktionen, um die Verkehrswege von Schlamm und Geröll zu befreien.

Die Vorsorge besteht aus der Risikoanalyse, der Vorbeugung und Bereitschafterhöhung. Die Risikoanalyse befasst sich vor allem mit der Berechnung der Gefährdung und der Vulnerabilität von Objekten. Entsprechendes bei Menschen durchzuführen ist ethisch nicht unproblematisch und wird deshalb nicht näher betrachtet. Die Gefährdung gibt die Wahrscheinlichkeit an, mit der ein Objekt bei einem 100-jährigen Hochwasser überflutet wird. Bei der Vulnerabilität spielen noch sozioökonomischen Faktoren des Objektes eine Rolle; z.B. hat eine Lagerhalle für Stahl eine geringere Vulnerabilität als ein Regierungsgebäude oder eine Chemiefabrik, da die Folgen eines Ausfalles hier schwerwiegender sind.

Die Vorbeugung steht in Relation zu den Konzepten des Umweltmanagements, weil dort die entsprechenden strukturellen und nichtstrukturellen Maßnahmen zur Minderung der Flutfolgen aufgeführt sind.

Die Bereitschaftserhöhung gehört ebenfalls zur Vorsorge, denn unter ihr versteht man zum Beispiel das Durchführen von Übungen der einzelnen Rettungsdienste und ähnliches.

4.3.4 Administratives

Unter dem Konzept der Administration sind administrative Notwendigkeiten zu den bereits vorgestellten Konzepten untergebracht (siehe Abb. 4.7). Eine große konzeptionelle Einheit bilden die administrativen räumlichen Einheiten. Darunter ist die hierarchische Aufteilung der Einzugsgebiete der Flüsse in Länder (USA, Deutschland, etc.), Gemeinde, Städte und ähnliches zu verstehen. Diese Konzepte stehen normalerweise immer in Relation zu den Einzugsgebieten der Flüsse, und zeigen so deren geographische Lokation an. Diese geographischen Daten wurden aus Erweiterungsgründen als externe Konzepte modelliert, da somit einfach zusätzliche Information zu diesen Konzepten hinzugefügt werden kann. Das kann zum Beispiel die Bevölkerungszahl eines Landes sein, aufgrund derer man dann Inferenzen bezüglich der

Landnutzung durchführen könnte. Desweiteren müssen bei der Angabe der geographischen Daten nicht alle Angaben vorhanden sein, so dass in diesem Falle keine Relation zu dem entsprechenden Konzept besteht. Wenn die geographischen Daten als Attribut modelliert wären, müsste man für diesen Fall einen speziellen Wert finden und diesen Wert beim automatischen Schließen auch immer überprüfen. Als weitere geographische Daten wurden wasserschutzrechtliche Einheiten, wie Wasserschutzgebiete und hydrologische Einheiten modelliert.

Ferner wurden so genannte Datenarchive eingeführt. Sie modellieren in der Ontologie vor allem die Pegelmessstationen. Modelliert wurden bei solch einer Station zum einen administrative Gegebenheiten, wie eindeutige Identifikationsnummer, ihrer geographischen Position und die verantwortlichen Personen für die Wartung der Station. Zum anderen wurde die wichtigste hydrologische Einheit, nämlich der Abfluss, den die Station misst, als Relation zu dem Konzept Abfluss modelliert. Auch wurde es ermöglicht historische Abflussdaten, welche durch eine diskrete Zeitreihe gegeben sind, zu modellieren.

Zu den Datenarchiven gehören auch die meteorologischen Stationen, die in ähnlicher Weise wie die Pegelmessstationen modelliert sind. Ebenfalls modelliert sind Metadaten, welche es ermöglichen, die historischen Zeitreihen der Messstationen anzugeben.

4.3.5 Datentypen

Unter dem Konzept Datentypen (siehe Abb. 4.8) sind verschiedene Einheiten und Datentypen zusammengefasst, die von anderen Konzepten und Relationen benutzt werden können. Das Konzept Zeit dient vor allem dazu, die unterschiedlichen Modellierungen der Zeit, z.B. kontinuierliche und diskrete Zeitreihen, für alle anderen Konzepte konsistent zu modellieren. Dieses Konzept dient auch zur Erfassung der unterschiedlichen Modellierungen bezüglich des Zeitrahmens einer Messreihe. Desweiteren ist das Zeitkonzept auch für die Modellierung der unterschiedlichen zeitlichen Einheiten wie Stunde, Jahr, Jahrzehnt, zuständig und könnte in einer überarbeiteten Fassung durch entsprechende Konzepte einer Upper Ontology ersetzt werden. Das Konzept der Frequenz ist auch unter dem Konzept Zeit untergebracht worden, weil es die unterschiedlichen Möglichkeiten zur Darstellung der Häufigkeiten von Ereignissen modelliert.

Die Datentypen repräsentieren auch grundlegende geographische und metrische Maße, wie beispielsweise Exposition, welche die vier Himmelsrichtungen darstellt und für Längen- und Breitenangabe in einigen metrischen Maßen benötigt wird. An metrischen Maßen existieren die Höhe über dem Meeresspiegel, Neigung und geographische Länge bzw. Größe eines Gebietes. Auch hier können unterschiedliche Einheiten benutzt werden und zukünftig durch standardisierte Konzepte aus einer Upper Ontology ersetzt werden.

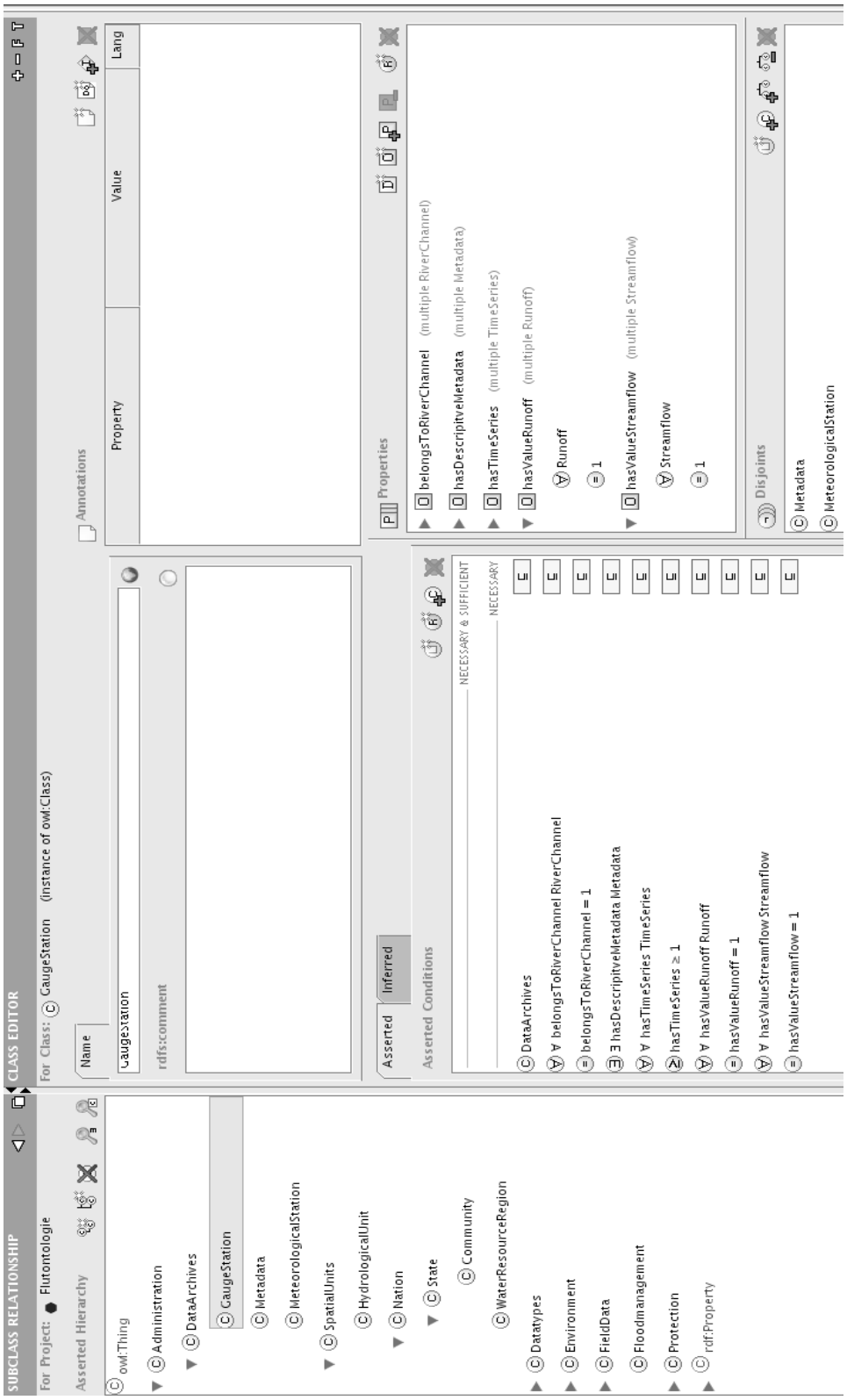


Abbildung 4.7: Modellierung der administrativen Daten. Gezeigt ist beispielhaft die Klasse *Gaugestation*.

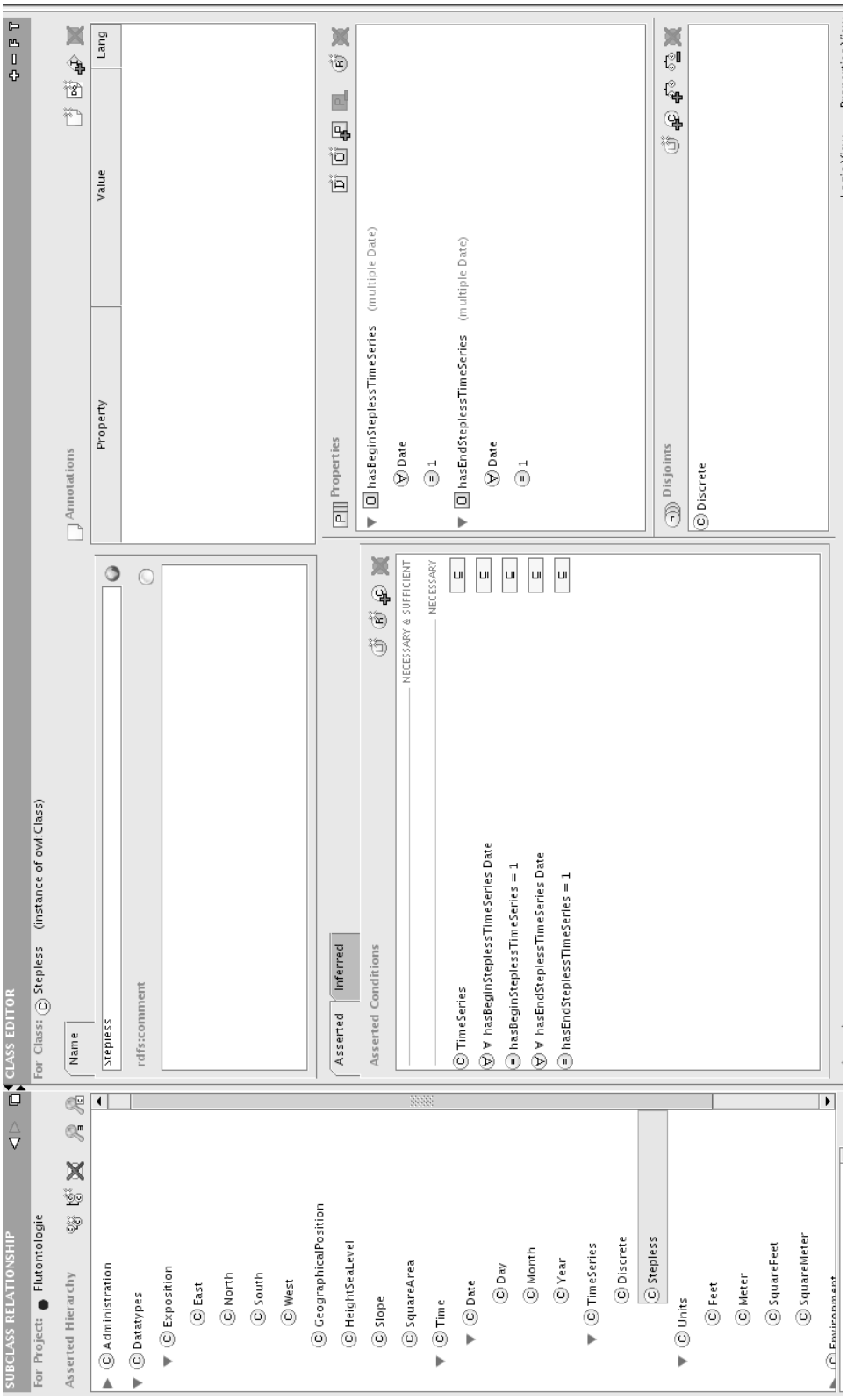


Abbildung 4.8: Modellierung der Datentypen. Gezeigt ist beispielhaft die Klasse *Stepless*.

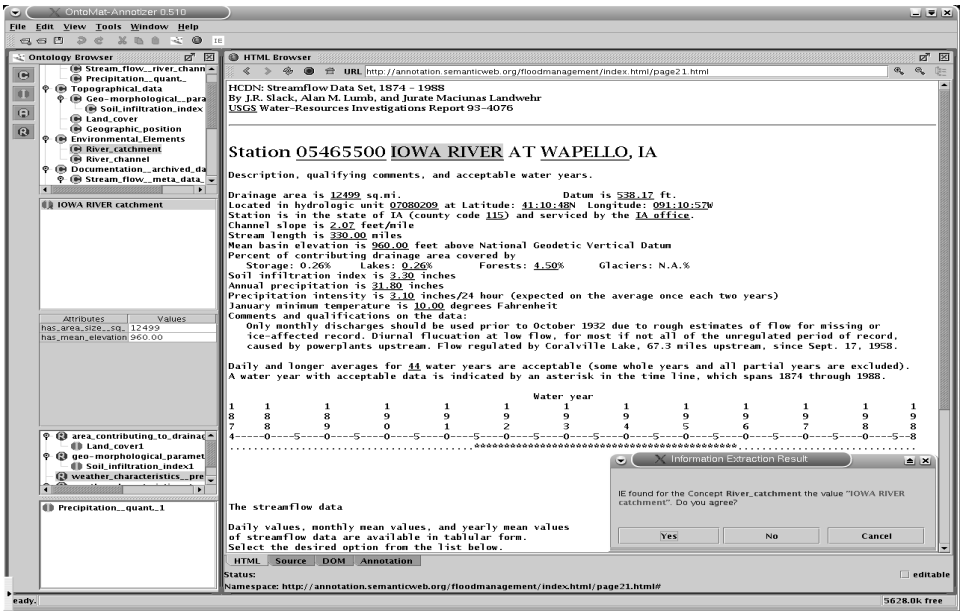


Abbildung 4.9: Webseite des USGS.

4.4 Annotationstest

Um die Einsatzmöglichkeiten der Ontologie zu testen, wurde eine Annotation der Webseiten des USGS Hydro-Climatic Data Network (HCDN)[2] vorgenommen. Bei der Annotation wurden die auf den Webseiten befindlichen Daten mit der Semantik der Ontologie versehen. Die Annotation der Webseiten erfolgte mit einem vom AIFB bereitgestelltem Werkzeug, dem OntoMat-Annotizer[3] [Handschuh u. a. 2003].

Auf diesen Webseiten sind sehr strukturierte Informationen über die Abflussdaten und sämtlicher zugehörigen, relevanten Parameter enthalten (siehe Abb. 4.9). Zur Annotation wurden das Wissen auf der Webseite manuell mit der Semantik der Ontologie versehen, d. h. den Informationen auf der Webseite wurde Meta-Information mittels der Ontologie hinzugefügt. Beispielsweise wurde der Flussname IOWA RIVER dem Attribut *Flussname* der Instanz *Fluss* zugeordnet. Damit hat er eine eindeutige semantische Bedeutung erhalten, weil das Konzept *Fluss* in der Ontologie durch seine Relationen eindeutig bestimmt ist. Die restlichen Daten der Webseite wurden per Hand mittels des Tools OntoMat-Annotizer den übrigen Konzepten und Attributen der Ontologie zugeordnet (siehe Abb. 4.10).

Die Relationen zwischen den Konzepten wurden dann ebenfalls mit Hilfe des beschriebenen Werkzeuges hergestellt. Beispielsweise wurden Relationen

[2] http://water.usgs.gov/pubs/wri/wri934076/1st_page.html
[3] http://annotation.semanticweb.org/ontomat/index.html

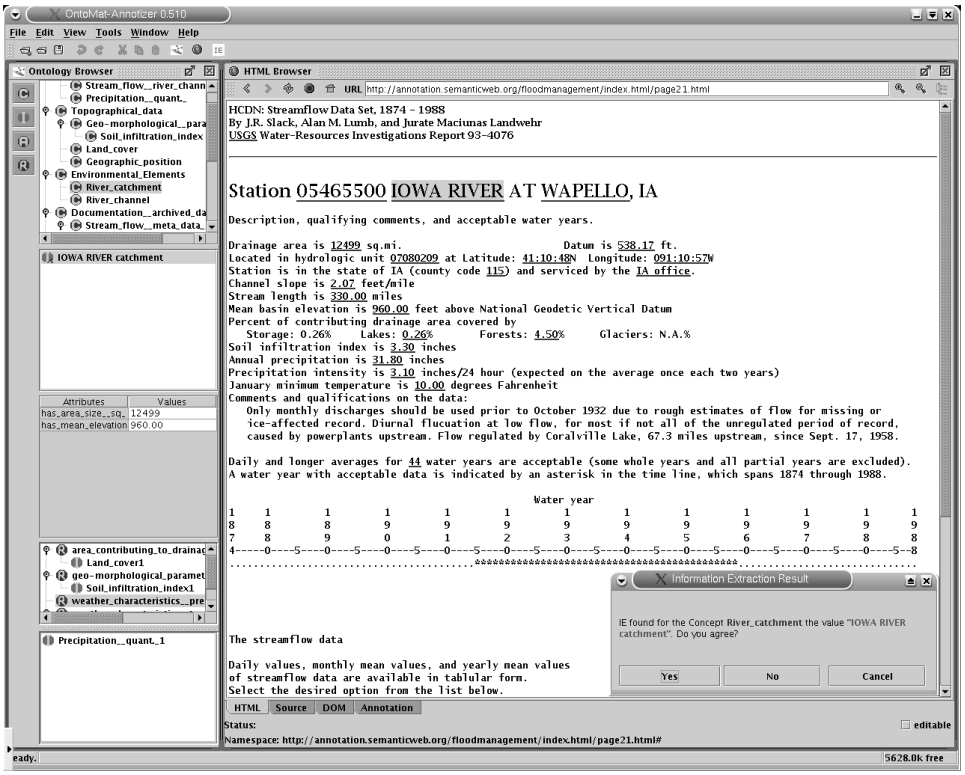


Abbildung 4.10: Annotierte Webseite des USGS.

zwischen den einzelnen *Landbedeckungen* (Storage, Lakes, Forests, Glaciers) und dem Konzept *Abfluss* hergestellt. Um dieses Verfahren in großem Stil anwenden zu können, muss die Annotation automatisch oder zumindest semiautomatisch erfolgen [Erdmann u. a. 2000]. Für eine Annotation der Webseiten des USGS erwies sich die Flutontologie hierbei mehr als ausreichend.

Kapitel 5

Relative Entropie für Ontologien

Im vorigen Kapitel wurde eine Ontologie für das Risikomanagement von Flutkatastrophen vorgestellt. Damit ist es möglich, Ursachen einer Flut zu identifizieren und entsprechende Gegenmaßnahmen zu erkennen. Nun besteht aber, trotz semantischer Eindeutigkeit das Problem, dass verschiedene Personenkreise mit unterschiedlichem Wissensstand eine Ontologie und den durch sie beschriebenen Diskursbereich unterschiedlich bewerten können. Beispielsweise könnten Politiker andere Schutzmaßnahmen als Hydrologen für wichtig erachten. Zur Feststellung solcher Unterschiede, und der damit einhergehenden Strukturierung des durch die Ontologie repräsentierten Wissens, wird in diesem Kapitel vorgeschlagen, die relative Entropie zu benutzen [Daemi u. Calmet 2004a].

Die relative Entropie wurde aus mehreren Gründen für die Strukturierung von Ontologien bzw. des durch sie modellierten Wissen ausgewählt. Ein Grund besteht darin, dass die relative Entropie den informationstheoretischen Abstand zwischen Datenkompressionsmodellen, dargestellt durch die Wahrscheinlichkeitsverteilungen $\mathbf{p}$ und $\mathbf{q}$, widergibt, wobei die statistischen Eigenschaften der Wahrscheinlichkeitsverteilungen berücksichtigt werden [Cover u. Thomas 1991]. Der Abstand ist genau dann Null, wenn beide Modelle exakt gleich sind. Ansonsten ist der Abstand größer Null und wird umso größer, je unterschiedlicher die Modelle zueinander sind. Die relative Entropie gibt somit die durchschnittliche Information an, dass ein Ereignis aus einem Wahrscheinlichkeitsmodell $\mathbf{p}$ für das Wissen nicht aus einem Modell $\mathbf{q}$ stammt. Damit lässt sich die relative Entropie auch mit der in Abschnitt 3.2 gegebenen Definition von Wissen als semantischer Information verbinden, welche als Möglichkeitsausschluss im Wissensraum definiert wurde. Weiterhin ist die relative Entropie additiv, d.h. für zwei unabhängige Wahrscheinlichkeiten ist die Distanz der Verbundverteilungen gleich der Summe der jeweiligen Randverteilungen. Eine weitere nützliche Eigen-

schaft für die Strukturierung von Ontologien ist das *Data Processing Theorem (DPT)*. Es besagt, dass keine statistische Verarbeitung des Wissen die relative Entropie erhöht [Kullback u. Leibler 1951]. Dies beinhaltet Operationen wie zum Beispiel Mittelung, Gruppierung oder Verdichtung des Wissen. Ein letzter Grund für die Verwendung der relativen Entropie ist ihre schnelle und einfache Berechenbarkeit gegenüber anderen, informationstheoretischen Distanzmaßen.

5.1 Voraussetzungen

Die zur Berechnung der relativen Entropie benötigten Wahrscheinlichkeitsverteilungen, welche auf der Ontologie in geeigneter Weise definiert sind, müssen lediglich die Kolmogorov Axiome erfüllen. Diese Axiome lauten:

- Nichtnegativität: $\forall A \in \mho \ : \ P(A) \geq 0$
- Normierung: $P(\Upsilon) = 1$
- Additivität: $\forall A_i, A_j \in \mho$ mit $A_i \cap A_j = \emptyset$ für $i \neq j \ : P(\sum_i) A_i = \sum_i P(A_i)$

wobei $\mho$ eine σ-Algebra von Teilmengen $A \subseteq \Upsilon$ ist. Für die konkrete Bedeutung der Wahrscheinlichkeiten bedeutet dies, dass neben objektiven, frequenzbasierten Wahrscheinlichkeiten auch beliebige subjektive Wahrscheinlichkeiten zugelassen sind, welche eine rein kognitive Bedeutung besitzen [Klir u. Wierman 1998]. Die Wahrscheinlichkeiten berechnen sich hierbei aus dem der Ontologie zugrundeliegenden Wissen, welches strukturiert werden soll. Das können, für den Fall der Flutontologie, beispielsweise Fragebögen sein, welche den unterschiedlichen Personengruppen zu dem Diskursbereich der Ontologie gegeben wurden, oder Dokumente, welche mittels einer Häufigkeitsanalyse auf in der Ontologie vorkommende Begriffe hin untersucht wurden.

Die berechneten Wahrscheinlichkeiten werden anschließend den Kanten zwischen den Konzepten, d.h. den Relationen zwischen ihnen, zugewiesen. Die Zuweisung der Wahrscheinlichkeitsverteilungen und der zugehörigen Wahrscheinlichkeiten zu den Kanten in einer Ontologie kann in verschiedenster Art und Weise erfolgen, hängt allerdings primär von der Bedeutung der Wahrscheinlichkeitsverteilung ab. In unserem Beispielfall für die Flutontologie wurde pro Ebene der Ontologie eine Wahrscheinlichkeitsverteilung definiert. In jeder Ebene der Ontologie werden bestimmte Aspekte des Hochwasserschutzes betrachtet, welche eine Spezialisierung der darüberstehenden Ebene sind. Die Definition einer Wahrscheinlichkeitsverteilung pro Ebene bedeutet somit, die *Wichtigkeit* der Konzepte für jeden Spezialisierungsgrad der Ontologie umfassend festzulegen.

Für die Anwendung der relativen Entropie als Strukturierungsmaß für Ontologien sind jedoch beliebige Zuweisungen der Wahrscheinlichkeiten zu den Kanten einer Ontologie möglich, solange Wahrscheinlichkeitsverteilungen nach den oben genannten Kolmogorov Axiomen gegeben sind. Beispielsweise kann für die gesamte Ontologie oder sämtliche Kanten eines Konzeptes eine Wahrscheinlichkeitsverteilung bestimmt werden. Die Art und Weise, wie die Wahrscheinlichkeiten den Kanten zu gewiesen werden, ist auch von dem zu strukturierenden Wissen abhängig, dass durch die Ontologie dargestellt wird. Die Wahrscheinlichkeitsverteilungen stellen ja in gewisser Hinsicht eine zusammenfassende Modellierung des Wissens dar.

5.2 Anwendung

Nachdem die Wahrscheinlichkeitsverteilungen den Kanten einer Ontologie zugewiesen wurden, kann mittels der relativen Entropie

$$D(\mathbf{p}||\mathbf{q}) = \sum_{x \in \Upsilon} p(x) \log \frac{p(x)}{q(x)}$$

die Distanz zwischen Ontologien bzw. des durch sie modellierten Wissens bestimmt, und damit auch eine Strukturierung durchgeführt werden. Die Wahrscheinlichkeitsverteilungen

$$\mathbf{p} = p(x_1), \ldots, p(x_n), \quad n = |\Upsilon|$$

und

$$\mathbf{q} = q(x_1), \ldots, q(x_n), \quad n = |\Upsilon|$$

ergeben sich hierbei jeweils aus den miteinander zu vergleichenden Ontologien bzw. deren Wissensbasen. Das zugrundeliegende Wissen kann beispielsweise das implizite Wissen einer Person sein, welche aufgrund ihrer Erfahrung den Konzepten bestimmte Wichtigkeiten bezüglich des Hochwasserschutzes zuweist. Das wäre eine Beispiel für eine rein subjektive Zuweisung der Wahrscheinlichkeiten. Eine objektive Wahrscheinlichkeitsverteilung würde sich ergeben, wenn die Wichtigkeiten anhand von Fragebögen oder ähnlichem bestimmt werden.

Von der Art der Zuweisung der Wahrscheinlichkeitsverteilung hängt die Berechnung der Gesamtdistanz zwischen den Ontologien ab. Falls eine globale Wahrscheinlichkeitsverteilung auf der Ontologie definiert wurde, müssen keine weiteren Berechnungen erfolgen. Wenn dagegen die Wahrscheinlichkeitsverteilungen pro Knoten (bzw. von ihm ausgehenden Kanten) festgelegt wurden, existieren viele verschiedene Möglichkeiten, die Gesamtdistanz zu berechnen. Beispielsweise könnte man die Distanzen aufsummieren und dabei zusätzliche Gewichtungen, anhand des Grades der Knoten oder der Ebene, vornehmen. Die Berechnung der Gesamtdistanz in unserem Beispielfall

für die Flutontologien erfolgt durch die Summation der einzelnen Distanzen, welche sich pro Ebene ergeben.

Bei der konkreten Berechnung der relativen Entropie muss beachtet werden, dass kein $q(x)$, mit $x \in \Upsilon$, gleich Null ist. In diesem Falle ist die Distanz zwischen $\mathbf{p}$ und $\mathbf{q}$ unendlich:

$$\lim_{p(x)\to 0} \log \frac{p(x)}{q(x)} = \infty$$

Es sollte noch angemerkt werden, dass die relative Entropie kein Distanzmaß im klassischen Sinne ist, denn sie erfüllt weder die Symmetrieeigenschaft noch die Dreiecksungleichung. Eine symmetrische Form der relativen Entropie erhält man, falls das geometrische Mittel der Summe

$$D(\mathbf{p}||\mathbf{q}) + D(\mathbf{q}||\mathbf{p})$$

betrachtet wird [Johnson u. Sinanovic 2001].

Die exakte Bedeutung der mittels der relativen Entropie berechneten Distanz hängt allerdings von den Wahrscheinlichkeitsverteilungen und den Ontologien ab, welche mit ihrer Hilfe strukturiert werden sollen, wie man anhand der unterschiedlichen Interpretationen in folgenden Abschnitt und in Kapitel 6 sehen kann.

5.3 Strukturierung anhand der Flutontologie

Die Vorgehensweise zur Strukturierung von Ontologien mittels der relativen Entropie soll anhand eines Beispieles mit der im vorigen Kapitel vorgestellten Flutontologie gezeigt werden. Hierbei wird nur die grundlegende Strukturierung nach den Vererbungsrelationen betrachtet. Die weiteren Relationen werden in diesem Beispiel außer Acht gelassen.

5.3.1 Bedeutung der Wahrscheinlichkeitsverteilung

Dazu wird zunächst die Bedeutung der Wahrscheinlichkeitsverteilung festgelegt. Sie soll für dieses Beispiel die *Wichtigkeit* der Konzepte i für das durch die Ontologie modellierte Ziel der Flutkatastrophenvorsorge darstellen. Die Basis der Wahrscheinlichkeiten soll das Wissen einer Person oder Personengruppe sein, aus dem sich die Wichtigkeiten ableiten lassen. Weiterhin wurde festgelegt, dass sich die Wahrscheinlichkeiten pro Ebene j der Ontologie zu eins summieren und damit pro Ebene eine Wahrscheinlichkeitsverteilung bilden (siehe Abb. 5.1):

$$\sum_{i \in Ebene_j} p_j(i) = 1$$

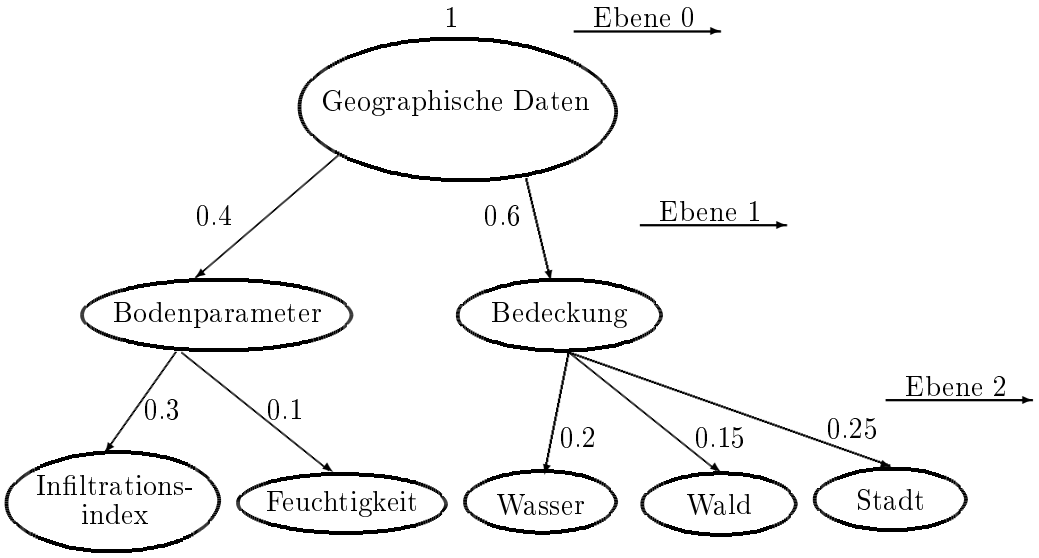


Abbildung 5.1: Alle Wahrscheinlichkeiten summieren sich pro Ebene zu 1.

Desweiteren soll die Summe der Wahrscheinlichkeiten der direkten Kindkonzepte k eines Elternkonzeptes e gleich der Wahrscheinlichkeit des Elternkonzeptes sein:

$$\sum_{k \in \{Kind\ von\ e\}} p(k) = p(e)$$

Diese Einschränkung besagt, dass spezialisierte Kindkonzepte nicht wichtiger bezüglich des definierten Zieles sein können, als das allgemeinere Elternkonzept.

5.3.2 Festlegung von p

Nun können die Wichtigkeiten **p** der einzelnen Konzepte bezüglich der Flutkatastrophenvorsorge festgelegt werden. Dazu wird ein Ausschnitt der Flutontologie in Abb. 5.2 betrachtet. Die zugewiesenen Wichtigkeiten wurden von den Experten (Hydrologen), aufgrund ihres Wissens, als optimal zur Erreichung der Flutkatastrophenvorsorge definiert. Sie stellen also ihr Modell hinsichtlich optimaler Flutkatastrophenvorsorge dar. Für die Hydrologen ist die langfristige Vorsorge wesentlich wichtiger, als kurzfristige Maßnahmen, die jeweils nur im akuten Notfall greifen. Von den Experten als besonders wichtig eingestuft wurden Polder, weil mit ihnen der Scheitel einer drohenden Flutwelle verringert werden kann. Ebenfalls signifikant sind die so genannten Flutpläne. In diesen wird dargestellt, welche Gebiete entlang eines Flusses bei einem Hochwasser gefährdet sind. Dadurch kann man beispielsweise bereits im vorraus eine Bebauung solcher Gebiete mit kritischen Objekten, wie

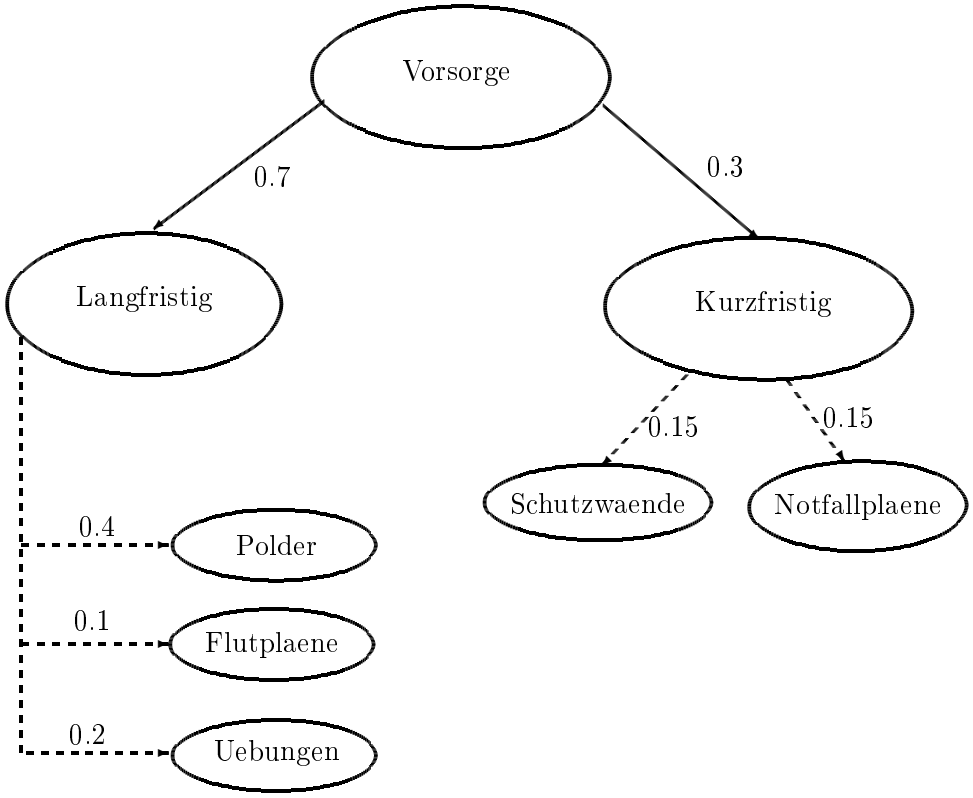


Abbildung 5.2: Wichtigkeit der Konzepte für **p**. Die gestrichelten Pfeile stellen die Relation *hatMaßnahme* dar, die durchgezogenen *is-a* Relationen.

Wohnhäusern oder Industrieanlagen, vermeiden. Ebenfalls nicht zu unterschätzen sind regelmäßige Übungen der Rettungskräfte.

Bei den kurzfristigen Maßnahmen wird die Erstellung von Notfallplänen, in denen Anweisungen für die jeweils beteiligten Akteure im Katastrophenfall enthalten sind, als etwas wichtiger erachtet als das Bereitstellen von Schutzmaßnahmen wie Flutschutzwänden oder Sandsäcken.

Die Wahrscheinlichkeiten, respektive Wichtigkeiten $\mathbf{p}$, repräsentieren somit ein Datenmodell der Hydrologen für die Flutkatastrophenvorsorge, welches auf ihrem persönlichen Wissen basiert und durch die Ontologie formalisiert wurde.

5.3.3 Festlegung von q

Die in Abb. 5.3 dargestellte Zuweisung der Wichtigkeiten $\mathbf{q}$ geschieht analog zu denjenigen von $\mathbf{p}$, welche im vorigen Abschnitt vorgestellt wurde. Die Wichtigkeiten stellen hier allerdings beispielsweise das Modell von Politikern oder beliebigen anderen Agenten wie Bevölkerung und Rettungskräften bezüglich der Flutkatastrophenvorsorge dar. Diese stimmen mit den Experten bei der Einschätzung der Wichtigkeit der langfristigen und kurzfristigen Maßnahmen überein. Allerdings sind in ihrem Modell die Polder und Übungen wichtiger für die Flutkatastrophenvorsorge als die Flutpläne. Das mag im Falle der Politiker daran liegen, dass Polder und Übungen sichtbare Maßnahmen sind, welche der Bevölkerung zeigen, dass aktiv etwas zu ihrem Schutz getan wird. Die Flutpläne bewerten sie dagegen als deutlich weniger wichtig, weil ihnen durch die Offenlegung der Gefährdungen von Objekten eventuelle Einnahmen oder dergleichen entgehen könnten.

5.3.4 Berechnung der Distanz

Mittels der relativen Entropie soll die Distanz D bestimmt werden, mit der das Modell der Hydrologen bezüglich der Flutkatastrophenvorsorge nicht mit dem Modell der anderen Agenten übereinstimmt. Falls sich eine kleine Distanz ergibt, sind informationstheoretisch ähnliche Modelle vorhanden, womit auch das Ziel der Flutkatastrophenvorsorge ähnlich gut erreicht wird wie bei dem Modell der Hydrologen. Eine krosse Distanz hingegen deutet auf andere Datenmodelle hin, so dass unter Umständen das Ziel der Flutkatastrophenvorsorge gefährdet ist (da die Hydrologen ja ein optimales Modell vorgeben). Die Distanz berechnet sich in diesem Fallbeispiel aus der Summe derjenigen Distanzen, welche sich jeweils aus den einzelnen Ebenen j der Ontologie berechnen:

$$D = \sum_j D_j(\mathbf{p}||\mathbf{q})$$

Für die erste Ebene ergibt sich aufgrund gleicher Wichtigkeit der Konzepte eine Distanz von Null zwischen den Teilontologien $\mathbf{p}$ der Experten und

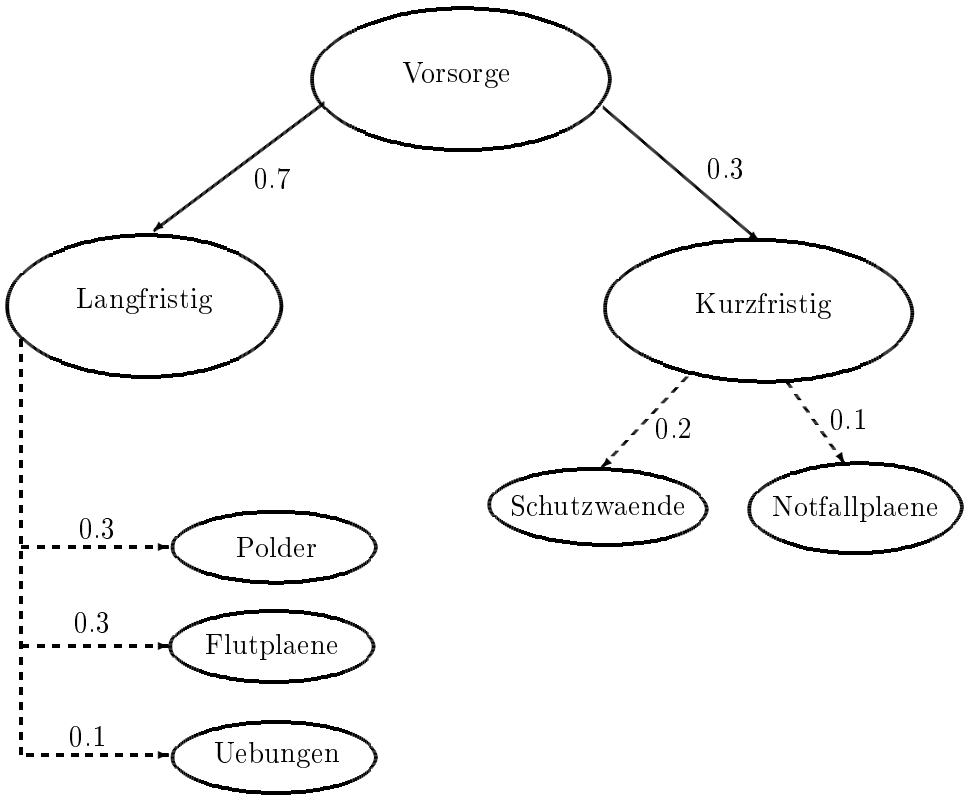


Abbildung 5.3: Wichtigkeit der Konzepte für $\mathbf{q}$. Die gestrichelten Pfeile stellen die Relation *hatMaßnahme* dar, die durchgezogenen *is-a* Relationen.

q der Politiker:

$$\begin{aligned} D_1(\mathbf{p}||\mathbf{q}) &= 0.7 \log \frac{0.7}{0.7} \\ &+ 0.3 \log \frac{0.3}{0.3} \\ &= 0 \end{aligned}$$

Ein anderes Bild ergibt sich bei Betrachtung der zweiten Ebene, denn hier treten Unterschiede in der Wichtigkeit der Konzepte auf.

$$\begin{aligned} D_2(\mathbf{p}||\mathbf{q}) &= 0.3 \log \frac{0.3}{0.4} \\ &+ 0.3 \log \frac{0.3}{0.05} \\ &+ 0.1 \log \frac{0.1}{0.25} \\ &+ 0.15 \log \frac{0.15}{0.2} \\ &+ 0.15 \log \frac{0.15}{0.1} \\ &= 0.54 \end{aligned}$$

Eine nähere Betrachtung der Kinderkonzepte der langfristigen Maßnahmen offenbart, dass hier eine Distanz von ca. 0.52 besteht. Bei den kurzfristigen Maßnahmen ist hingegen die Distanz hingegen deutlich kleiner, sie beträgt lediglich ca. 0.02. Somit ergibt sich für die Distanz D zwischen den dargestellten Ontologien der Politiker und Hydrologen der Abstand:

$$\begin{aligned} D &= D_1(\mathbf{p}||\mathbf{q}) \\ &+ D_2(\mathbf{p}||\mathbf{q}) \\ &= 0 + 0.54 \\ &= 0.54 \end{aligned}$$

5.3.5 Interpretation der Distanz

Die exakte Bedeutung der konkreten Werte, beispielsweise ab einer Distanz zwischen den Ontologien von 1.0 besteht eine signifikante Gefährdung der Gebäude in Flussnähe, muss empirisch bestimmt werden. Dies war aufgrund des begrenzten Zeitrahmens der Dissertation nicht möglich.

Wie man anhand des gezeigten Beispieles für die Flutontologie und der in Kapitel 6 vorgestellten Strukturierung von Musikstücken sehen kann, ist die konkrete Bedeutung der mit der relativen Entropie bestimmten Distanz zwischen Ontologien von der Interpretation der gewählten Wahrscheinlichkeitsverteilung sowie dem durch die Ontologie modellierten Wissen abhängig.

Dies gilt insbesondere bei der Verwendung von subjektiven Wahrscheinlichkeiten.

Weiterhin spielt die Anwendung, in dessen Rahmen die vorgeschlagene Methodik verwendet wird, eine wichtige Rolle bei der Interpretation der Distanz. In einem sicherheitskritischen Kontext wird die akzeptable Distanz zwischen zwei Ontologien bzw. dem durch sie modellierten Wissen sicherlich geringer sein, als beispielsweise in dem vorigen Beispiel oder der in Kapitel 6 vorgestellten Anwendung für die musikalische Wahrnehmung.

5.4 Implementierung

Zur Berechnung der Distanz zwischen Ontologien mittels der in diesem Kapitel vorgestellten Verfahrensweise, wurde ein prototypisches System implementiert. Dieses System ermöglicht nicht nur die Berechnung der Distanz zwischen zwei Ontologien, sondern es erlaubt weiterhin, ausgehend von einer gegebenen Wahrscheinlichkeitsverteilung $\mathbf{p}$, die Wahrscheinlichkeiten $\mathbf{q}$ zu variieren und die daraus resultierenden Distanzen zu berechnen. Damit kann ein Überblick über diejenigen Wahrscheinlichkeitsverteilungen gegeben werden, die eine geringe Distanz aufweisen und somit ähnliche Datenmodelle besitzen und solche die krosse Distanzen, also von $\mathbf{p}$ verschiedene Datenmodelle ergeben. Solche Berechnungen lassen sich aufgrund der einfachen Beschaffensweise der relativen Entropie schnell durchführen.

Die Implementierung des Prototyps erfolgte mit Groovy[1], einer auf Java[2] basierenden Skriptsprache, die für alle weitverbreitete Betriebssysteme zur Verfügung steht.

Zur Berechnung der möglichen Belegungen für die Wahrscheinlichkeitsverteilungen der von einem Knoten ausgehenden Kanten wurden numerische Partitionen[3] verwendet. Eine numerische Partition einer Zahl n ist eine Sequenz

$$p_1 \geq p_2 \geq \cdots \geq p_k$$

der Länge k, so dass

$$p_1 + p_2 + \cdots + p_k = n$$

gilt. Die p_i sind die möglichen k Wahrscheinlichkeitsbelegungen der Kanten, die in Summe die Wahrscheinlichkeit des Elternkonzeptes n ergeben müssen. Zur Berechnung der möglichen Wahrscheinlichkeitsbelegungen für die Kanten eines Konzeptes mussten die Wahrscheinlichkeiten zunächst um den Faktor zehn skaliert werden, damit die numerische Partition von n berechnet werden konnte. Sämtliche verfügbaren Implementierungen zur Berechnung von numerischen Partition basieren nämlich auf ganzen Zahlen. Aus Effizienzgründen wurden die numerischen Partitionen für jedes n mit zugehöriger

[1] http://groovy.codehaus.org
[2] http://java.sun.com
[3] http://www.theory.cs.uvic.ca/ cos/

Länge k zwischengespeichert, so dass die Ergebnisse für die Belegung der Kanten bei gleichem n und k wiederverwendet werden konnte. Nach dem die Belegungen berechnet wurden, muss wieder durch den Faktor zehn dividiert werden, um die Wahrscheinlichkeiten zu erhalten.

Ein kleines Beispiel soll die Vorgehensweise verdeutlichen. Anhand Abb. 5.1 sollen die Belegungsmöglichkeiten für *kurzfristig* berechnet werden:

1. Die Summe der Wahrscheinlichkeiten beider Kinderkonzepte darf nicht größer als $0.3 \cdot 10 = 30$ sein. Damit muss eine numerische Partition für $n = 30$ der Länge $k = 2$ (Anzahl ausgehender Kanten) berechnet werden.

2. Die minimale Wahrscheinlichkeiten muss $0.05 \cdot 10 = 5$ pro Kindkonzept sein.

3. Die Veränderung der Wahrscheinlichkeiten erfolgt in fünfer Schritten (siehe Abschnitt 5.4.1).

4. Nun werden alle möglichen Belegungen berechnet, die diesen Restriktionen genügen:

   (a) $15, 15$
   (b) $10, 20$
   (c) $5, 25$
   (d) $25, 5$
   (e) $20, 10$

5. Im letzten Schritt werden diese Werte wieder durch zehn dividiert um die korrekten Wahrscheinlichkeiten zu erhalten.

Die Berechnung dieser numerischen Partitionen erfolgt durch eine frei verfügbare Java Bibliothek[4] des JASA[5] Projektes.

5.4.1 Restriktionen

Um die Effektivität der Berechnung bei der Variation der Wahrscheinlichkeiten zu gewährleisten, können die Wahrscheinlichkeiten nur in Schritten von 0.05 geändert werden. Beispielsweise sind für die Kante zwischen *kurzfristig* und *Schutzwall* aus Abb. 5.2 nur die Werte $0.05, 0.1, 0.15, 0.2, 0.25$ und 0.3 möglich. Weiterhin wurde als Minimum für die Wahrscheinlichkeiten der Konzepte 0.05 und nicht 0 vorgesehen, da die relative Entropie sonst gegen unendlich strebt. Ansonsten gelten die bereits beschriebenen Restriktionen

[4] http://jasa.sourceforge.net/doc/uk/ac/liv/util/Partitioner.html
[5] http://jasa.sourceforge.net

für die Zuweisung der Wahrscheinlichkeiten. Namentlich ist dies die Einschränkung, dass die Summe der Überzeugungen der Kinderkonzepte nicht größer als die des Elternkonzeptes sein darf und die Summe der Wahrscheinlichkeiten sich pro Ebene auf ein summiert.

5.4.2 Benutzung

Die Benutzung des Programmes ist sehr einfach. Die Eingabe der Ontologie mit ihren Wahrscheinlichkeitsverteilungen $\mathbf{p}$ erfolgt durch eine Adjazenzmatrix, in der die Elemente a_{ij} die Wahrscheinlichkeiten des Kindkonzeptes j, welches mit einem Elternkonzept i verbunden ist, darstellt. 0 bedeutet, dass keine Verbindung zwischen dem Konzept i und j besteht. Die Ausgabe der Distanz zu den Ontologien mit variierenden Wahrscheinlichkeitsverteilungen q erfolgt in eine vom Benutzer vorgegebene Textdatei. Dabei wird die Adjanzenzmatrix mit den Wahrscheinlichkeitsverteilungen q und die Distanz zu der Ontologie mit Wahrscheinlichkeitsverteilungen $\mathbf{p}$ ausgegeben.

Kapitel 6

Ontologien und Entropie in der Musik

Im Laufe dieser Arbeit schlug ein bekannter Komponist, Herr Boris Yoffe, ein Modell für die Kreativität in der Musik vor. Dieses Modell zeigte eine starke Ähnlichkeit mit den Modellen aus der Informatik, insbesondere aus dem Bereich der Agenten [Weiss 2000] und „Emotionen in Multi-Agenten Systemen“ [Petta u. Trappl 2001]. Ein wesentlicher Bestandteil dieser Modelle sind Ontologien, mit denen es möglich ist, einen Kontext und daraus eine (spezialisierte) Semantik zu entwickeln. Im Hinblick auf den musikalischen Kontext bedeutet dies, dass Kreativität eine Art „Erweiterung“ der zugrundeliegenden Ontologie ist, und damit auch des Kontextes und der Semantik, durch hinzufügen neuer Konzepte. Im musikalischen Bereich ist die daraus neu entstehende Semantik persönlich, da Kunst im allgemeinen sehr subjektiv ist. Ein Problem, dass sich hierbei stellte, war eine bislang fehlende Datenstruktur zur Formalisierung von gehörter Musik.

In der Musik ist die Verwendung von Computern heutzutage alltäglich, sei es zur Erstellung eines Notensatzes, zur Livebearbeitung des Auftrittes einer Musikgruppe oder zur Komposition von elektronischer Musik. All diesen Anwendungen ist gemeinsam, dass sie den Musiker bei der Erschaffung oder Bearbeitung von Musik unterstützen. Den umgekehrten Weg, d.h. ein gehörtes Musikstück zu formalisieren (Transkription), ist dagegen bedeutend schwieriger [Temperley 2002]. Hierbei soll anhand eines gehörten Musikstückes eine geeignete, formale Repräsentation, beispielsweise durch Noten, gefunden werden. Erste Ansätze zur Transkription decken lediglich einen Teilbereich eines Musikstückes ab, wie beispielsweise die Extraktion eines Rhythmus [Gouyon u. Dixon 2005]. Eine formale Darstellung eines Musikstückes, anhand dessen eine abstrakte Repräsentation im Computer möglich ist, soll im folgenden durch Ontologien ermöglicht werden.

Die entwickelte Ontologie soll eine formale Darstellung eines gehörten Musikstückes ermöglichen, wobei bei ihrer Erstellung insbesondere die mensch-

liche, musikalische Wahrnehmung berücksichtigt wurde. Die Zuverlässigkeit der Modellierung gewährleistete dabei eine enge Zusammenarbeit mit dem Institut für neue Musik und Medien der Musikhochschule Karlsruhe, insbesondere mit Herrn Prof. Dr. T. Troge. Durch die Ontologie dargestellte Musikstücke sind somit Wissen und dieses Wissen kann wiederum strukturiert werden. Zur Strukturierung der Ontologien soll das in Kapitel 5 vorgestellte Distanzmaß angewandt und damit auch validiert werden. Die konkrete Bedeutung der Wahrscheinlichkeitsverteilungen, welche für die Benutzung der relativen Entropie als Distanzmaß gegeben sein müssen, ist für die Validierung eine frequenzbasierte. In weiteren Anwendungen können allerdings auch subjektive Wahrscheinlichkeiten (Interessantheit einer Interpretation) verwendet werden.

6.1 Erzeugen von Musik

Bei der bisherigen Nutzung von Rechnern in der Musik stand vor allem die Unterstützung des Musikers bei seiner kreativen Arbeit im Mittelpunkt, im speziellen die Erzeugung von Noten und Klängen mit dem Computer.

Das MIDI Datenformat (Musical Instruments Digital Interface), welches den Musiker bei der Komposition und Notation unterstützt [Gorges 1997], ist seit mehreren Jahren etabliert. MIDI ist ein Datenformat, welches genau festlegt, wie die damit beschriebene Musik klingen soll. Die Datenstruktur ist einfach aufgebaut und besteht unter anderem aus den folgenden Befehlen:

- Note On: Gibt den Anschlagszeitpunkt einer Note an. Vergleichbar mit dem Drücken einer Taste auf einer Klaviatur.
- Note Off: Gibt den Endzeitpunkt einer Note an, vergleichbar mit dem Loslassen einer Taste auf einer Klaviatur.
- Velocity: Gibt die Geschwindigkeit an, mit der eine Klaviertaste gedrückt wurde. Dies ist ein Indikator für die Lautstärke des gespielten Tones.
- Pitchbend: Gibt die Änderung der Tonhöhe eines Tones an, während er gespielt wird. Diese Tonhöhenänderung beschränkt sich lediglich auf einen relativ kleinen, tonalen Bereich.

Wie aus dieser Datenstruktur ersichtlich wird, legt MIDI die *Notenschrift* für den Rechner fest. Wie ein Stück dann in Wirklichkeit erklingt, hängt von den Programmen und Geräten ab, die MIDI interpretieren und anschließend in hörbare Töne umsetzen. Diese synthetisch erzeugten Klänge liefern Programme und Geräte wie zum Beispiel Synthesizer [Braut 1993].

Systeme wie *Csound*[1] [Boulanger 2000] erlauben dem Musiker verschiedene Klänge zu generieren und miteinander zu kombinieren. Die Datenstruktur

[1] http://www.ugrad.physics.mcgill.ca/reference/Csound/INDEX.html

von Csound lehnt sich an die Programmiersprache *C* an. Sie erlaubt dem Musiker die Entwicklung von Musikstrukturen und Klangdesign sowie beliebige Kombinationen solcher Strukturen. Beispielsweise kann das Programm angewiesen werden, einen Sinuston in einer bestimmten Lautstärke einige Sekunden lang zu spielen, um ihn anschließend mit weiteren Werkzeugen (Verzerrung, Modulation) manipulieren zu können. Bei der Produktion von Musik spielen Sequenzer wie Logic Pro von Apple Computers[2] oder Protools der Firma Digidesign[3] eine zentrale Rolle. Sie erlauben, zum Teil in Echtzeit, die Manipulation eingespielter Musik. Beispielsweise kann bei einem Auftritt eines Popstars dessen Gesungenes um einige Hertz verändert werden, um so eventuelle Fehler auszugleichen. Diese Programme können ebenfalls zum arrangieren der Aufnahmen von Musikern verwendet werden. Die einzelnen, aufgenommen Teile werden dann zu einer Gesamtkomposition zusammengefügt (Studioaufnahmen).

Eine wichtige Rolle bei der Unterstützung von Musikern durch Rechner spielen Notensatzprogramme wie Finale[4]. Diese vereinfachen das Erstellen von Partituren erheblich und liefern diese auch in einer ansprechenden äußeren Form ab. Auch erlauben sie das einfache Einspielen der Noten mittels dem zuvor besprochenen MIDI Format.

6.2 Wahrnehmung von Musik

All den im vorigen Abschnitt vorgestellten Techniken ist gemeinsam, dass sie den Musiker bei seiner kreativen Arbeit unterstützen. Der umgekehrte Weg, die musikalische Wahrnehmung des Menschen zu erfassen und zu formalisieren (Transkription) bereitet dagegen grosse Schwierigkeiten. Hierbei soll anhand eines gehörten Musikstückes eine geeignete, formale Repräsentation, beispielsweise durch Noten, gefunden werden.

Erste Ansätze zur Transkription decken lediglich einen Teilbereich eines Musikstückes ab, wie beispielsweise die Extraktion eines Metrums und Rhythmuses [Gouyon u. Dixon 2005]. Dieser Ansatz konzentriert sich dabei auf frequenzbasierte Analysen der Musikstücke mittels eines Multiagentensystems. Dazu versuchen die Agenten herausstechende, akustische Ereignisse zu erkennen und diese in Gruppen zusammenzufassen. Das Klassifikationskriterium für die Gruppenzugehörigkeit ist ein Zeitintervall zwischen zwei Ereignissen. Die in den jeweiligen Gruppen zusammengefassten Hypothesen für das Metrum bzw. den Rhythmus werden für das gesamte Musikstück überprüft. Falls sich eine Übereinstimmung für das ganze Musikstück findet, wurde ein Metrum bzw. Rhythmus erkannt. [Goto 2001] verwenden ebenfalls ein Multiagentensystem zur Erkennung von Metrum und Rhythmus, welches

[2] http://www.apple.com/logic
[3] http://www.digidesign.com
[4] http://www.finalemusic.com

auf drei musikalischen Parametern aufbaut. Das sind die Anschlagzeiten der Noten, Akkordwechsel und bekannte Rhythmusmuster von Schlagzeugern. Bei der Benutzung dieses Systems werden bereits stark einschränkende Annahmen über die Stücke gefällt. So wird beispielsweise für das Grundmetrum ein 4/4 Takt angenommen. Die Qualität der Ergebnisse dieser Ansätze schwankt stark, je nachdem welcher Musikrichtung das analysierte Musikstück angehört. So wird das Metrum bei Rock- oder Popmusik gut erkannt ($90 - 100\%$ Erkennungsrate), bei Jazzmusik ist die Erkennung jedoch deutlich schlechter (75%).

Für die Transkription von einstimmigen Melodien existiert ein System [Viitaniemi u. Eronen 2003], welches auf einem Hidden-Markov Modell für die Tonhöhen, einem musikwissenschaftlichen Modell für Vorzeichen und Tonarten sowie den Zeitdauern einzelner Klänge basiert. Die Fehlerraten liegen, selbst bei einfachen, einstimmigen Melodien, zwischen 10 und 20%. Ein System zur Transkription von Klavierstücken mittels eines Blackboardsystems und neuronalen Netzen wird in [Bello u. a. 2000] beschrieben. Hierbei werden allerdings starke Einschränkungen gemacht. So dürfen die Klänge nur in einem bestimmten Intervall liegen und es dürfen nicht mehr als drei Klänge gleichzeitig erklingen. In [Temperley 2002] wird ein regelbasiertes System erstellt, um grundlegende, musikalische Strukturen bei der menschlichen Wahrnehmung zu erkennen. Für weiteregehende Informationen hierzu siehe [Kink 2005].

6.3 Ontologie für die menschliche, musikalische Wahrnehmung

Aufgrund der bestehenden Beschränkungen der bisherigen Ansätze zur formalen Repräsentation von Musik wurde in dieser Arbeit ein anderer Ansatz für die formale Darstellung eines Musikstückes gewählt. Die formale Repräsentation eines Musikstückes erfolgt durch eine Ontologie, welche die menschliche, musikalische Wahrnehmung berücksichtigt. Abb. 6.1 zeigt einen Überblick über die wichtigsten Konzepte der Ontologie, deren Implementierung in OWL erfolgte. Die gezeigte Ontologie ist allerdings nur für die Darstellung europäischer Musik geeignet. In diesem Zusammenhang zeichnet sich europäische Musik unter anderem durch das Vorhandensein eines Metrums und Rhythmuses sowie Tönen und Klängen aus. Weiterhin müssen die Töne und Klänge eine stimmige Melodie bilden und eine mögliche Begleitung sollte auf „europäischen“ Harmonien basieren. Das schließt beispielsweise einige Kirchenchoräle aus, da die dort verwendeten Harmonien auf reinen Molldreiklängen basieren. Auch atonale Musik, beispielsweise indische Stücke, sowie elektronische Kompositionen können mit der Ontologie in ihrer jetzigen Ausbauphase nicht formalisiert werden.

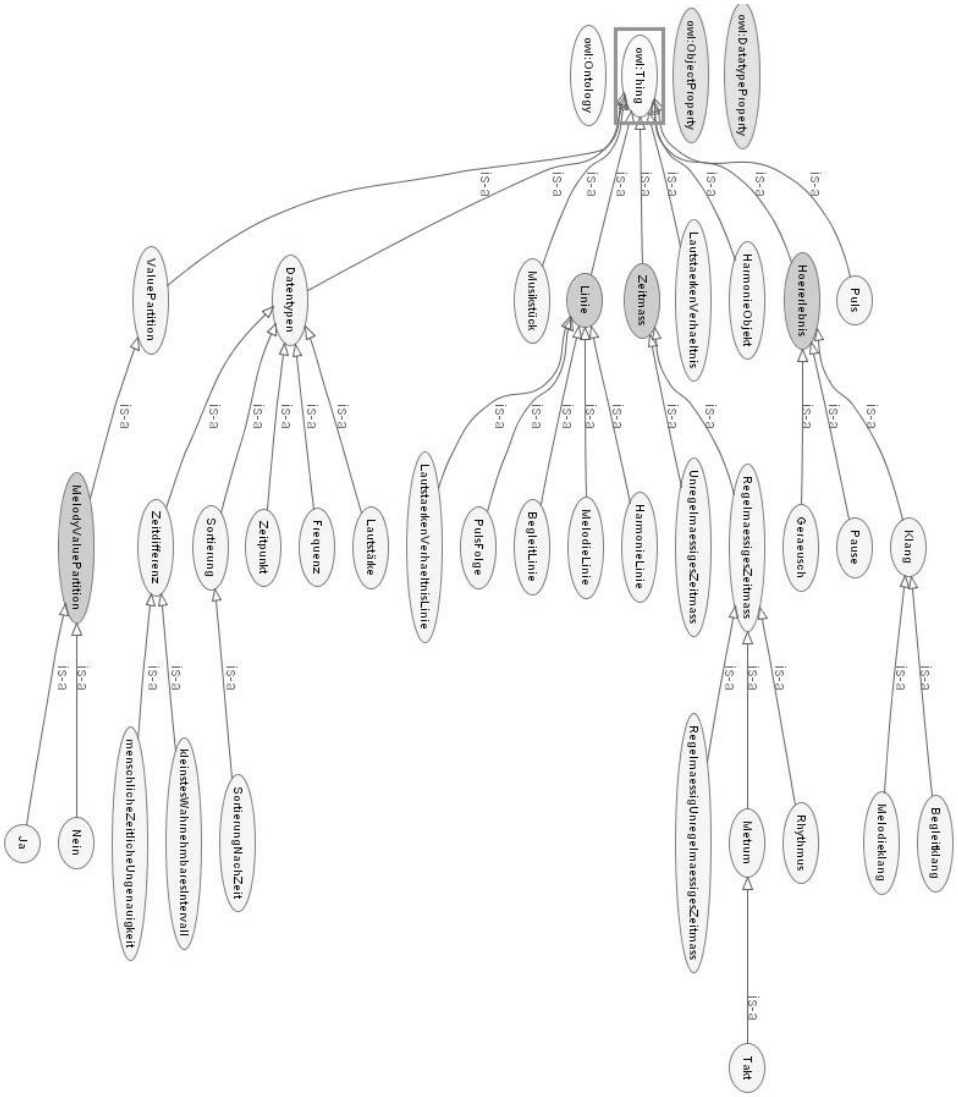


Abbildung 6.1: Ontologie für die musikalische Wahrnehmung

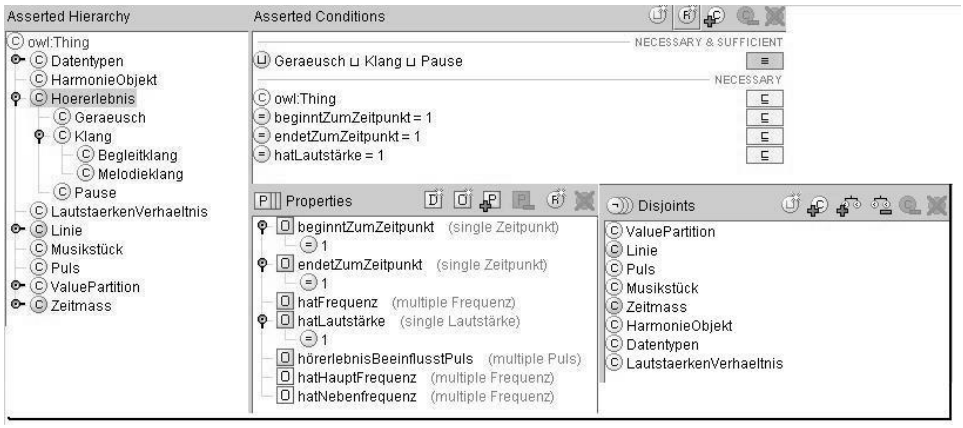


Abbildung 6.2: Die Klasse Hörerlebnis.

6.3.1 Hörerlebnis

Eine der wichtigsten Klassen in der Ontologie ist das *Hörerlebnis* (siehe Abb. 6.2). Ein Hörerlebnis stellt alles hörbare dar und unterteilt sich in drei weitere, disjunkte Untermengen:

1. Ein Klang beschreibt ein Hörerlebnis, welches eine Tonhöhe hat.

2. Ein Geräusch ist ein Hörerlebnis, dass mindestens zwei Frequenzen aufweist, aber keine Tonhöhe besitzt. Falls nur eine Hauptfrequenz hörbar ist, handelt sich um einen Klang, denn durch eine hörbare Hauptfrequenz ist die Tonhöhe bereits eindeutig festgelegt.

3. Die dritte Untermenge beschreibt ein Hörerlebnis, bei dem *nichts* zu hören ist. Diese Menge repräsentiert somit eine *Pause*, die in der Musik oftmals als rhythmisches und stilistisches Mittel eingesetzt wird.

Ein Hörerlebnis besitzt genau einen Anfangs- und Endzeitpunkt, da es über einen begrenzten Zeitraum stattfindet. Mittels der Attribute Frequenz und Lautstärke wird der Klang eines Hörerlebnisses festgelegt. Die drei Parameter Zeitdauer, Frequenz und Lautstärke bestimmen physikalisch, welche musikalischen Ereignisse das menschliche Ohr wahrnehmen kann. Die Frequenz untergliedert sich hierbei noch in eine Haupt- und Nebenfrequenz. Töne, welche durch Instrumente wie Posaune, Klavier oder Trompete erzeugt werden, bestehen aus einem Grundton (Grundfrequenz) und mehreren gleichzeitig erklingenden Obertönen (Nebenfrequenzen). Die Gesamtheit aller noch zusätzlich zu dem Grundton schwingenden Obertöne ergibt dann das Frequenzspektrum des Tones (Obertonreihe), welches den charakteristischen Klang eines Instrumentes ausmacht. Für weitere Informationen hierzu siehe [Kink 2005].

Hörerlebnisse beeinflussen direkt das Metrum- und Rhythmusempfinden des Menschen, weshalb sie in direkter Relation zu ihnen stehen. Indirekt werden das Metrum- und Rhythmusempfinden durch die Lautstärkenverhältnisse zweier Hörerlebnisse beeinflusst, welches durch die Beziehung *beeinflusstPuls* modelliert wird.

Klang

Ein Klang wird in der Ontologie vereinfacht durch eine festgelegte Tonhöhe und Lautstärke dargestellt. Ein realer Klang, welcher sich während seiner Spieldauer ändert, wird in der Ontologie durch mehrere Klänge kürzerer Zeitdauer modelliert. Ein Beispiel soll diese Vorgehensweise näher verdeutlichen: Ein Geiger spielt einen sehr langen Ton an, versieht ihn nach und nach mit einem Vibrato und wird dabei lauter (crescendo). In der menschlichen Wahrnehmung wird dies als nur ein Klang empfunden. In der Ontologie müsste er jedoch in kleine Zeitintervalle aufgeteilt werden mit dazugehöriger Tonhöhe, Frequenzen und entsprechender Lautstärke.

Ein Klang innerhalb eines Musikstückes lässt sich weiter in einen Melodieklang und einen Begleitklang unterteilen. Innerhalb der europäischen Musik ist diese Aufteilung disjunkt, denn jeder Klang gehört entweder zu einer Melodielinie oder einer Harmonielinie.

Geräusch

Als Geräusche werden in der Ontologie all diejenigen Hörerlebnisse dargestellt, die nicht als Klang oder Pause wahrgenommen werden. Geräusche werden insbesondere im rhythmischen Bereich eingesetzt. Beispielsweise erzeugen Trommeln, im Gegensatz zu Pauken, keine Töne mit definierter Tonhöhe. In manchen Musikstücken werden Geräusche aus der Umwelt auch als Stilmittel eingesetzt, wie zum Beispiel Kanonenschläge in Tschaikowsky Ouvertüre 1812.

Pause

Eine Pause beschreibt ein Hörerlebnis, bei dem über einen bestimmten Zeitraum „nichts“ zu hören ist. Die Pause ist ein besonderes Stilmittel in der Musik und wird oft als Überraschungsmoment verwendet. Auch ist sie ein wichtiges Mittel zur Gestaltung eines interessanten und abwechslungsreichen Rhythmuses.

6.3.2 Linien

Die Klasse *Linie* (siehe Abb. 6.3) dient als Behälter für Ansammlungen von Objekten gleichen Typs, die eine chronologische Reihenfolge aufweisen. Demzufolge besitzen alle Linien einen Anfangs- und Endzeitpunkt, sowie eine

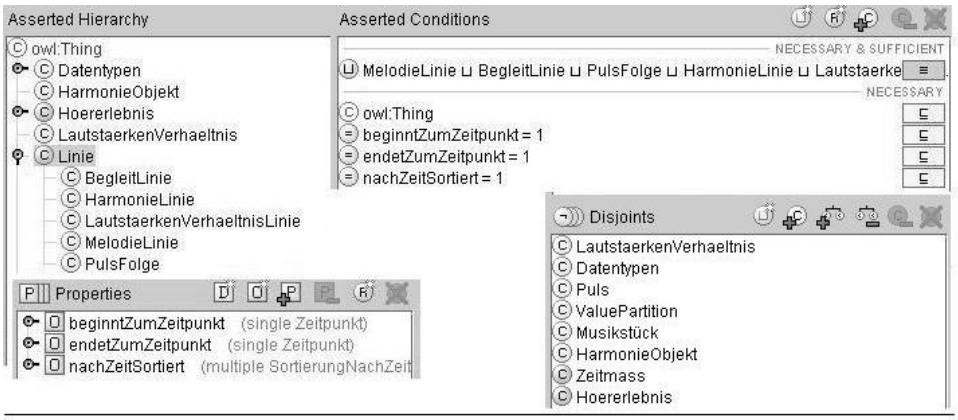


Abbildung 6.3: Die Klasse Linie.

zeitliche Sortierung der in ihnen enthaltenen Objekte. Aus offensichtlichen Gründen ist dies besonders bei der Melodie- und Begleit- und Harmonielinie wichtig. Die von *Linie* abgeleiteten Klassen sind die Melodie-, Begleit-, Harmonie- und Lautstärkenverhältnislinie sowie die Pulsfolge.

Melodielinie

Eine Melodielinie beschreibt die fortlaufende Melodie eines Stückes, welche aus einer Menge von Melodieklängen besteht und in der Regel einstimmig ist. Menschen erkennen ein Melodiemuster bereits ab drei Klängen, so dass in der Ontologie mindestens drei aufeinanderfolgende Klänge für die Definition einer Melodielinie vorhanden sein müssen. Eine Melodielinie bildet sich also aus einer „horizontalen“ Folge von Melodieklängen. Die Darstellung mehrerer Melodielinien, beispielsweise bei einem Kanon, ist ebenfalls möglich. Bei polyphonen Stücken variiert der Anfangs- und Endzeitpunkt einer Melodielinien von Mensch zu Mensch, da es keine objektive, eindeutige Zuordnung der Klänge zu Melodie oder Begleitung gibt. Auch kann ein Melodieklang mehreren Melodielinien angehören, beispielsweise wenn sich bei einem Kanon zwei Melodien kurzzeitig auf einem Ton treffen.

Eine Melodielinie übt einen erheblichen Einfluss auf den Rhythmus eines Stückes aus, weshalb eine entsprechende Relation zu dem Konzept *Rhythmus* in der Ontologie besteht. Durch die Relation *beeinflusstZeitmaß* wird weiterhin berücksichtigt, dass eine Melodielinie einen eigenen Rhythmus besitzt und unter Umständen zusammen mit der Begleitlinie sogar einen eigenen Rhythmus bildet.

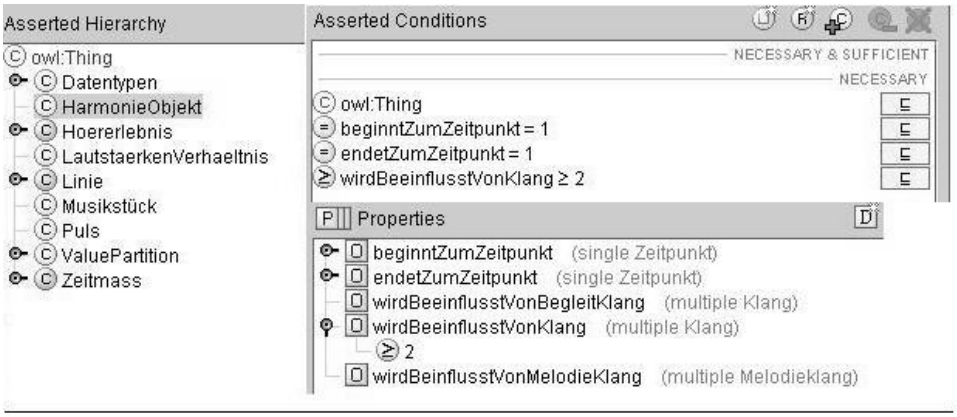


Abbildung 6.4: Die Klasse Harmonieobjekt.

Begleitlinie

Alle Klänge die nicht zur Melodielinie gehören, werden als Begleitklänge aufgefasst und bilden die Begleitlinie. Im Gegensatz zur Melodielinie sind Begleitlinien in der Regel mehrstimmig und bestehen aus mindestens einem Klang. Die Beziehungen der Begleitlinie zum Rhythmus ähneln denen der Melodielinie, so dass hier keine nähere Beschreibung erfolgt.

Harmonielinie

Ein Harmonieobjekt (siehe Abb. 6.4) beschreibt eine Harmoniebeziehung zwischen mindestens zwei Klängen und erklingt immer in einem bestimmten Zeitraum. In diesem Zeitraum müssen die an der Harmonie beteiligten Klänge gleichzeitig erklingen. Im Gegensatz zur Melodielinie besteht eine Harmonielinie also aus einer „vertikalen“ Folge von Klängen, nämlich den einzelnen Akkorden. Mit Hilfe der Relationen *wirdBeeinflusstVonMelodieklang* und *wirdBeeinflusstVonBegleitklang* lässt sich ausdrücken, ob die Harmonien an Melodieklängen oder nur an Begleitklängen beteiligt sind. Mehrere Harmonieobjekte, es müssen mindestens zwei sein, bilden schließlich eine Harmonielinie.

Lautstärkenverhältnislinie und Pulsfolge

Die Lautstärkenverhältnislinie fasst die *Lautstärkenverhältnisse* (siehe Abschnitt 6.3.4) zu einer zeitlich geordneten Linie zusammen. Aus diesen werden dann die *Pulse* bestimmt, deren chronologische Abfolge die *Pulsfolge* bildet.

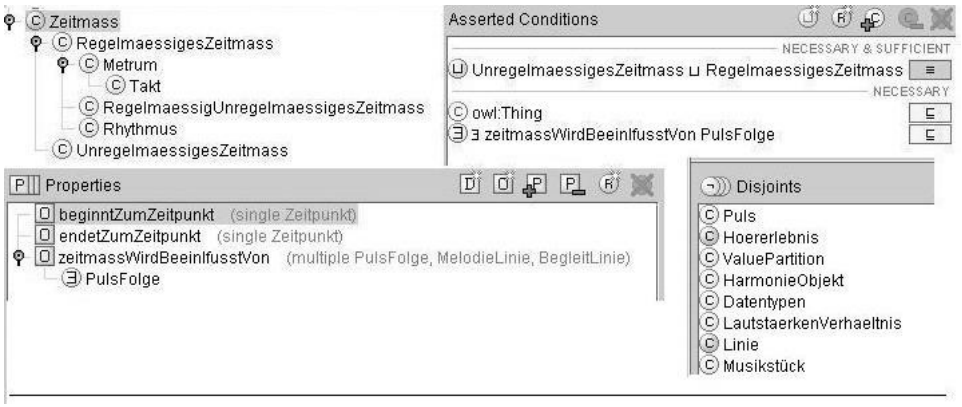


Abbildung 6.5: Die Klasse Zeitmaß.

6.3.3 Zeitmaß

Die Klasse Zeitmaß (siehe Abb. 6.5) beschreibt die möglichen Zeiteinteilungen eines Musikstückes. Sie bildet eine Oberklasse für alle musikalischen Zeiteinteilungen eines Stückes und gilt immer für ein bestimmtes Zeitintervall, weshalb jedes Zeitmaß einen Anfangs- und Endzeitpunkt besitzt. Jedes Zeitmaß wird von einer Pulsfolge beeinflusst, weshalb eine *beeinflusst* Relation zu der Klasse *Pulsfolge* besteht. Das Zeitmaß untergliedert sich in die regelmäßigen und unregelmäßigen Zeitmaße. Unter das regelmäßige Zeitmaß fallen das Metrum und der Rhythmus. Zeitliche Stilmittel der Musik, welche sich linear verändern (accelerando - schneller werdend oder ritardando - langsamer werdend) werden unter der Klasse *regelmäßig-unregelmäßiges* Zeitmaß zusammengefasst.

Metrum

In der Musik bezeichnet das Metrum den gleichmäßigen, unbetonten Grundschlag eines Musikstückes. Damit legt das Metrum die Geschwindigkeit des Stückes fest. In der Ontologie ist das Metrum durch ein Zeitintervall definiert, welches die Zeitdifferenz zwischen zwei Schlägen angibt. Falls die Musik elektronisch erzeugt wurde, muss lediglich die Startzeit angegeben werden, aus welcher sich dann alle weiteren Intervalle berechnen lassen. Die Zeiteinteilung wird aber ungenau, sobald die Musik von Menschen erzeugt wird, denn das menschliche Zeitgefühl ist nicht perfekt. Aus diesem Grund variieren die Zeitintervalle zwischen zwei Klängen bei von Menschen erzeugter Musik. Die Toleranz der menschlichen Wahrnehmung, bis zu der unterschiedlich lange Zeitintervalle als gleich erkannt werden, hängt im wesentlichen von zwei Faktoren ab:

1. Je langsamer ein Stück ist, desto größer können die Ungenauigkeiten

zwischen den Intervallen sein.

2. Jeder Mensch ist unterschiedlich empfindlich für zeitliche Schwankungen.

In der Ontologie wird dieser Umstand durch einen *Unsicherheitsfaktor* modelliert.

Ein Takt ist definiert als die Zusammenfassung von mehreren Schlägen des Metrums. Diese Schläge bilden eine metrisch geordnete Einheit von betonten und unbetonten Zählzeiten. In der Ontologie ist es mögliche, alle möglichen Taktarten zu modellieren. Sie müssen allerdings immer auf ein Zeitintervall beschränkt sein.

Rhythmus

Die Klasse Rhythmus ist ein Behälter für alle möglichen Arten von Rhythmen oder Rhythmusmustern. Unter einem Rhythmus versteht man die unterschiedlichen Bedeutungen der Tondauern innerhalb eines oder mehrerer Takte. Diese Tondauern bilden, zum Teil wiederkehrende, Akzentmuster. Nur selten kommt es vor, dass Rhythmus und Metrum identisch sind. In der Regel ist der Rhythmus eines Musikstückes wesentlich komplexer und trägt somit dessen Spannungaufbau bei. Im Jazz oder Blues werden beispielsweise häufig Synkopierungen (gegen das Metrum laufende Schläge) und Pausen zur Gestaltung eines interessanten Rhythmus verwendet.

Regelmäßig-unregelmäßiges Zeitmaß

In der Klasse Regelmäßig-unregelmäßiges Zeitmaß werden alle musikalischen Ereignisse zusammengefasst, bei denen die Zeitabstände untereinander zwar nicht mehr gleich sind, aber dennoch einen Zusammenhang aufweisen. In einem Musikstück sind dies beispielsweise solche Stellen, an denen das Tempo verlangsamt (ritardando) oder beschleunigt (accelerando) wird.

Unregelmäßiges Zeitmaß

Das unregelmäßige Zeitmaß fasst alle Zeitmaße zusammen, die nicht unter die regelmäßigen Zeitmaße fallen. Solche Stellen sind in einem Musikstück eher selten. Ein Beispiel hierfür wäre eine Kadenz, in welcher ein Musiker völlig frei improvisieren kann.

6.3.4 Lautstärkenverhältnisse

Bei der Bestimmung eines Zeitmaßes spielen die Lautstärkenverhältnisse zwischen zwei Klängen eine wichtige Rolle (siehe Abb. 6.6). Wenn auf einen sehr lauten bzw. leisen Klang plötzlich ein leiser bzw. lauter Klang folgt, wird dies

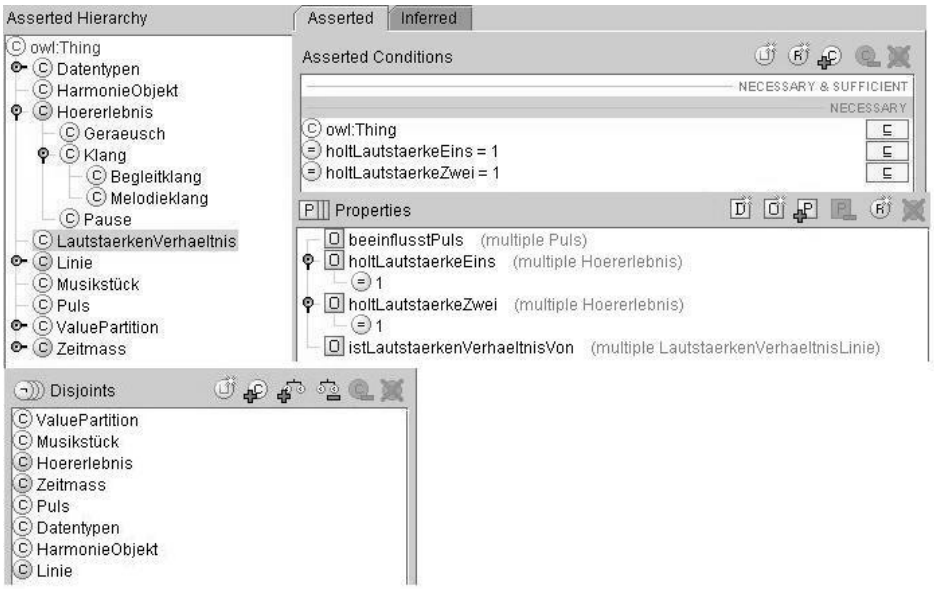


Abbildung 6.6: Die Klasse Lautstärkenverhältnisse.

von Menschen als Puls empfunden. Daraus entsteht dann eine Folge von betonten und unbetonten Stellen, welche eine Pulsfolge ergeben. Diese Beziehungen werden durch entsprechende Relationen mit den Klassen *Puls* bzw. den *Pulsfolgen* formalisiert. Da in OWL keine ternären Relationen möglich sind, wurde die Verbindung des Lautstärkenverhältnisses mit den dazugehörigen Klängen über Kardinalitätsrestriktionen realisiert.

6.3.5 Puls

Der Begriff *Puls* (siehe Abb. 6.7) steht für eine Art Impuls, der vor allem gefühlt wird. Er entsteht hauptsächlich durch plötzliche Lautstärkeveränderungen, welche bestimmte Stellen in einem Musikstück besonders wichtig erscheinen lassen. Beispielsweise entsteht ein Puls, wenn ein sehr langer, leiser Ton durch einen kurzen, lauteren Ton, unterbrochen wird. Das Anspielen eines Tones erzeugt ebenfalls einen Puls, allerdings können solche Pulse unterschiedlich stark ausfallen. Die Stärke eines Pulses hängt primär davon ab, wie ein Ton angespielt wird. Wenn der Ton hart angespielt wird, empfinden Menschen diesen Ton wichtiger als einen weich angestoßenen. Aufgrund dessen steht die Klasse Puls in enger Relation zu den Lautstärkeverhältnissen, denn aus ihnen bestimmt sich der Puls. Desweiteren wird der Puls auch direkt durch Hörerlebnisse, beispielsweise einen Pauken- oder Beckenschlag beeinflusst. Aus der chronologische Anordnung der Pulse ergibt sich eine Pulsfolge,welche wiederum das Zeitmaß, insbesondere Metrum und Takt,

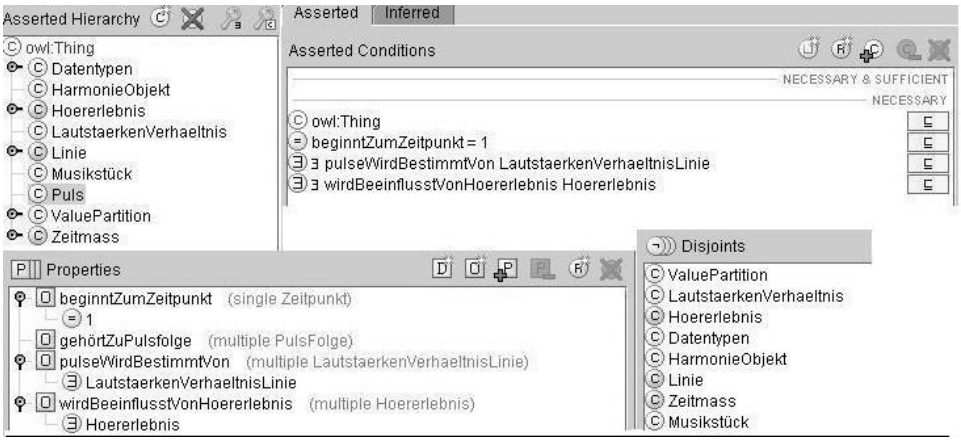


Abbildung 6.7: Die Klasse Puls.

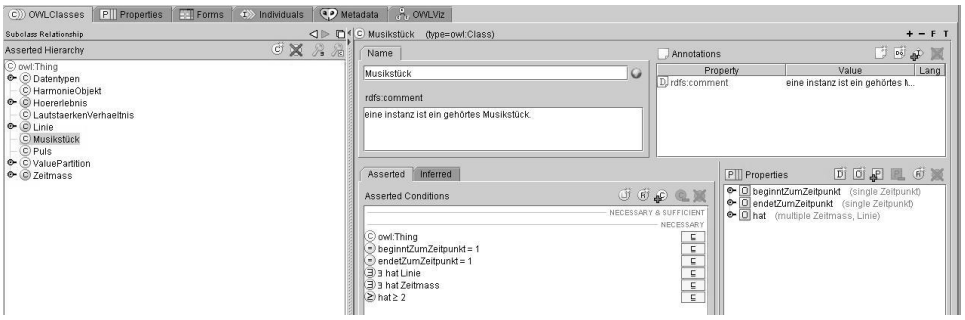


Abbildung 6.8: Die Klasse Musikstück.

beeinflusst.

6.3.6 Musikstück

Die Klasse Musikstück (siehe Abb. 6.8) beschreibt ein gehörtes Musikstück. Es ist allerdings schwierig, grundlegende Gemeinsamkeiten von Musikstücken festzulegen, da sie in der Regel sehr unterschiedlich gestaltet sind. Die Musikstücke, welche mit der Ontologie modelliert werden sollen, müssen eine gewisse Zeitdauer aufweisen, womit ein Start- und Endzeitpunkt bestimmt werden kann. Ferner wird das Vorhandensein von mindestens einem Zeitmaß sowie einer Linie (Melodie-, Begleit- oder Harmonielinie) vorausgesetzt. Das lässt sich damit begründen, dass einfache Lieder meistens alle eine Melodielinie und ein Metrum aufweisen. Hintergrundmusik, bei der eine Melodielinie fehlt, besteht im Gegensatz dazu aus einer oder mehreren Begleitlinie und einem Metrum. Selbst das Aneinanderreihen von Akkorden (Harmonielinie) kann als Musik angesehen werden.

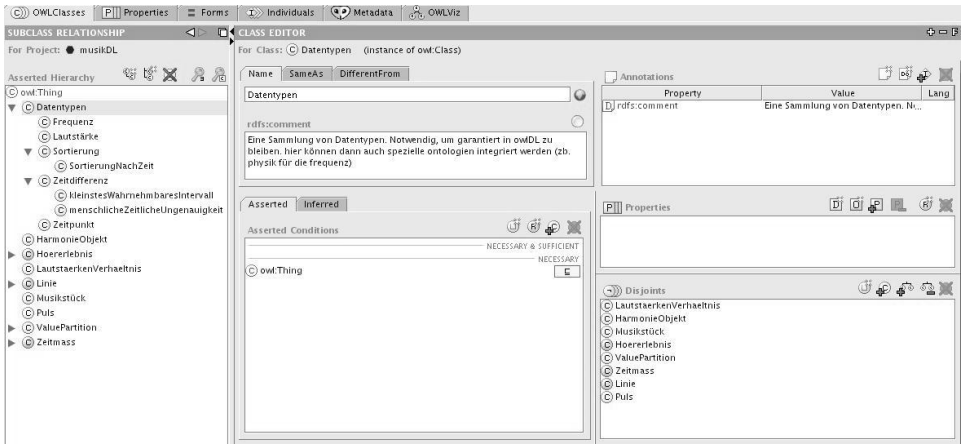


Abbildung 6.9: Die Klasse Datentypen.

6.3.7 Datentypen

Die Klasse Datentypen (siehe Abb. 6.9) aggregiert grundlegende Datentypen, welche von den übrigen Klassen durch unterschiedliche Relationen genutzt werden können. Diese Vorgehensweise ermöglicht einen einfachen Austausch der verwendeten Datentypen. Die Frequenz wird momentan beispielsweise durch eine einfache Zahl dargestellt. Sie könnte jedoch durchaus komplexer modelliert werden, falls eine genauere Definition in einer anderen Ontologie gegeben wäre, welche dann einfach den Datentyp „Frequenz“ in dieser Klasse ersetzt. Somit ist eine Erweiterbarkeit der Ontologie gegeben. Folgende Datentypen sind in der Klasse *Datentypen* enthalten:

- Frequenz
- Lautstärke
- Zeitpunkt
- Zeitdifferenz
- Sortierung

Die *Frequenz* zeigt die Frequenz eines Klanges in Hertz an. Je nach verwendeter Relation werden damit die Haupt- oder Nebenfrequenzen eines Klanges dargestellt.

Die *Lautstärke* gibt die Lautstärke eines Klanges in *dB* an. Die fest definierte Lautstärke *Unhörbar* dient hierbei zur Modellierung von Pausen.

Die Zeitdifferenz gliedert sich in zwei weitere Unterklassen auf:

1. *KleinstesWahrnehmbaresIntervall* beschreibt das kleinste Zeitintervall, welches Menschen noch als solches empfinden können.

2. *MenschlicheZeitlicheUngenauigkeit* gibt die Abweichung zweier Zeitintervalle an, bis zu der sie vom Menschen als regelmäßig empfunden werden. Hierfür existiert kein exakter Wert, denn diese Abweichung unterscheidet sich bei jedem Menschen geringfügig.

Die *Sortierung* ist eine Hilfskonstruktion, welches es ermöglicht, Inhalte bestimmter Klassen nach verschiedenen Kriterien zu sortieren. Die chronologische Sortierung von Klangereignissen ist die einzige, für diese Ontologie relevante Sortierung, weshalb hierzu eine entsprechende Unterklasse existiert.

6.3.8 Implementierung der Klasse Hörerlebnis

Anhand der Klasse *Hörerlebnis* (siehe Abb. 6.2) soll exemplarisch die Implementierung der Ontologie mit Protégé in OWL gezeigt werden. Da ein Hörerlebnis alles hörbare darstellt, überdeckt es sämtliche Unterklassen. Dieser Sachverhalt wird durch das Covering Axiom verdeutlicht:

$$\cup \text{ Geräusch } \cup \text{ Klang } \cup \text{ Pause}$$

Relationen, welche zu der Domäne eines Hörerlebnisses gehören, werden im folgenden dargestellt. Die kursiv gedruckten Klassen rechts der Pfeile stellen den Bildbereich der Relationen dar.

- beginntZumZeitpunkt $\longrightarrow$ *Zeitpunkt*
- endetZumZeitpunkt $\longrightarrow$ *Zeitpunkt*
- hatFrequenz $\longrightarrow$ *Frequenz*
- hatLautstärke $\longrightarrow$ *Lautstärke*
- höererlebnisBeeinflussPuls $\longrightarrow$ *Puls*

Die Tatsache, dass jedes Hörerlebnis genau einen Startpunkt aufweist, wird durch eine Kardinalitätsrestriktion auf der Relation *beginntZumZeitpunkt* ausgedrückt:

$$= \text{ beginntZumZeitpunkt } = 1$$

Die Unterklasse *Klang* verwendet das Designmuster *ValuePartition* (siehe Abschnitt 2.3), um zu erkennen, ob eine Klang ein Melodie- oder Begleitklang ist. Dazu wurde eine MelodyValuePartition, mit den zwei Unterklassen JA und NEIN (siehe Abb. 6.10), gebildet. Die Relation *gehörtZuMelodie* verbindet einen *Klang* mit einer *MelodyValuePartition*. Durch die Restriktion

$$\exists \text{ gehörtZuMelodie } JA$$

wird ausgedrückt, dass ein *Klang* zu einer Melodie gehört. Um einen *Begleitklang* darzustellen wird analog dazu folgende Restriktion verwendet:

$$\exists \text{ gehörtZuMelodie } NEIN$$

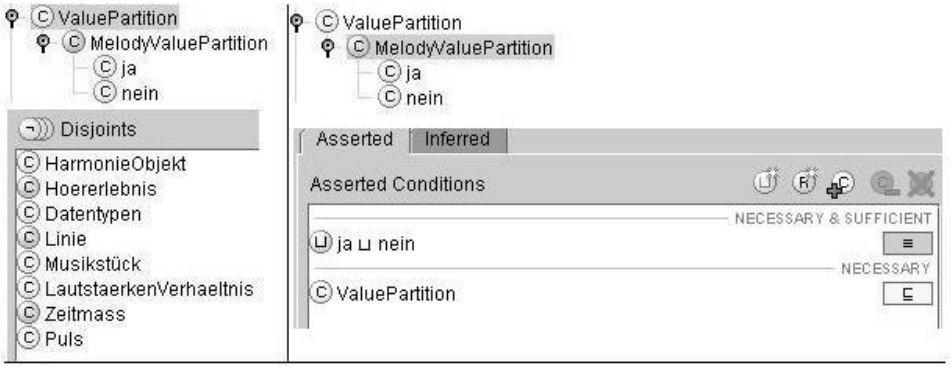


Abbildung 6.10: Designmuster ValuePartition.

Insgesamt besteht die Ontologie in ihrer jetzigen Form aus 140 Klassen und ebensovielen Relationen und Attributen. Eine OWL-Implementierung der Klasse Klang ist in Abb. 6.11 gezeigt.

6.4 Distanz zwischen Thema und Variation

Zur Validierung des in Kapitel 5 vorgestellten Distanzmaßes wurden verschiedene Musikstücke mittels der Ontologie formalisiert und miteinander verglichen. Die ausgewählten Musikstücke sind das Thema der Klaviersonate in A-Dur (KV 331) von Wolfgang Amadeus Mozart, sowie zwei Variationen hiervon und das Rondo Alla Turca. Das Thema der Sonate soll mit der ersten und vierten Variation des Themas verglichen werden. Da musikalisch das Thema in den Variationen in verschiedener Art und Weise verarbeitet wird, sollte das Distanzmaß Ähnlichkeiten zwischen den Stücken erkennen. Ein Vergleichsstück, der türkische Tanz, welcher nicht zur Sonate gehört, sollte hingegen keine Ähnlichkeit mit dem Thema aufweisen. Für einen Vergleich der Stücke untereinander wurde jeweils eine Periode, das sind in den betrachteten Musikstücken acht Takte, ausgewählt. Eine Periode bezeichnet in der Musikwissenschaft eine Struktureinheit in Ablauf und Gliederung eines Musikstücks und eignet sich damit zum Vergleich der Stücke.

6.4.1 Strukturierungskriterien

Zur Strukturierung der Musikstücke durch die relative Entropie wurden bestimmte Aspekte der Musikstücke durch die Ontologie formalisiert und näher untersucht. Insbesondere wurden die Melodie-, Harmonie- und Begleitlinien näher betrachtet. Die Melodielinie wurde anhand der Melodieführung, des Pulses und der Noten in der Melodie bzw. deren Umspielung in den Variationen analysiert. Die Harmonien, deren Pulse und die tatsächlich verwendeten

```
<owl:Class rdf:ID="Klang">
    <owl:disjointWith>
      <owl:Class rdf:ID="Geraeusch"/>
    </owl:disjointWith>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:someValuesFrom>
          <owl:Class rdf:about="#Frequenz"/>
        </owl:someValuesFrom>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="hatHauptFrequenz"/>
        </owl:onProperty>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#Hoererlebnis"/>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:cardinality rdf:datatype="int">1</owl:cardinality>
        <owl:onProperty>
          <owl:ObjectProperty rdf:about="#hatHauptFrequenz"/>
        </owl:onProperty>
      </owl:Restriction>
    </rdfs:subClassOf>
    <owl:disjointWith>
      <owl:Class rdf:ID="Pause"/>
    </owl:disjointWith>
  </owl:Class>
```

Abbildung 6.11: Darstellung eines Ausschnittes der Klasse *Klang* in OWL Syntax.

Noten in den Akkorden dienten zur Untersuchung der Begleitlinie. Für weitere Informationen zur Analyse eines Musikstückes siehe [Frisius 1984].

Melodieführung

Die Melodieführung betrachtet die Auf- und Abwärtsbewegungen sowie die gleichbleibende Fortführung einer Melodie. Für den Vergleich der Musikstücke wird die Anzahl der jeweiligen Melodiebewegungen pro Takt betrachtet, immer in Relation zu der Gesamtzahl der erklingenden Töne. Falls in einem Takt nur wenige Töne vorkommen, sind die Gestaltungsmöglichkeiten hinsichtlich der Melodieführung geringer als bei dem Vorhandensein vieler Töne. Die Untersuchung dieses Aspektes einer Melodielinie erlaubt somit eine Aussage über die Ähnlichkeit des Verlaufes der Melodien. Die Melodieführung ist als einziges Kriterium nicht direkt durch eine Klasse modellierbar, sondern ergibt sich indirekt aus dem Tonhöhenunterschied zweier aufeinanderfolgender Klänge.

Melodietöne

Hier werden die Töne des Themas und deren Umspielung in den Variationen untersucht. Dazu wird der relative Anteil der melodiegebenden Töne des Themas pro Takt betrachtet und mit den Variationen bzw. der Vergleichsperiode aus Alla Turca verglichen. Dieser Vergleich gibt an, welche Noten des Themas in den Variationen wiederverwendet, hinzugekommen oder weggelassen wurden.

Puls

Die Analyse des Pulses geschieht bei der Melodie- als auch bei der Begleitlinie gleich. Es werden hierbei die Anschläge pro Takt gezählt, d.h. wieviele Noten insgesamt pro Takt die Melodie bzw. die Begleitlinie bilden. Die musikalische Aussagekraft des Pulses ist aufgrund seiner Einfachheit nicht so hoch wie die der anderen, untersuchten Aspekte.

Harmonien

Die Harmonien stellen die relative Anzahl und Art der Akkorde pro Takt dar. Ähnliche Akkorde, beispielsweise ein Dominant- und Dominant-Sept Akkord, werden von dem Distanzmaß als völlig unterschiedlich bewertet, was bei einer weitergehenden Strukturierung berücksichtigt werden müsste.

Töne der Begleitung

Bei der Analyse der Begleittöne wird ähnlich wie bei den Melodietönen vorgegangen. Es wird untersucht, welche Töne des Akkordes wirklich erklingen

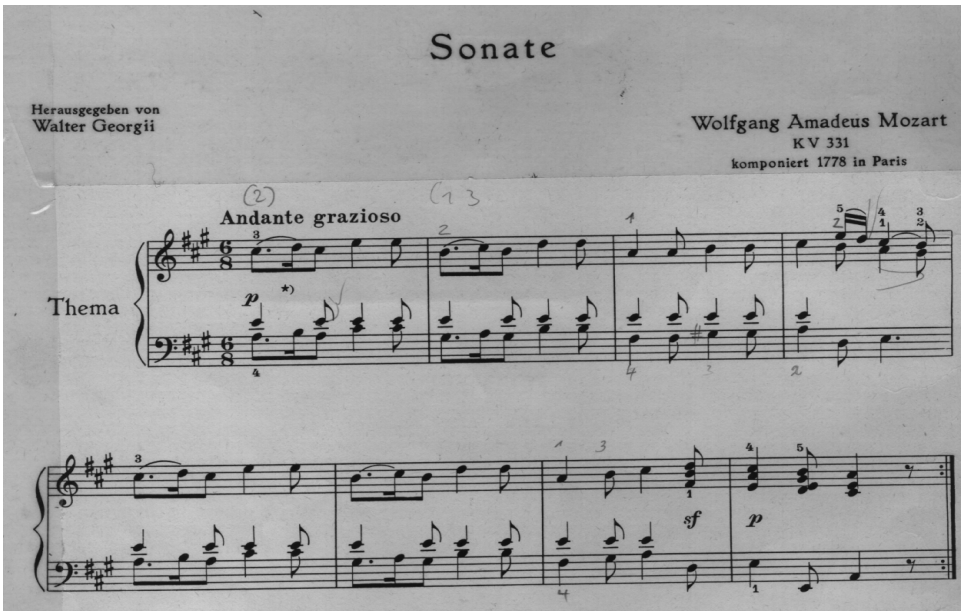


Abbildung 6.12: Erste Periode des Themas der Klaviersonate in A-Dur von Wolfgang Amadeus Mozart.

und ob Übergangstöne zwischen den Akkorden eines Taktes existieren, falls mehrere Akkorde in einem Takt vorhanden sind.

6.4.2 Formalisierung

In Abb. 6.12 ist die erste Periode des Themas der Sonate in A-Dur von Wolfgang Amadeus Mozart dargestellt. Aus Gründen der Übersichtlichkeit sind in Abb. 6.13 nur die ersten vier Takte der Formalisierung des Stückes mittels der Ontologie für die menschliche, musikalische Wahrnehmung dargestellt. Ebenso wurde auf eine Darstellung der Instanzen der Klasse *Puls* sowie der Melodieführung verzichtet. Der Wurzelknoten der Ontologie ist eine Instanz der Klasse *Musikstück*, welche das Thema der Sonate in A-Dur verkörpert. Für das Thema wurden die *Melodielinie*, die *Harmonielinie* und die *Begleitline* formalisiert, aus denen das Musikstück aufgebaut ist. Die dargestellten Instanzen gehören zu den gleichnamigen Klassen Melodielinie, Harmonilinie und Begleitlinie.

Den Strukturierungskriterien entsprechend wurden bei der Darstellung der Melodielinie die vorkommenden Melodieklänge gleicher Tonhöhe der Einfachheit halber zu einer Instanz zusammengefasst. In dem ersten Takt des Themas kommen beispielsweise die Töne *cis*, *d* und *e* vor, so dass sie jeweils eine Instanz der Klasse *Melodieklang* bilden. Mit der Relation *hatMelodie-Klang* wird angezeigt, dass sie der Melodie des Themas angehören. Der erste

Teil des Namens der Instanz zeigt dabei den konkreten Ton an, der letzte Teil, in welchem Takt er vorkommt, so dass gleiche Töne, welche in unterschiedlichen Takten vorkommen eindeutig zu unterscheiden sind und visuell schnell erfassbar sind. Die konkreten Tonhöhen und andere wichtige Parameter werden durch Attribute der Instanzen repräsentiert, welche aber aus Übersichtlichkeitsgründen in der graphischen Darstellung weggelassen wurden. Die Relation *gehörtZuTakt* gibt an, in welchem Takt die Töne vorkommen. Die Bedeutung dieser Relation ist für alle weiteren Instanzen gleich, d.h. sie zeigt die Zugehörigkeit zu einem bestimmten Takt an, was eine einfache zeitliche Sortierung darstellt.

Die Instanzen der Harmonielinie, welche zur Klasse *Harmoniebjekt* gehören, geben die in dem Stück vorkommenden Harmonien an, dargestellt durch die Relation *hatHarmonie*. Die Beschriftung der Instanzen für den letzten Teil des Namens gleicht dem der Melodielinie. Der erste Teil zeigt an, welche Harmonie erklingt, wobei die Akkordnotation aus der Musiktheorie übernommen wurde [Frisius 1984]. Großbuchstaben zeigen einen Dur-Akkord an, Kleinbuchstaben einen Moll-Akkord und die Zahl *7* direkt hinter einem Akkordnamen weist darauf hin, dass es sich um einen Sept-Akkord handelt.

Die Darstellung der Begleitlinie erfolgt analog zu derjenigen der Melodielinie. Es werden ebenfalls die vorkommenden Noten des Akkordes und mögliche Durchgangsnoten zusammenfassend für jeden Takt dargestellt. Die Instanzen gehören hier also der Klasse *Begleitklang* an. Der Buchstabe „B“ zwischen dem ersten und letzten Teil gibt an, dass die Töne zur Begleitung gehören. In Abb. 6.14 und Abb. 6.15 ist das Notenbild der ersten Variation des Themas der Sonate sowie die dazugehörige Formalisierung durch die Ontologie dargestellt. Abb. 6.16 stellt die vierte Variation des Themas dar. Die Formalisierung ist in Abb. 6.17 zu sehen. Die Vergleichsperiode aus dem Mittelteil von Alla Turca gibt Abb. 6.18 wieder, die dazugehörige Formalisierung durch die Ontologie findet sich in Abb. 6.19 wieder.

6.4.3 Ergebnisse

Die relative Entropie wurde auf die im vorigen Abschnitt beschriebenen Ontologien angewandt, um die Distanz zwischen dem Thema der Sonate und den Variationen bzw. einer Vergleichsperiode aus Alla Turca zu bestimmen. Die relative Entropie gibt in diesem Anwendungfall die durchschnittliche Information in Bits an, dass das Thema mit Wahrscheinlichkeitsverteilungen $\mathbf{p}$ nicht in den Variationen mit Wahrscheinlichkeitsverteilungen $\mathbf{q}$ verarbeitet wird. Eine kleine Distanz bedeutet somit eine weitgehende Verarbeitung des Themas in den Variationen. Falls keine Verarbeitung des Themas erfolgte, ergibt sich eine große Distanz.

Der Vergleich der Musikstücke wurde taktweise, jeweils für ein in Abschnitt 6.4.1 beschriebenes Kriterium, durchgeführt und die sich daraus ergebenden Distanzen für die gesamte Periode aufsummiert. Falls in den Va-

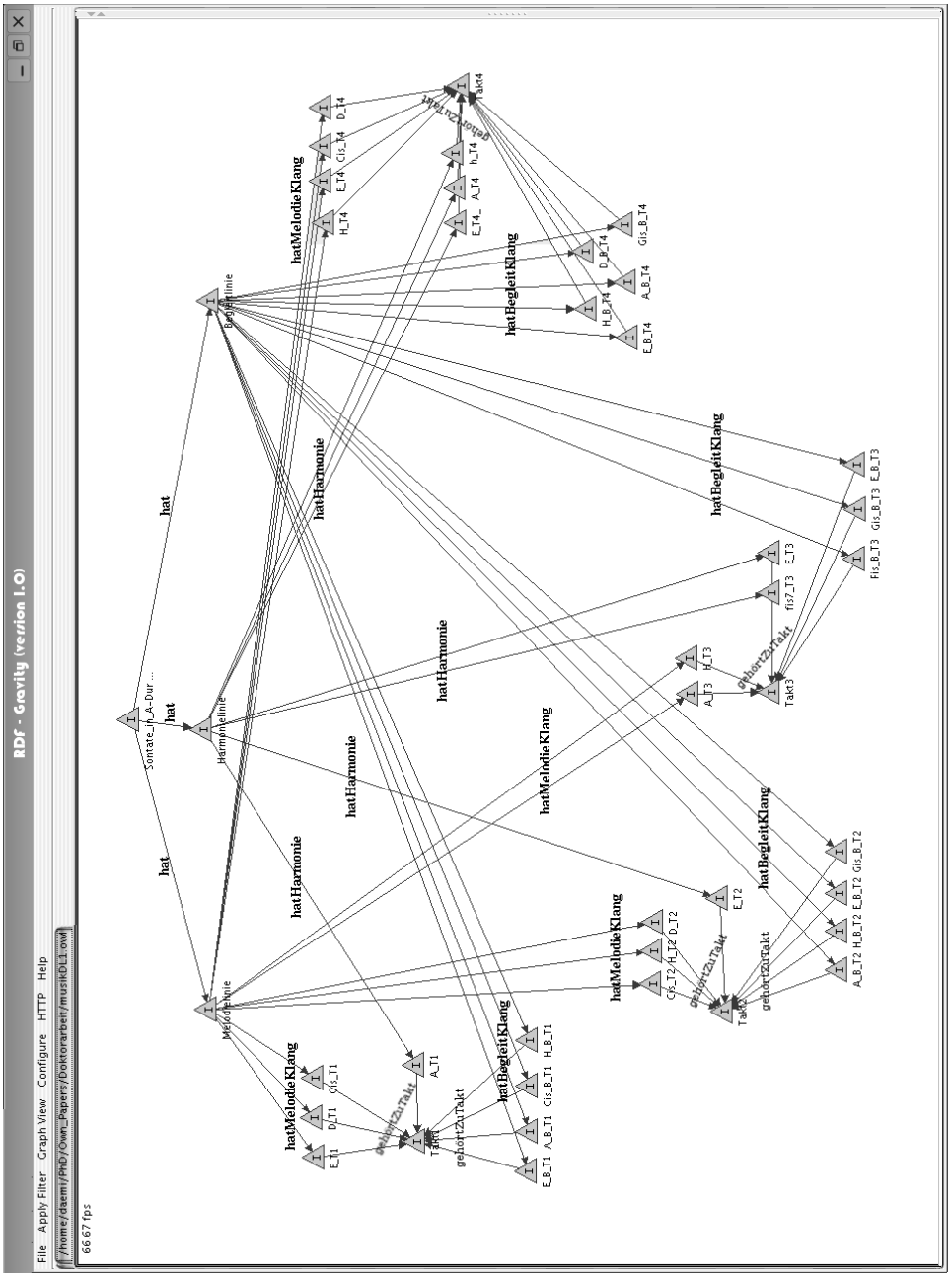


Abbildung 6.13: Formalisierung des Vordersatzes des Themas der Klaviersonate in A-Dur durch die Ontologie für musikalische Wahrnehmung. Aus Übersichtlichkeitsgründen sind nur die Instanzen der ersten vier Takte dargestellt. Weggelassen wurden die Attribute der Instanzen und die dazugehörigen Klassen.

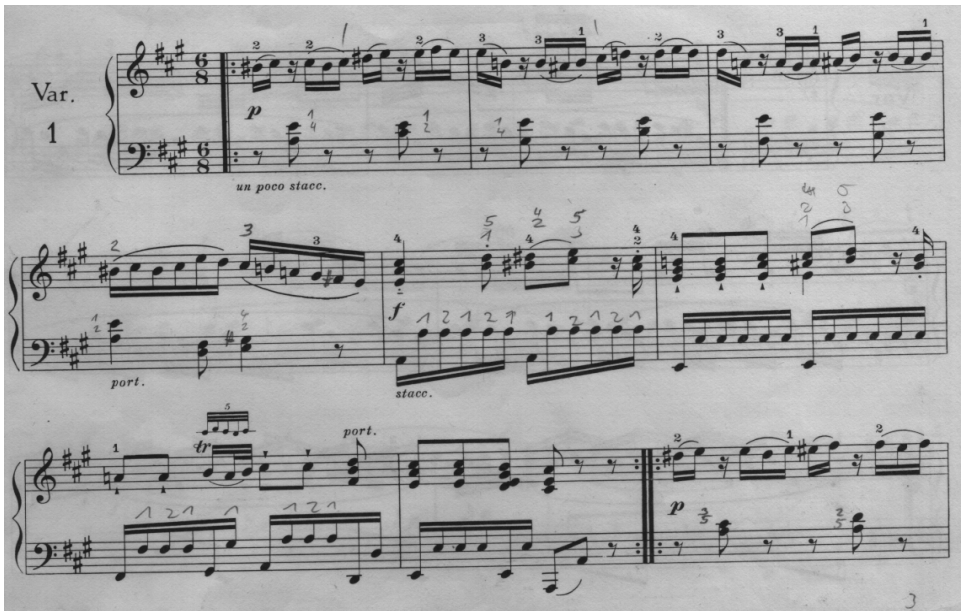


Abbildung 6.14: Erste Variation des Themas der Sonate in A-Dur von Wolfgang Amadeus Mozart.

riationen und der Periode aus Alla Turca die relative Häufigkeit eines Merkmales nicht vorhanden, also Null war, wurde die Wahrscheinlichkeit auf den Wert 10^{-5} gesetzt. Die Annäherung an Null ist für die betrachtete Anwendung ausreichend, da der Ereignisraum pro Takt mit maximal 25 Ereignissen im Vergleich klein ist. Damit wird eine Berechnung der relativen Entropie garantiert, und gleichzeitig eine hohe Distanz erreicht, falls ein Merkmal in $\mathbf{p}$ aber nicht in $\mathbf{q}$ vorhanden ist. Die Distanz beträgt ca. 16 bits, falls $\mathbf{p}$ aus einem Ereignis mit Wahrscheinlichkeit 1 besteht, und das entsprechende Ereignis in $\mathbf{q}$ mit 10^{-5} näherungsweise angegeben wird.

Eine beispielhafte Darstellung für den Vergleich der Melodietöne findet sich in Abb. 6.20. Die Gewichte an den Kanten der Relation *hatMelodie-Klang* zeigen die relative Häufigkeit des Tones in dem dazugehörigen Takt an. Pro Takt summieren sich die einzelnen Gewichte zu eins, da die Distanzen taktweise berechnet werden. Die schwarz beschrifteten Gewichte repräsentieren dabei die Häufigkeiten der Klänge im Thema, die grün beschrifteten in diesem Beispiel die Häufigkeiten der Melodieklänge in Variation vier. Die Gewichte der Variation müssen sich hierbei nicht auf eins summieren, da die gezeigte Ontologie das Thema formalisiert und in der Variation unter Umständen zusätzliche Töne mit dazugehöriger Häufigkeit hinzugekommen sein können. Da die zusätzlichen Töne im Thema $\mathbf{p}$ nicht vorkommen, ergibt sich

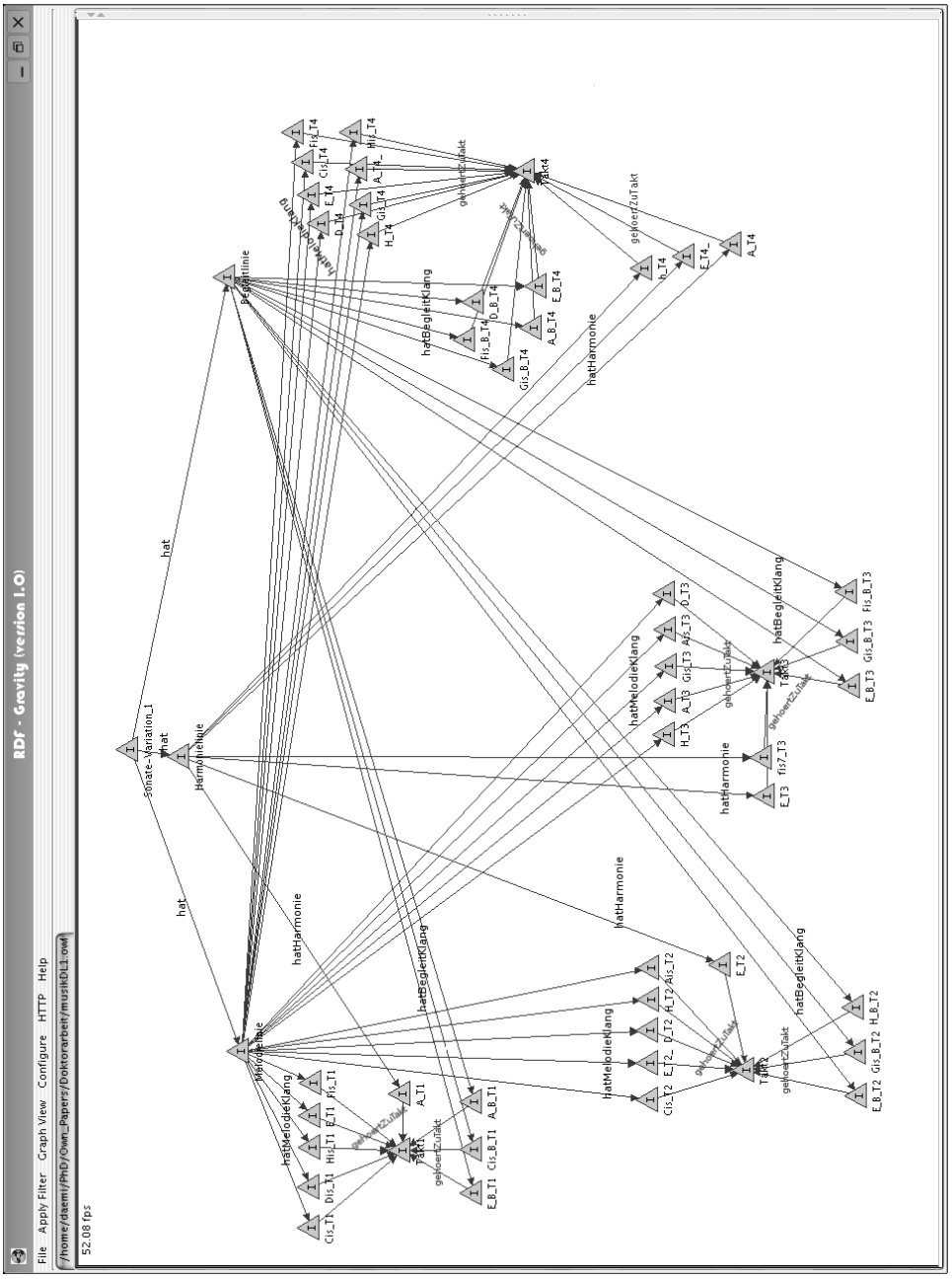


Abbildung 6.15: Formalisierung des Vordersatzes der ersten Variation des Themas der Sonate in A-Dur durch die Ontologie für musikalische Wahrnehmung. Aus Übersichtlichkeitsgründen sind nur die Instanzen der ersten vier Takte dargestellt. Weggelassen wurden die Attribute der Instanzen und die dazugehörigen Klassen.

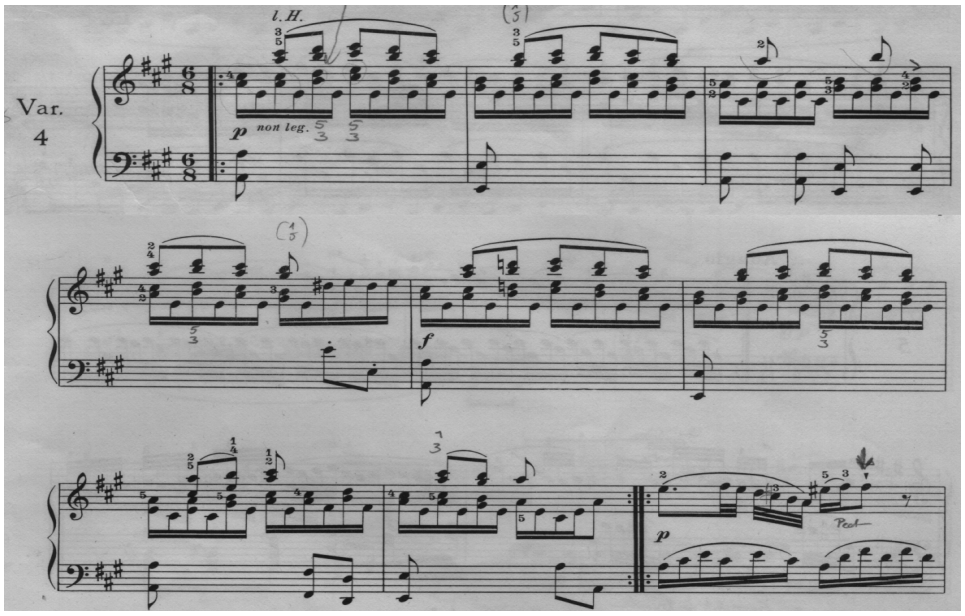


Abbildung 6.16: Vierte Variation des Themas der Sonate in A-Dur von Wolfgang Amadeus Mozart.

$p(x) = 0$, wobei x der zusätzliche Ton in der Variation sei. Mit

$$\lim_{x \to 0} x \log x = 0$$

folgt, dass die relative Entropie für diesen Ton Null ist. Die Relation *gehört-ZuTakt* wird mit keinem Gewicht versehen, denn sie zeigt die Zugehörigkeit der Töne an, was kein Vergleichskriterium war. Ebensowenig sind Gewichte an den drei *hat* Relationen zu finden, da für die Linien kein Vergleichskriterium festgelegt wurde. Analog zu Variation vier wurde Variation eins sowie die Vergleichsperiode aus Alla Turca dargestellt und verglichen. Die Wahrscheinlichkeitsverteilungen pro Takt für die einzelnen Kriterien und die dazugehörigen Distanzen finden sich in übersichtlicher, tabellarischer Form in Anhang A.

Die Strukturierung der Ontologien anhand der Melodietöne zwischen Thema der Sonate und den Variationen eins und vier, sowie dem Ausschnitt aus Alla Turca, ergab folgende Distanzen:

$$D_{Melodieklang}(Thema||Variation_1) = 5.86$$

$$D_{Melodieklang}(Thema||Variation_4) = 2.19$$

$$D_{Melodieklang}(Thema||AllaTurca) = 61.17$$

Die erste Variation umspielt die Töne der Melodie des Themas mit Sechzehnteln, kommt dabei aber immer wieder auf die Melodietöne zurück. Die vierte

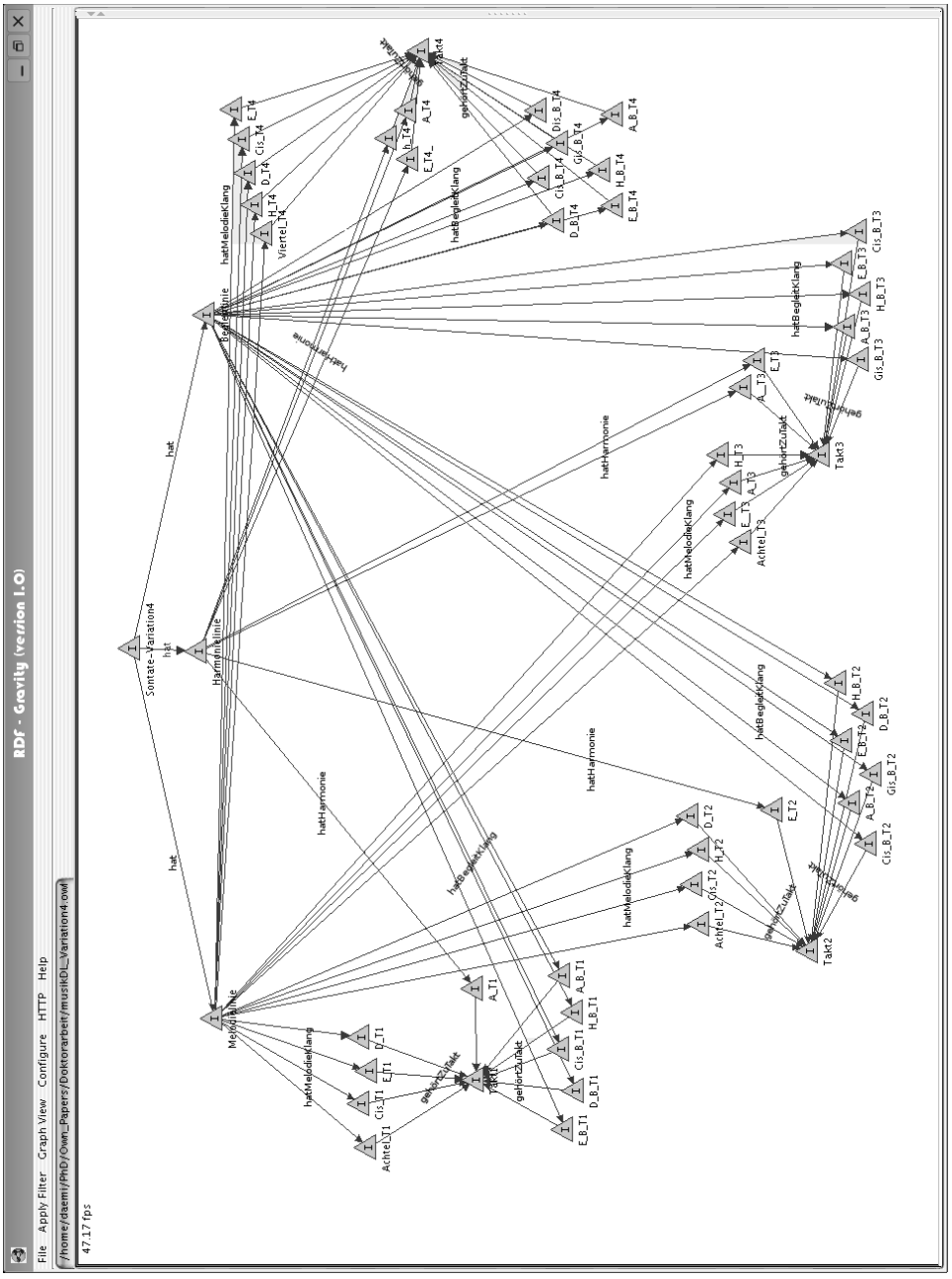


Abbildung 6.17: Formalisierung des Vordersatzes der vierten Variation des Themas der Sonate in A-Dur durch die Ontologie für musikalische Wahrnehmung.Aus Übersichtlichkeitsgründen sind nur die Instanzen der ersten vier Takte dargestellt. Weggelassen wurden die Attribute der Instanzen und die dazugehörigen Klassen.

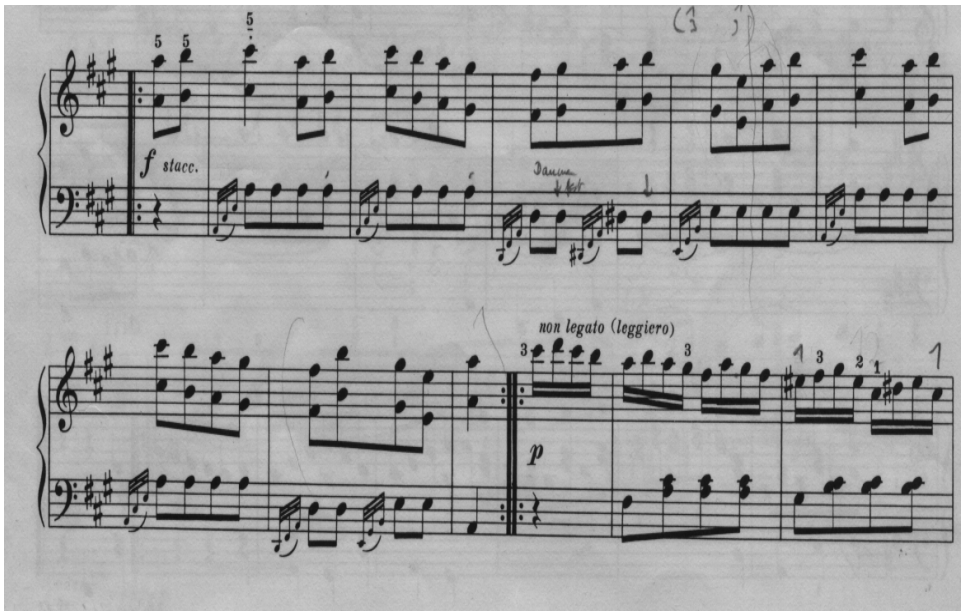


Abbildung 6.18: Eine Periode des Türkischen Tanzes von Wolfgang Amadeus Mozart als Vergleichsstück.

Variation weist im wesentlichen die gleichen Melodietöne wie das Thema auf, und umspielt diese nur mit sehr wenigen Durchgangstönen, weshalb hier die Distanz zur Melodie des Thema geringer ist als bei Variation eins. Der türkische Tanz (Alla Turca) weist hingegen keine Ähnlichkeit mit den Tönen der Melodie auf. Aufgrund der gleichen Tonart erklingen aber in einigen Takten des Rondos Töne der Themamelodie.

Ein Vergleich der Melodieführung, d.h. der tonalen Aufwärts-, Abwärts- und gleichbleibenden Bewegungen der Melodienoten ergab ein anderes Bild:

$$D_{Melodiefuehrung}(Thema||Variation_1) = 6.16$$

$$D_{Melodiefuehrung}(Thema||Variation_4) = 16.03$$

$$D_{Melodiefuehrung}(Thema||AllaTurca) = 57.48$$

Aufgrund der Sechzehntelbewegungen in der ersten Variation ergeben sich für die Melodiebewegung sehr viele Möglichkeiten. Sie folgt jedoch größtenteils der Melodiebewegung des Themas, wobei die gleichbleibenden Weiterbewegungen der Melodie meistens umspielt werden, was eine Vergrößerung der Distanz bewirkt. Die Melodiebewegung in der vierten Variation unterscheidet sich dagegen erheblich von derjenigen des Themas, weil insbesondere in den ersten beiden Takten des Vorder- und Nachsatzes nur eine Auf- und Abwärtsbewegung der Melodie stattfindet. Das Thema hingegen umfasst noch

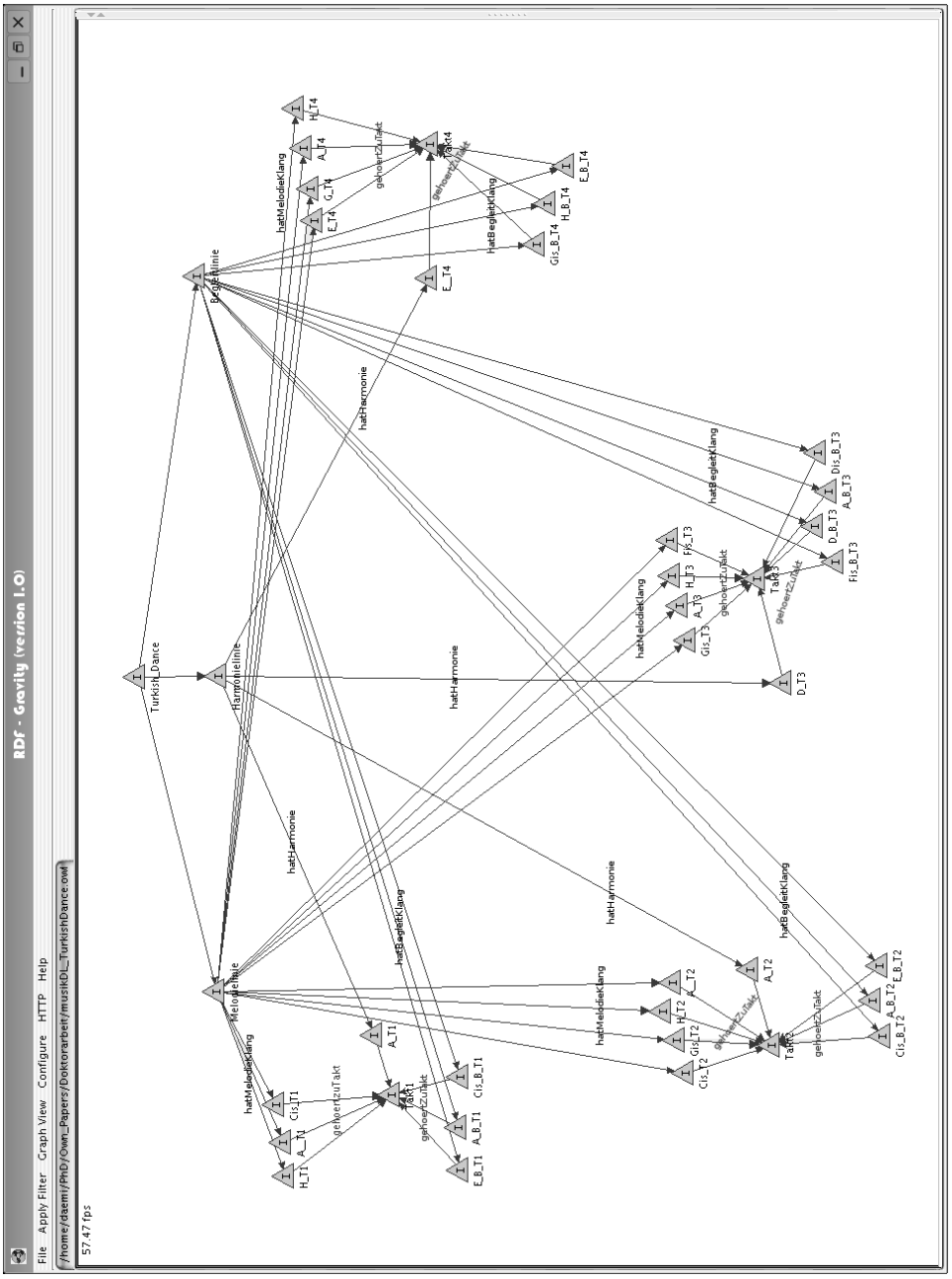


Abbildung 6.19: Formalisierung einer Periode des Türkischen Tanzes durch die Ontologie für musikalische Wahrnehmung. Aus Übersichtlichkeitsgründen sind nur die Instanzen der ersten vier Takte dargestellt. Weggelassen wurden die Attribute der Instanzen und die dazugehörigen Klassen.

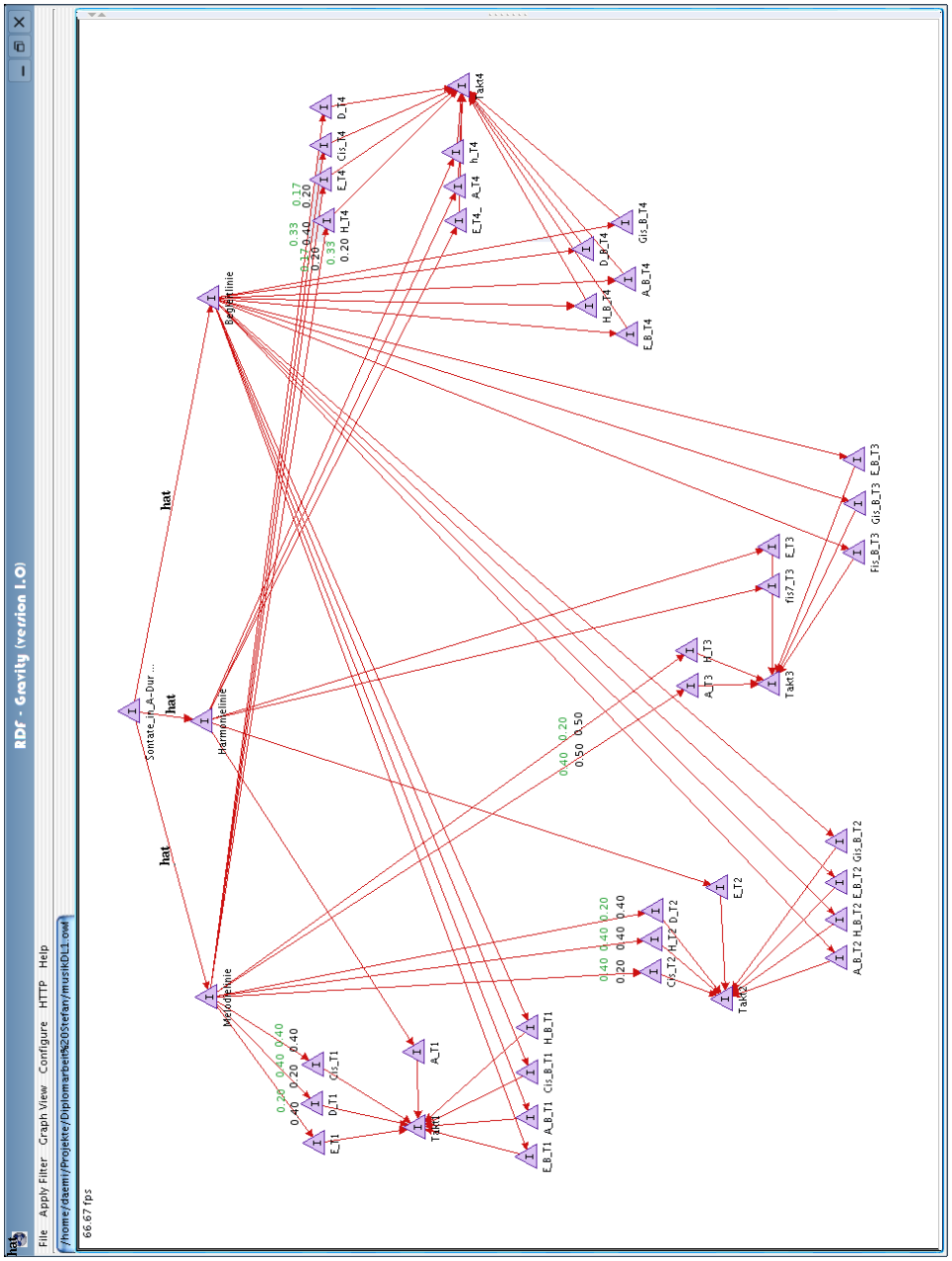


Abbildung 6.20: Gewichtung der relativen Häufigkeiten der Melodietöne pro Takt. Die schwarzen Zahlen geben die relative Häufigkeit der Melodietöne desThemas **p** an, die grünen Zahlen diejenigen der vierten Variation **q**. Analog wurden die Häufigkeiten für die anderen Relationen bzw. Strukturierungskriterien dargestellt.

eine zusätzliche Aufwärts- sowie Gleichbleibende Bewegung. Auch hier findet sich keine Ähnlichkeit der Melodiebewegung des Themas in dem türkischen Tanz wieder, da es pro Takt, bis auf eine Ausnahme, entweder nur Auf- oder Abwärtsbewegungen gibt.

Die letzte Vergleichsmöglichkeit für die Melodie betrachtet die relative Anzahl der Pulse pro Takt. Damit ergibt sich für die erste Variation eine größere Distanz als für die vierte, da die erste sehr viele Sechzehnteltöne aufweist. In der vierten Variation ist dagegen die Anzahl der Pulse der Melodie denen des Themas sehr ähnlich. Auf einen Vergleich mit Alla Turca wurde verzichtet, da das Grundmetrum bereits verschieden ist.

$$D_{MelodiePuls}(Thema||Variation_1) = 3.36$$

$$D_{MelodiePuls}(Thema||Variation_4) = 0.13$$

Zum Vergleich der Begleitung wurde zunächst die Harmonielinie der Stücke betrachtet. Hier ergab sich kein Unterschied der ersten Variation zum Thema, da die gleichen Akkorde erklingen. Der Unterschied zur vierten Variation erscheint recht groß, jedoch hat sich nur eine Harmonie geändert. Die große Änderung resultiert daher, dass der Ereignisraum des verantwortlichen Taktes aus nur zwei Akkorden besteht. Das Nichtvorhandensein eines Akkordes löst demzufolge bereits eine große Änderung aus. In Alla Turca sind dagegen bis auf zwei Takte sämtliche Harmonien unterschiedlich:

$$D_{Harmonie}(Thema||Variation_1) = 0.00$$

$$D_{Harmonie}(Thema||Variation_4) = 7.80$$

$$D_{Harmonie}(Thema||AllaTurca) = 75.06$$

Im nächsten Schritt wurde untersucht, welche Töne der Harmonien angespielt wurden und wie oft. Hier ergaben sich für beide Variationen ähnliche Distanzen. Die Ähnlichkeit rührt vor allem daher, dass in beiden Variationen gebrochene Akkorde verwendet werden. Ein Großteil des Unterschiedes zwischen dem Thema und der vierten Variation lässt sich auf eine unterschiedliche Harmonie in Takt drei zurückführen, da diese zum Teil aus anderen Tönen besteht. In der ersten Variation wurde hingegen in Takt sieben der Grundton eines Akkordes weggelassen, weshalb sich in diesem Takt im Vergleich zu den anderen eine große Distanz ergibt. Der markante Unterschied zu Alla Turca bleibt bestehen, wobei in einigen Takten aufgrund der gleichen Tonart nur moderate Unterschiede auftreten:

$$D_{Begleitklang}(Thema||Variation_1) = 9.20$$

$$D_{Begleitklang}(Thema||Variation_4) = 9.05$$

$$D_{Begleitklang}(Thema||AllaTurca) = 49.04$$

Der letzte Vergleich befasst sich, wie bei dem Vergleich der Melodielinie, mit dem Puls der Begleitung. Die Unterschiede hierbei sind vor allem auf die gebrochenen Akkorde im sechzehntel Rhythmus zurückzuführen. In der ersten Variation werden die gebrochenen Akkorde erst im Nachsatz eingesetzt, so dass sich hier eine kleine Distanz ergibt. Aufgrund des unterschiedlichen Metrums wurde, wie bei der Melodie, auf einen Vergleich mit Alla Turca verzichtet.

$$D_{BegleitPuls}(Thema||Variation_1) = 5.90$$

$$D_{BegleitPuls}(Thema||Variation_4) = 11.69$$

Insgesamt betrachtet ergibt sich für den Vergleich der Melodien ein ähnlicher Wert der Distanz zwischen dem Thema und den Variationen. Die Distanz zu Alla Turca ist bedeutend größer, obwohl der Puls nicht berücksichtigt wurde.

$$D_{Melodie}(Thema||Variation_1) = 15.38$$

$$D_{Melodie}(Thema||Variation_4) = 18.30$$

$$D_{Melodie}(Thema||AllaTurca) = 118.65$$

Bei der Betrachtung der Begleitung ergibt sich ein ähnliches Bild. Die Distanz zwischen Thema und Variation vier ist hier allerdings größer als zwischen Thema und Variation eins, was sich zum Teil aus dem etwas verfälschenden Vergleich der Harmonien ergibt. Die Distanz zu Alla Turca ist ebenfalls wieder markant größer.

$$D_{Begleitung}(Thema||Variation_1) = 15.10$$

$$D_{Begleitung}(Thema||Variation_4) = 28.54$$

$$D_{Begleitung}(Thema||AllaTurca) = 124.10$$

Die Distanzen für die Melodie und Begleitung repräsentieren also zuverlässig den gegebenen, musikalischen Sachverhalt. Eine Verarbeitung des Themas erfolgt in den Variationen eins und vier, woraus sich die kleine Abstände ergeben. Eine großer Abstand zwischen Alla Turca und dem Thema zeigt keinerlei Verarbeitung des Themas in Alla Turca an, was den musikalischen Tatsachen entspricht.

6.5 Fazit

Die Vorgehensweise zur Darstellung der menschlichen, musikalischen Wahrnehmung mittels Ontologien stellt einen neuen Ansatz zur formalen Repräsentation von gehörter Musik (Transkription) dar. Bisherige Ansätze basieren auf Frequenzanalysen der Musikstücke und anschließenden Lernverfahren [Audi 2004]. Hierbei werden meistens lediglich einzelne Aspekte des Musikstückes analysiert, beispielsweise Melodie, Begleitung oder Rhythmus, oder

es bestehen Einschränkungen hinsichtlich der analysierten Musikstücke, d.h. es sind nur bestimmte Musikstücke zur Analyse mit dem betrachteten Verfahren geeignet.

Der vorgestellte Ansatz ermöglicht es, ein europäisches Musikstück mittels einer Ontologie vollstandig in einem Rechner zu repräsentieren. Dazu muss ein menschlicher Zuhörer, u.U. mit Hilfe einer Partitur, ein gehörtes Musikstück durch die Ontologie formalisieren. Mit dieser Darstellung können die Musikstücke anschließend mittels der relativen Entropie, wie gezeigt, effektiv und zuverlässig strukturiert werden. Die Strukturierungskriterien sind dabei frei wählbar, müssen aber durch explizite oder (automatisch) abgeleitete Relationen in der Ontologie darstellbar sein.

Mit einer Erweiterung der Ontologie sollte es auch möglich sein, den wichtigen Aspekt der *Emotionen* bei der Wahrnehmung von Musik mitaufzunehmen. Mit dieser Erweiterung und der Tatsache, dass durch Ontologien ein Kontext gegeben ist, aus dem eventuell eine Semantik ableitbar ist, kann ein erster Schritt zur Modellierung von Kreativität erfolgen. Kreativität ist hierbei definiert als das Hinzufügen oder Ändern von Konzepten, und damit auch Ideen, zu einer bestehenden Semantik.

Kapitel 7

Zusammenfassung und Ausblick

In dieser Arbeit wurden entropiebasierte Distanzmaße für ihre Anwendung zur Strukturierung von Ontologie bzw. des durch sie aufgespannten Wissensraumes näher untersucht. Dazu werden zunächst in Abschnitt 2.1 Ontologien als formale, explizite Konzeptualisierung einer bestimmten Wissensdomäne vorgestellt sowie deren mögliche Repräsentationen vorgestellt. Abschnitt 2.2 ging näher auf bekannte Strukturierungsmaße aus der Computerlinguistik und dem Information-Retrieval (IR) ein und grenzt diese von den untersuchten, entropiebasierten Distanzmaßen ab. In Abschnitt 2.3 wurden die historischen und theoretischen Grundlagen der Entropie vorgestellt, als auch die untersuchten, entropiebasierten Distanzmaße eingeführt.

In Kapitel 3 wurde ein theoretisches Modell zur Strukturierung von Wissen vorgeschlagen, dem vergleichend eine klassische Vorgehensweise zur Strukturierung von Daten vorangestellt wurde. Die vom Karlsruher Ansatz für Wissensforschung übernommene Definition von Information als Möglichkeitsausschluss im Wissenraum und Wissen als semantischer Information fügt sich gut in das vorgestellte Modell ein. Der Möglichkeitsausschluss wird vereinfachend Wahrscheinlichkeitsverteilung angesehen, welcher somit die Grundlage für eine weitergehende Strukturierung des Wissens bildet. Die Wahrscheinlichkeitsverteilungen müssen dabei nicht notwendigerweise eine objektive, d.h. frequenzbasierte Interpretation aufweisen, sondern können auch subjektive (degree-of-belief) Bedeutungen besitzen. Mit Hilfe der Wahrscheinlichkeitsverteilungen können entropiebasierte Distanzmaße definiert werden, welche für die Strukturierung des durch die Ontologien aufgespannten Wissensraumes untersucht wurden.

Als erstes entropiebasiertes Distanzmaß wurde die gegenseitige Information für ihre Anwendung auf Ontologien untersucht. Dafür musste zuerst die Bedeutung von Zufallsvariablen und die Zuweisung der Wahrscheinlichkeitsverteilungen zu den Konzepten der Ontologie definiert werden. Die be-

dingten Wahrscheinlichkeiten, welche für die Berechnung der gegenseitigen Information erforderlich sind, werden den Relationen, also Kanten zwischen den Konzepten, zugewiesen. Damit kann die gegenseitige Information als die Reduktion der Unsicherheit eines Konzeptes durch Kenntnis eines anderen, davon abhängigen Konzeptes, definiert werden. Die Definition wurde auf mehrere Konzepte erweitert und in einem letzten Schritt für die Anwendung auf Graphen, durch welche Ontologien dargestellt werden können, verfeinert. Hierzu wird nicht nur der kürzeste Pfad zwischen den Konzepten betrachtet, sondern auch längere Pfade, wobei diese einen hinreichend großen Informationsgewinn liefern müssen. Damit ist eine genauere Strukturierung von Ontologien möglich als bei bisherigen Maßen, weil diese immer nur die kürzesten Pfade zwischen Konzepten betrachten. Anschließend wurden noch zwei weitere Anwendungsmöglichkeiten der gegenseitigen Information auf Ontologien vorgestellt. Die erste ermöglicht eine optimale, informationstheoretische Strukturierung des Graphen. Mit Hilfe der zweiten, vorgestellten Anwendungsmöglichkeit kann das kleinste, gemeinsame Elternkonzept zwischen zwei Konzepten berechnet werden. Ein Problem bei der Verwendung der gegenseitigen Information als Distanzmaß sind die benötigten, bedingten Wahrscheinlichkeiten, welche sich aus den Verbundverteilungen berechnen lassen. Die Komplexität solcher Verteilungen ist bei großen Ontologien sehr hoch, was eine Berechnung der bedingten Wahrscheinlichkeiten schwierig gestaltet. Aus diesem Grund wurde in Kapitel 5 die einfacher zu berechnende relative Entropie als Strukturierungsmaß untersucht.

In Kapitel 4 wurde eine Ontologie für die Flutkatastrophenvorsorge (Flutontologie) erstellt, die in Zusammenarbeit mit dem Institut für Wasserwirtschaft und Kulturtechnik entstanden ist. Der Entwicklungprozess der Ontologie folgte einem Standardentwicklungsmodell und sie wurde in einer standardisierten, XML-basierten Sprache (DAML+OIL) implementiert. Die Flutontologie gewährleistet eine eindeutige Betrachtung der verschiedenen, beteiligten Wissensdomänen auf das Gebiet des Risikomanagements für Hochwässer. Damit ist eine einheitliche, semantische Grundlage für die Lokalisierung und Bereitstellung von domänenspezifischem Wissen für Wissenschaftler, welche in diesem Bereich tätig sind, geschaffen worden. Der Diskursbereich der Flutontologie muss dazu die grundlegenden Aspekte der Flutkatastrophenvorsorge sowie Präventionsmaßnahmen abdecken. Die wesentlichen meteorologischen als auch hydrologischen Prozesse, welche bei der Hochwasserbildung eine Rolle spielen, mussten für eine zuverlässige Repräsentation der Flutkatastrophenvorsorge ebenfalls mit in die Flutontologie aufgenommen werden.

In Kapitel 5 wurde die Anwendung der relativen Entropie auf Ontologien untersucht. Es wurden mehrere Gründe für die Benutzung der relativen Entropie als Strukturierungsmaß für Ontologien angegeben. Ein wichtiger Grund stellte die Ähnlichkeit der informationstheoretischen Definition der relativen Entropie mit der Wissensdefinition des Karlsruher Ansatzes als

Möglichkeitsausschluß im Wissensraum dar. Für die Benutzung der relativen Entropie als Distanzmaß müssen, wie bei der gegenseitigen Information, Wahrscheinlichkeitsverteilungen auf der Ontologie definiert werden. Die Bedeutung der Wahrscheinlichkeitsverteilungen wurde bewusst offen gelassen, sie müssen lediglich die Kolmogorov Axiome erfüllen. Die einzelnen Wahrscheinlichkeiten werden dann den Kanten zwischen den Konzepten, also den Relationen zwischen ihnen, zugewiesen.

Nachdem die Wahrscheinlichkeitsverteilungen zugewiesen wurden, kann eine Distanz zwischen den zu vergleichenden Ontologien bestimmt werden. Eine kleine Distanz bedeutet, dass sich die Ontologien ähnlich sind, eine große Distanz zeigt wesentliche Unterschiede an. Die Bedeutung der Distanz ergibt sich jeweils aus der Interpretation der Wahrscheinlichkeitsverteilungen und der zugrundeliegenden Ontologie. Generell lässt sich lediglich sagen, dass die relative Entropie die durchschnittliche Information angibt, mit der ein Ereignis aus einer Ontologie nicht aus der anderen Ontologie stammt. Für subjektive Wahrscheinlichkeiten wurde ein Beispiel anhand der Flutontologie gezeigt. Die Wahrscheinlichkeiten stellen hierbei die subjektiven Wichtigkeiten einzelner Konzepte bezüglich der Flutkatastrophenvorsorge dar. Die Wissenbasis der Ontologien sind also die Einschätzungen einzelner Personen oder Personenkreise. Die Distanz gibt in diesem Beispiel an, inwieweit das Wichtigkeiten eines Personenkreises, wie zum Beispiel der Experten, mit denjenigen anderer Personenkreise übereinstimmt. Bei einer großen Distanz wird das Ziel der Flutkatastrophenvorsorge nur noch suboptimal erreicht, da in diesem Beispiel die Einschätzung der Experten als optimal angesehen wurde. Ein Vorteil bei der Benutzung der relativen Entropie als Distanzmaß gegenüber anderen, informationstheoretischen Maßen ist ihre effektive und schnelle Berechnung. Eine Validierung des Distanzmaßes anhand von objektiven Wahrscheinlichkeitsverteilungen erfolgte anschließend in Kapitel 6.

In Kapitel 6 wurde eine Ontologie für die menschliche, musikalische Wahrnehmung erstellt. Diese ermöglicht es erstmals, ein gehörtes Musikstück anhand der menschlichen Wahrnehmung in einer formalen Datenstruktur zu modellieren. Zentrale Konzepte der menschlichen, musikalischen Wahrnehmung, welche durch die Ontologie modelliert werden sind Hörerlebnisse, Linien, Zeitmaße und Lautstärkenverhältnisse sowie Puls und Harmonien. Diese Konzepte werden in der Ontologie noch weiter verfeinert und in Beziehung zueinander gesetzt. Damit ist es möglich, eine grundlegende Formalisierung europäische Musik durchzuführen.

Mit Hilfe der beschriebenen Ontologie wurden einige Musikstücke formalisiert und mittels der relativen Entropie strukturiert. Die durch die Ontologie formalisierten Musikstücke waren das Thema der Klaviersonate in A-Dur von Wolfgang Amadeus Mozart, sowie die erste und vierte Variation dieses Themas und als weiteres Vergleichsstück das Rondo Alla Turca. Für die Anwendung der relativen Entropie wurden zunächst nach einfachen,

musikwissenschaftlichen Kriterien objektive Wahrscheinlichkeitsverteilungen erstellt, anhand derer Melodie, Begleitung, Harmonien und Puls der Musikstücke verglichen werden konnten. Die so berechneten Wahrscheinlichkeiten wurden den Kanten der Ontologien, welche die Musikstücke darstellten, zugewiesen. Die daraus berechnete Distanz zwischen den Musikstücken bedeutet in diesem Fall, inwieweit das Thema in den Variationen bzw. dem Teilstück aus Alla Turca nicht verarbeitet wurde. Die Distanz zwischen Thema und den beiden Variationen blieb für die einzelnen Kriterien jeweils zwischen zwei und zehn Bits, so dass hier die aus musikalischer Sicht erwartete Ähnlichkeit bestätigt werden konnte. Die Distanzen zu dem Vergleichsstück aus Alla Turca lagen hingegen zwischen 50 und 70 Bits, weil hier keine Verarbeitung des Themas stattfand. Mit diesen Ergebnissen konnte die relative Entropie erfolgreich für die Strukturierung von Ontologien eingesetzt werden.

Ausblick

Im folgenden Abschnitt werden zunächst mögliche Erweiterungen der beiden vorgestellten Ontologien diskutiert. Anschließend werden einige Ideen für weitere Anwendungen der entropiebasierten Distanzmaße gegeben.

Erweiterung der Flutontologie

Im Rahmen dieser Arbeit wurden einige Konzepte der Flutontologie, vor allem im Bereich des Umwelt- und Katastrophenmanagement, nicht vollständig oder ausreichend detailliert modelliert, weil nicht genügend Expertise zur Verfügung stand, so dass hier noch Erweiterungmöglichkeiten bestehen. Auch sind bisher versicherungstechnische Aspekte des Risikomanagements unberücksichtigt geblieben. Die für ein Hochwasser relevanten Versicherungen, deren Bedingungen und Einschränkungen, müssten noch in die Ontologie mit aufgenommen werden, um eine vollständige Modellierung des Risikomanagements zu erreichen. Die Versicherung federn nämlich einen Teil des vorhandenen Risikos ab, wie beispielsweise Produktionsausfall durch Hochwasserschäden. In der Flutontologie sollten auch rechtliche Vorschriften und Gesetze, welche insbesondere bei der Erstellung von baulichen Sicherungsmaßnahmen beachtet werden müssen, mit aufgenommen werden.

Auch entspricht die in der Flutontologie modellierte Welt im jetzigen Zustand der Sichtweise der Experten. Für eine verbesserte, semantisch einheitliche, Kommunikation sollten zumindest die Sichtweisen und damit auch die Begrifflichkeiten der Bevölkerung und der Rettungskräfte modelliert werden. Damit könnten mit der Methode der relativen Entropie auch die unterschiedlichen Wahrnehmungen verschiedener Bevölkerungsgruppen hinsichtlich des Flutrisikos dargestellt werden. Anhand der berechneten Distanzen könnten anschließend gezielte Aufklärungskampagnen gestartet werden.

Erweiterung der Musikontologie

Bei der Ontologie der musikalischen Wahrnehmung existieren ebenfalls noch einige Erweiterungsmöglichkeiten. Zum einen sind in der Ontologie die einen Klang erzeugenden Instrumente nicht berücksichtigt. Man könnte verschiedene Klangerzeuger einführen, welche die unterschiedlichen Instrumente repräsentieren. Den Klängen werden dann die entsprechenden Instrumente zugeordnet, so dass zusätzliches Wissen über die Klänge verfügbar ist. Hierbei muss aber beachtet werden, dass es bei gleichen Instrumenten zum Teil erhebliche Materialunterschiede gibt, die einen Klang entscheidend beeinflussen. Ein gutes Beispiel hierfür sind die Stradivariusgeigen, die wegen ihres Materials und ihrer Form einen besonders schönen Klang erzeugen.

In der Ontologie wurden die verschiedenen, existierenden Musikstile, die Musikstücke in bestimmte Kategorien einteilen, bisher nicht berücksichtigt. Klassische Stücke lassen sich noch einfach den einzelnen Stilen zuordnen, weil sie meistens entstehungszeitlich klassifiziert sind. Bei moderner Unterhaltungsmusik wie Rock, Pop, Jazz oder Blues wird es schon bedeutend schwieriger, eindeutig definierbare Kriterien für zuverlässige Klassifizierung anzugeben. Das liegt u.a. daran, dass manche Musikstücke mehreren Stilen zugeordnet werden können, weil deren Definition nicht eindeutig ist.

Ein weiterer Aspekt, der in die Ontologie mit aufgenommen werden könnte, ist die Betrachtung des „Kontextes“ eines Menschen, wenn er ein Musikstück hört. Dieser Kontext beeinflusst die Wahrnehmung von Musik erheblich und hängt von vielen, unterschiedlichen Faktoren ab. Das sind beispielsweise bisher schon gehörten Musikstücke, der kulturelle Hintergrund und der persönliche Geschmack einer Person. Wichtig ist ebenso die musikalische Bildung, denn je höher diese ist, desto mehr Eigenheiten eines Stückes werden erkannt.

Erweiterung der gegenseitigen Information

Eine mögliche Erweiterung der Verfeinerung der gegenseitigen Information auf Graphen könnte darin bestehen, mittels dieser Methodik die minimale Beschreibung eines Graphen, das heisst also seine Kolmogorovkomplexität zu erhalten. Dazu wird zunächst zwischen zwei Knoten des Graphen die gegenseitige Information nach der in Kapitel 3 beschriebenen Vorgehensweise berechnet. Nun wird durch geschicktes Umsortieren der Knoten auf dem Pfad, welches die Semantik des Graphen nicht verändert, versucht, die gegenseitige Information zu maximieren. Damit wird jegliche Redundanz zwischen den Knoten minimiert. Dieses Umsortieren geschieht solange, bis sich keine Änderung der gegenseitigen Information mehr ergibt, womit die minimale Beschreibung dieses Pfades gegeben wäre. Diese Vorgehensweise kann auf den gesamten Graphen oder einen interessierenden Teilgraphen erweitert werden, um dessen minimale Beschreibung zu erhalten.

Anwendungen der relativen Entropie

Eine Anwendungsmöglichkeit der relativen Entropie auf Ontologien ergibt sich in Zusammenhang mit Bayesnetzwerken. Sei dazu eine Ontologie mit verschiedenartigen Relationen gegeben. Anhand dieser Ontologie möchte man für eine probabilistische Entscheidungsfindung ein Bayesnetzwerk generieren. Das geschieht zum Beispiel dadurch, dass aus der Ontologie einfach die Vererbungsrelationen entfernt werden. Die übrigen Relationen bilden bereits ein kausales Netzwerk, welches für Bayesinferenz genutzt werden kann. Zusätzlich müssen noch die bedingten Wahrscheinlichkeitstabellen (CPT) pro Konten angegeben werden, die aber bereits in der Ontologie pro vorhandenem Konzept angeben werden können. Jedoch besteht bei dieser Vorgehensweise das Problem, dass die Wahrscheinlichkeitstabellen an einzelnen Knoten sehr groß geraten, weil sämtliche Relationen des Knotens berücksichtigt werden müssen. Eine Möglichkeit dieses Problem abzumildern, wäre eine automatische Abstraktion der Konzepte in der Ontologie bei der Umwandlung in ein Bayesnetzwerk. Die Entscheidung, ob auf eine höhere Ebene abstrahiert wird, könnte die relative Entropie liefern. Wenn die Distanz zwischen den Wahrscheinlichkeitstabellen der detaillierten Konzepte und dem dazugehörigen Elternkonzept gering ist, sind die zugrundeliegenden Wahrscheinlichkeitsverteilungen ähnlich und es geht durch eine Abstraktion nur wenig Information verloren. Je nach Abstraktionsgrad, also wieviele einzelne Konzepte zusammengefasst werden, ergibt sich aber eine beträchtliche Reduktion der Komplexität in dem Bayesnetzwerk. Falls die Distanz dagegen groß ist, sind die zugrundeliegenden Wahrscheinlichkeitsverteilung verschieden, und bei einer Abstraktion könnte dann zu viel Information verloren gehen.

Ein weiterer, interessanter Einsatz der relativen Entropie könnte es in Zusammenhang mit nichtmonotonem Schließen und dem Zusammenführen von unterschiedlichen Ontologien geben. Dazu müssen zwei Ontologien O_1 und O_2 zunächst in eine probabilistische, konditionale Wissensbasis [Kern-Isberner 2001] umgewandelt werden, was eine nicht triviale Aufgabe darstellt. Idealerweise ist diese Transformation ein Isomorphismus. Damit erhält man Wissensbasen P_1 und P_2, welche die entsprechenden Ontologien O_1 und O_2 repräsentieren. Die Schnittmenge zwischen P_1und P_2 sei ungleich Null, es existieren demzufolge gewisse Gemeinsamkeiten. Anhand der zwei Wissensbasen werden im nächsten Schritt konkrete Daten inferiert. Wenn hinreichend viele Daten akkumuliert wurden, kann nach [Kern-Isberner u. Fisseler 2004] wiederum eine Wissensbasis P_3 berechnet werden. Diese stellt die Zusammenführung der Wissensbasen P_1 und P_2 dar. Die Wissensbasis P_3 kann anschließend wieder in eine Ontologie O_3 überführt werden, falls die erste Transformation isomorph war. Die Ontologie O_3 stellt somit eine Zusammenführung der Ontologien O_1 und O_2 dar, wobei aufgrund der Eigenschaften des verwendeten Verfahrens keinerlei zusätzliche Information

mit eingeflossen ist.

Eine interessante Anwendung der relativen Entropie als Strukturierungsmaß würde sich im Hinblick auf die musikalische Wahrnehmung von Menschen ergeben, was eine Erweiterung der in Kapitel 6 vorgestellten Anwendung darstellt. Hierbei sollen mit dem Distanzmaß unterschiedliche Interpretationen ein und desselben Stückes miteinander verglichen werden. Die verschiedenen Interpretationen ergeben sich aus unterschiedlichen Dirigenten und Orchestern, welche ein Musikstück darbieten. Mit Hilfe der Ontologie für die menschliche, musikalische Wahrnehmung und, der sich durch Anwendung der relativen Entropie ergebenden Struktur, kann anschließend festgestellt werden, welche Interpretationen „interessant“ oder besonders „spannend“ sind und was für Faktoren dazu beitragen. Solch eine Strukturierung ist stets subjektiv, was sich auch in den Bedeutungen der Wahrscheinlichkeitsverteilungen (Interesse, Spannung) wiederspiegelt. Die relative Entropie würde in diesem Falle angeben, inwieweit sich Elemente einer Interpretation in einer anderen nicht wiederfinden. Für solch eine Untersuchung müssen allerdings „Emotionen“, die beim Hören von Musik eine entscheidende Rolle spielen, mit berücksichtigt werden. Damit wäre ein erster Schritt zur Darstellung und Erforschung von „Kreativität“ in der Musik gegeben.

Anhang A

Distanzen zwischen Thema und Variation

Die folgenden, tabellarischen Abbildungen geben übersichtlich die Wahrscheinlichkeitsverteilungen für die einzelnen Vergleichskriterien pro Takt an. Die Distanz der Variationen und des Vergleichsstückes aus Alla Turca zum Thema wird pro Takt und insgesamt, für das jeweilige Kriterium aufsummiert, angegeben.

Melodieklänge	**Takt 1**				**Takt 2**				**Takt 3**				**Takt 4**				***Distanz***
	cis	d	e		h	cis	d		a	h			cis	d	e	h	
Thema	0,40	0,20	0,40		0,40	0,20	0,40		0,50	0,50			0,40	0,20	0,20	0,20	
Distanz zu Var. 1		***3,18***				***0,53***				***0,52***					***0,85***		
Var. 1	0,30	0,00	0,30		0,30	0,10	0,30		0,40	0,30			0,25	0,08	0,17	0,08	
Distanz zu Var. 4		***0,20***				***0,20***				***0,82***					***0,06***		
Var. 4	0,40	0,40	0,20		0,40	0,40	0,20		0,40	0,20			0,33	0,17	0,33	0,17	
Distanz zu Alla Turca		***9,08***				***6,32***				***1,00***					***8,84***		
Alla Turca	0,33	0,00	0,00		0,25	0,25	0,00		0,25	0,25			0,00	0,00	0,25	0,25	
	Takt 5				**Takt 6**				**Takt 7**				**Takt 8**				
	cis	d	e		h	cis	d		a	h	cis	d	cis	h	a		
Thema	0,40	0,20	0,40		0,40	0,20	0,40		0,25	0,25	0,25	0,25	0,33	0,33	0,33		
Distanz zu Var. 1		***0,40***				***0,22***					***0,08***			***0,08***			***5,86***
Var. 1	0,40	0,20	0,20		0,50	0,33	0,17		0,38	0,25	0,25	0,13	0,50	0,25	0,25		
Distanz zu Var. 4		***0,20***				***0,20***					***0,32***			***0,19***			***2,19***
Var. 4	0,40	0,40	0,20		0,40	0,40	0,20		0,20	0,20	0,20	0,20	0,20	0,20	0,60		
Distanz zu Alla Turca		***9,08***				***6,32***					***10,96***			***9,57***			***61,17***
Alla Turca	0,33	0,00	0,00		0,25	0,25	0,00		0,00	0,25	0,00	0,00	0,67	0,00	0,00		

Abbildung A.1: Wahrscheinlichkeitsverteilungen und Distanzen pro Takt für die Melodieklänge. Angegeben sind die Noten, welche pro Takt vorkommen sowie deren relative Häufigkeiten. Die Gesamtdistanz ist in der letzen Spalte angegben.

Melodieführung	**Takt 1**				**Takt 2**				**Takt 3**				**Takt 4**				***Distanz***
	hoch	runter	gleich		hoch	runter	gleich		hoch	runter	gleich		hoch	runter	gleich		
Thema	0,50	0,25	0,25		0,50	0,25	0,25		0,33	0,00	0,67		0,25	0,75	0,00		
Distanz zu Var. 1		***0,01***				***0,04***				***0,94***				***0,01***			
Var. 1	0,56	0,22	0,22		0,44	0,33	0,22		0,44	0,33	0,22		0,27	0,73	0,00		
Distanz zu Var. 4		***3,40***				***3,40***				***1,41***				***0,02***			
Var. 4	0,50	0,50	0,00		0,50	0,50	0,00		0,20	0,40	0,20		0,33	0,67	0,00		
Distanz zu Alla Turca		***3,40***				***10,96***				***10,21***				***0,53***			
Alla Turca	0,50	0,50	0,00		0,00	1,00	0,00		1,00	0,00	0,00		0,67	0,33	0,00		
	Takt 5				**Takt 6**				**Takt 7**				**Takt 8**				
	hoch	runter	gleich		hoch	runter	gleich		hoch	runter	gleich		hoch	runter	gleich		
Thema	0,50	0,25	0,25		0,50	0,25	0,25		1,00	0,00	0,00		0,00	1,00	0,00		
Distanz zu Var. 1		***3,35***				***0,72***				***1,32***				***0,58***			***6,16***
Var. 1	0,75	0,25	0,00		0,40	0,20	0,40		0,40	0,20	0,40		0,00	0,67	0,33		
Distanz zu Var. 4		***3,40***				***3,40***				***1,00***				***0,00***			***16,03***
Var. 4	0,50	0,50	0,00		0,50	0,50	0,00		0,50	0,50	0,00		0,00	1,00	0,00		
Distanz zu Alla Turca		***3,40***				***10,96***				***1,59***				***16,43***			***57,48***
Alla Turca	0,50	0,50	0,00		0,00	1,00	0,00		0,33	0,67	0,00		0,50	0,00	0,50		

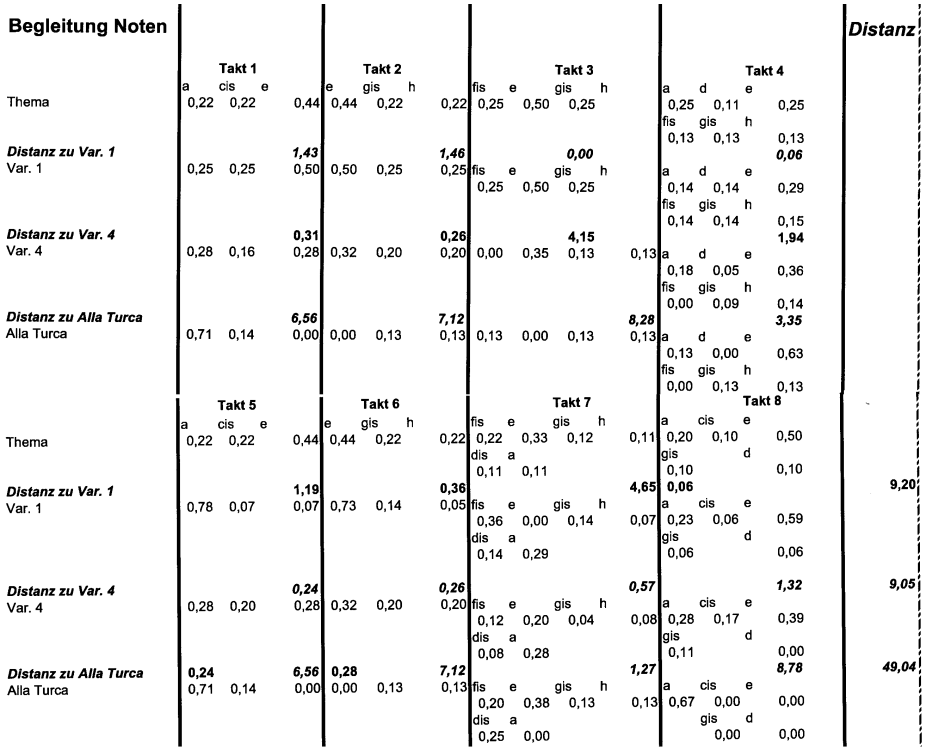
Abbildung A.2: Wahrscheinlichkeitsverteilungen und Distanzen pro Takt für die Melodieführung. Angegeben sind die relativen Häufigkeiten der Auf- und Abwärtsbewegungen sowie die gleichbleibende Fortführung der Melodie. Die Gesamtdistanz ist in der letzen Spalte angegben.

Puls Melodie									Distanz
	Takt 1		Takt 2		Takt 3		Takt 4		
	Puls	kein Puls	Puls	kein Puls	Puls	kein Puls	Puls	kein Puls	
Thema	0,38	0,62	0,38	0,62	0,31	0,69	0,38	0,62	
Distanz zu Var. 1		***0,50***		***0,50***		***0,69***	***1,35***		
Var. 1	0,77	0,23	0,77	0,23	0,77	0,23	0,92	0,08	
Distanz zu Var. 4		***0,00***		***0,00***		***0,02***	***0,02***		
Var. 4	0,38	0,62	0,38	0,62	0,38	0,62	0,46	0,54	
Distanz zu Alla Turca									
Alla Turca									
	Takt 5		**Takt 6**		**Takt 7**		**Takt 8**		
	Puls	kein Puls	Puls	kein Puls	Puls	kein Puls	Puls	kein Puls	
Thema	0,38	0,62	0,38	0,62	0,31	0,69	0,23	0,77	
Distanz zu Var. 1		**0,00**		**0,02**		**0,28**		**0,02**	**3,36**
Var. 1	0,38	0,62	0,46	0,54	0,62	0,38	0,31	0,69	
Distanz zu Var. 4		***0,00***		***0,00***		***0,02***		***0,07***	***0,13***
Var. 4	0,38	0,62	0,38	0,62	0,38	0,62	0,38	0,62	
Distanz zu Alla Turca									
Alla Turca									

Abbildung A.3: Wahrscheinlichkeitsverteilungen und Distanzen pro Takt für den Melodiepuls. Angegeben ist relative Häufigkeit der Anzahl der Notenanschläge pro Takt. Die Gesamtdistanz ist in der letzen Spalte angegben.

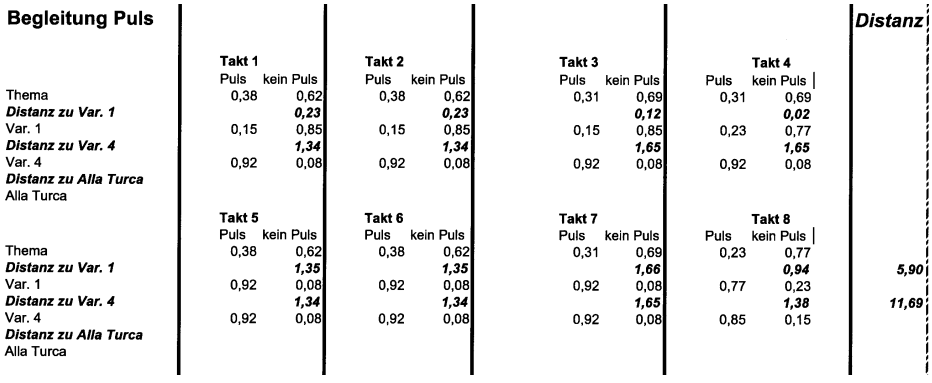
Begleitung Harmonien																Distanz
	Takt 1				**Takt 2**				**Takt 3**				**Takt 4**			
	A				E				fis7	E			A	E	h	
Thema	1,00				1,00				0,50	0,50			0,50	0,25	0,25	
Distanz zu Var. 1	***0,00***				***0,00***					***0,00***				***0,00***		
Var. 1	1,00				1,00				0,50	0,50			0,50	0,25	0,25	
Distanz zu Var. 4	***0,00***				***0,00***					***7,80***				***0,00***		
Var. 4	1,00				1,00				0,00	0,50			0,50	0,25	0,25	
Distanz zu Alla Turca	***0,00***				***16,60***					***15,60***				***11,00***		
Alla Turca	1,00				0,00				0,00	0,00				1,00		
	Takt 5				**Takt 6**				**Takt 7**				**Takt 8**			
	A				E				fis7	E	A	h	A	E7	A	
Thema	1,00				1,00				0,25	0,25	0,25	0,25	0,33	0,33	0,34	
Distanz zu Var. 1	***0,00***				***0,00***						***0,00***			***0,00***		***0,00***
Var. 1	1,00				1,00				0,25	0,25	0,25	0,25	0,33	0,33	0,34	
Distanz zu Var. 4	**0,00**				**0,00**						**0,00**			**0,00**		**7,80**
Var. 4	1,00				1,00				0,25	0,25	0,25	0,25	0,33	0,33	0,34	
Distanz zu Alla Turca	***0,00***				***16,60***						***10,70***			***4,57***		***75,06***
Alla Turca	1,00				0,00				0,00	0,50	0,00	0,00	1,00			

Abbildung A.4: Wahrscheinlichkeitsverteilungen und Distanzen pro Takt für die Begleitharmonien. Angegeben ist relative Häufigkeit der Anzahl der Akkorde pro Takt. Die Gesamtdistanz ist in der letzen Spalte angegben.



Begleitung Noten	Takt 1			Takt 2			Takt 3				Takt 4				Distanz
	a	cis	e	e	gis	h	fis	e	gis	h	a	d	e		
Thema	0,22	0,22	0,44	0,44	0,22	0,22	0,25	0,50	0,25		0,25	0,11	0,25		
											fis	gis	h		
											0,13	0,13	0,13		
Distanz zu Var. 1			***1,43***			***1,46***			***0,00***				***0,06***		
Var. 1	0,25	0,25	0,50	0,50	0,25	0,25	fis	e	gis	h	a	d	e		
							0,25	0,50	0,25		0,14	0,14	0,29		
											fis	gis	h		
											0,14	0,14	0,15		
Distanz zu Var. 4			***0,31***			***0,26***			***4,15***				***1,94***		
Var. 4	0,28	0,16	0,28	0,32	0,20	0,20	0,00	0,35	0,13	0,13	a	d	e		
											0,18	0,05	0,36		
											fis	gis	h		
											0,00	0,09	0,14		
Distanz zu Alla Turca			***6,56***			***7,12***				***8,28***			***3,35***		
Alla Turca	0,71	0,14	0,00	0,00	0,13	0,13	0,13	0,00	0,13	0,13	a	d	e		
											0,13	0,00	0,63		
											fis	gis	h		
											0,00	0,13	0,13		
	Takt 5			**Takt 6**			**Takt 7**				**Takt 8**				
	a	cis	e	e	gis	h	fis	e	gis	h	a	cis	e		
Thema	0,22	0,22	0,44	0,44	0,22	0,22	0,22	0,33	0,12	0,11	0,20	0,10	0,50		
							dis	a			gis		d		
							0,11	0,11			0,10		0,10		
Distanz zu Var. 1			1,19			0,36				4,65	0,06				9,20
Var. 1	0,78	0,07	0,07	0,73	0,14	0,05	fis	e	gis	h	a	cis	e		
							0,36	0,00	0,14	0,07	0,23	0,06	0,59		
							dis	a			gis		d		
							0,14	0,29			0,06		0,06		
Distanz zu Var. 4			***0,24***			***0,26***				***0,57***			***1,32***		***9,05***
Var. 4	0,28	0,20	0,28	0,32	0,20	0,20	fis	e	gis	h	a	cis	e		
							0,12	0,20	0,04	0,08	0,28	0,17	0,39		
							dis	a			gis		d		
							0,08	0,28			0,11		0,00		
Distanz zu Alla Turca	**0,24**		***6,56***	**0,28**		***7,12***				***1,27***			***8,78***		***49,04***
Alla Turca	0,71	0,14	0,00	0,00	0,13	0,13	fis	e	gis	h	a	cis	e		
							0,20	0,38	0,13	0,13	0,67	0,00	0,00		
							dis	a				gis	d		
							0,25	0,00				0,00	0,00		

Abbildung A.5: Wahrscheinlichkeitsverteilungen und Distanzen pro Takt für die Begleitklänge. Angegeben sind die vorkommenden Noten in den Begleitharmonien sowie deren relative Häufigkeit der pro Takt. Die Gesamtdistanz ist in der letzen Spalte angegeben.



Begleitung Puls	Takt 1		Takt 2		Takt 3		Takt 4		Distanz
	Puls	kein Puls	Puls	kein Puls	Puls	kein Puls	Puls	kein Puls	
Thema	0,38	0,62	0,38	0,62	0,31	0,69	0,31	0,69	
Distanz zu Var. 1		***0,23***		***0,23***		***0,12***		***0,02***	
Var. 1	0,15	0,85	0,15	0,85	0,15	0,85	0,23	0,77	
Distanz zu Var. 4		***1,34***		***1,34***		***1,65***		***1,65***	
Var. 4	0,92	0,08	0,92	0,08	0,92	0,08	0,92	0,08	
Distanz zu Alla Turca									
Alla Turca									
	Takt 5		**Takt 6**		**Takt 7**		**Takt 8**		
	Puls	kein Puls	Puls	kein Puls	Puls	kein Puls	Puls	kein Puls	
Thema	0,38	0,62	0,38	0,62	0,31	0,69	0,23	0,77	
Distanz zu Var. 1		***1,35***		***1,35***		***1,66***		***0,94***	***5,90***
Var. 1	0,92	0,08	0,92	0,08	0,92	0,08	0,77	0,23	
Distanz zu Var. 4		***1,34***		***1,34***		***1,65***		***1,38***	***11,69***
Var. 4	0,92	0,08	0,92	0,08	0,92	0,08	0,85	0,15	
Distanz zu Alla Turca									
Alla Turca									

Abbildung A.6: Wahrscheinlichkeitsverteilungen und Distanzen pro Takt für den Begleitpuls. Angegeben ist relative Häufigkeit der Anzahl der Notenanschläge pro Takt. Die Gesamtdistanz ist in der letzen Spalte angegeben.

Literaturverzeichnis

Aggarwal u. a. 1999
Aggarwal, C. C. ; Gates, S. C. ; Yu, P. S.: On the merits of building categorization systems by supervised clustering. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, 1999, S. 352 – 356

Agirre u. Rigau 1996
Agirre, E. ; Rigau, G.: Word sense disambiguation using conceptual density. In: *Proceedings of COOLING*, 16 – 22

Ali u. Silvey 1966
Ali, S. M. ; Silvey, S. D.: A General Class of Coefficients of Divergence of One Distribution from Another. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 28 (1966), Nr. 1, S. 131 – 142

Audi 2004
Audiovisual Institute, Universitat Pompeu Fabra (Veranst.): *ISMIR 2004 5th International Conference on Music Information Retrieval*. Barcelona, Spanien, 2004

Baader u. a. 2003
Baader, F. (Hrsg.) ; Calvanese, D. (Hrsg.) ; McGuinness, D. (Hrsg.) ; Nardi, D. (Hrsg.) ; Patel-Schneider, P. (Hrsg.): *The Description Logic Handbook : Theory, Implementation and Applications*. Cambridge University Press, 2003

Bashkirov 2003
Bashkirov, A. G.: On the Rényi entropy, Boltzmann Principle, Levy and Power-law distributions and Renyi parameter. In: *eprint arXiv:cond-mat/0211685* (2003)

Bello u. a. 2000
Bello, J. P. ; Monti, G. ; Sandler, M. B.: Techniques for Automatic Music Transcription. In: *International Symposium on Music Information Retrieval (ISMIR)*. Plymouth, Massachusetts, USA, 2000

Benferhat u. a. 1997
BENFERHAT, S. ; DUBOIS, D. ; PRADE, H.: Nonmonotonic reasoning, conditional objects and possibility theory. In: *Artificial Intelligence* 92 (1997), S. 259 – 276

Berners-Lee u. a. 2001
BERNERS-LEE, T. ; HENDLER, J. ; LASSILA, O.: The Semantic Web. In: *Scientific American* (2001), 05

Beyerer 1999
BEYERER, J.: *Verfahren zur quantitativen statistischen Bewertung von Zusatzwissen in der Meßtechnik.* VDI-Verlag, 1999 (Fortschritt-Berichte VDI : Reihe 8, Meß-, Steuerungs- und Regelungstechnik)

Billard u. Diday 2003
BILLARD, L. ; DIDAY, E.: From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. In: *Journal of the American Stistical Association* 98 (2003), Nr. 462, S. 470 – 487

Bittner u. a. 2004
BITTNER, T. ; DONNELLY, M. ; SMITH, B.: Individuals, Universals, Collections: On the Foundational Relations of Ontology. In: *Proceedings of FOIS 2004*

Boulanger 2000
BOULANGER, R.: *The Csound Book: Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming.* MIT Press, 2000

Bourne 2003
BOURNE, R. A.: Explaining default intuitions using maximum entropy. In: *Applied Logic* (2003), Nr. 1, S. 255 – 271

Braut 1993
BRAUT, C.: *Das MIDI Buch.* Sybex Verlag, 1993

Brickley u. Guha 2004
BRICKLEY, D. ; GUHA, R.V.: *RDF Vocabulary Description Language 1.0: RDF Schema.* http://www.w3.org/TR/rdf-schema/: W3C Recommendation, 2004

Calmet u. Daemi 2004
CALMET, J. ; DAEMI, A.: From Entropy to Ontology. In: TRAPPL, R. (Hrsg.): *Cybernetics and systems 2004 - AT2AI-4: From Agent Theory to Agent Implementation* Bd. 2, 2004, S. 547 – 551

Celeux u. a. 1989
CELEUX, G. ; DIDAY, E. ; GOVAERT, G.: *Classification automatique des donnees.* Dunod, 1989

Chaudhri u. a. 1998
CHAUDHRI, V. K. ; A.FARQUHAR ; FIKES, R. ; KARP, P.D. ; RICE, J. P.: OKBC: a programmatic foundation for knowledge base interoperability. In: *Proceedings of the fifteenth/tenth national conference on Artificial intelligence/Innovative applications of Artificial Intelligence*, 1998, S. 600 – 607

Chernoff 1952
CHERNOFF, H.: Measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. In: *Annal Math. Stat.* 23 (1952), S. 493 – 507

Ciszár 1995
CISZÁR, I.: Generalized Cutoff Rates and Rényi's Information Measures. In: *IEEE Transactions on Information Theory* 41 (1995), Nr. 1, S. 26 – 34

Corman u. a. 2002
CORMAN, S.R. ; MCPHEE, R. ; DOOLEY, K.: Studying complex discursive Systems: Centering Resonance Analysis of Communication. In: *Human Communication Research* (2002), S. 157 – 206

Cover u. Thomas 1991
COVER, T.M. ; THOMAS, J.A.: *Elements of Information Theory*. Wiley Series in telecommunications, 1991

Daemi u. Calmet 2004a
DAEMI, A. ; CALMET, J.: Assessing Conflicts in Ontologies. In: *WSEAS Transactions on Information Science and Applications* 1 (2004), Nr. 5, S. 1289 – 1294

Daemi u. Calmet 2004b
DAEMI, A. ; CALMET, J.: From Ontologies to Trust through Entropy. Luxembourg City, Luxembourg, 2004 (Proceedings of AISTA 2004 - International Conference on Advances in Intelligent Systems: Theory and Applications. In Cooperation with the IEEE Computer Society)

Doyle 1988
DOYLE, J.: On Universal Theories of Defaults. Version: 1988. `citeseer.ist.psu.edu/doyle88universal.html` (CMU-CS-88-111). – Forschungsbericht. – Elektronische Ressource

Dutoit u. Poibeau 2002
DUTOIT, D. ; POIBEAU, T.: Inferring knowledge form a large semantic network. In: *Proceedings of COLING*, 307-316

Erdmann u. a. 2000
ERDMANN, M. ; MAEDCHE, A. ; SCHNURR, H. P. ; STAAB, Steffan: From Manual to Semi-Automatic Annotation: About Ontology-based Text Annotation Tools. In: BUITELAAR, P. (Hrsg.) ; HASIDA, K. (Hrsg.): *Proceedings of COOLING*. Luxembourg, 2000

Fancis u. Kucera 1982
FANCIS, W. N. ; KUCERA, H.: Frequency Analysis of English Usage: Lexicon and Grammar. In: *Houghton Mifflin* (1982)

Farquhar u. a. 1997
FARQUHAR, A. ; FIKES, R. ; RICE, J.: The Ontolingua Server: A Tool for Collaborative Ontology Construction. In: *International Journal of Human Computer Studies* 46 (1997), Nr. 6, S. 707 – 727

Fensel 2000
FENSEL, D.: *Ontologies: A silver bullet for Knowledge Management and Electronic Commerce*. Springer Verlag, 2000

Frisius 1984
FRISIUS, R.: *Notation und Komposition*. Ernst Klett Verlag, 1984

Gaul 2000
GAUL, W.: *Data analysis : scientific modeling and practical application*. Springer, 2000

Genesereth u. Fikes 1992
GENESERETH, M. R. ; FIKES, R. E.: *Knowledge Interchange Format. Version 3.0. Reference Manual*. Stanford University: Computer Science Department, 1992

Gennari u. a. 2002
GENNARI, J. ; MUSEN, M. A. ; FERGERSON, R. W. ; GROSSO, W. E. ; CRUBÉZY, M. ; ERIKSSON, H. ; NOY, N. F. ; TU, S. W.: The Evolution of Protégé: An Environment for Knowledge-Based Systems Development / Stanford Medical Informatics. Version: 2002. `http://www.smi.stanford.edu/pubs/SMI_Reports/SMI-2002-0943.pdf` (SMI-2002-0943). – Forschungsbericht. – Elektronische Ressource

Gibbs 1878
GIBBS, J. W.: On the Equilibrium of Heterogeneous Substances. In: *The American Journal of Science and Arts* XVI (1878), Nr. 96, S. 441 – 458

Goldszmidt u. a. 1993
GOLDSZMIDT, M. ; MORRIS, P. ; PEARL, J.: A Maximum Entropy Approach to Nonmonotonic Reasoning. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15 (1993), March, Nr. 3

Gorges 1997
GORGES, A. Merck P.: *Keyboards, MIDI, Homerecording.* Carstensen Verlag, 1997

Goto 2001
GOTO, M.: An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds. In: *Journal of New Music Research* 30 (2001), Juni, Nr. 2, S. 159 – 171

Gouyon u. Dixon 2005
GOUYON, F. ; DIXON, S.: A Review of Automatic Rhythm Description Systems. In: *Computer Music Journal,* 29 (2005), Nr. 1, S. 34 – 54

Gruber 1993a
GRUBER, T. R.: A translation approach to portable ontology specification. In: *Knowledge Acquisition* 5 (1993), Nr. 2, S. 199 – 220

Gruber 1996
GRUBER, T. R.: Ontolingua: A mechanism to support portable ontologies / Stanford University. 1996 (Technical Report KSL-91-66). – Forschungsbericht

Gruber 1993b
GRUBER, T.R.: Towards principles for the design of ontologies used for knowledge sharing. In: *Knowledge Acquisition* (1993), S. 199 – 220

Gómez-Pérez u. a. 2004
GÓMEZ-PÉREZ, A. ; FERNÁNDEZ-LÓPEZ, M. ; CORCHO, O.: *Ontological Engineering.* Springer, 2004

Handschuh u. a. 2003
HANDSCHUH, S. ; STAAB, S. ; VOLZ, R.: On Deep Annotation. In: *Proceedings of WWW2003.* Budapest, Hungary, 2003

van Harmelen u. Patel-Schneider 2001
HARMELEN, F. van ; PATEL-SCHNEIDER, P.: *DAML+OIL.* http://www.daml.org/2001/03/daml+oil-index.html: DARPA und EU (IST), 2001

Harold 2002
HAROLD, E. R.: *Die XML Bibel.* mitp, 2002 (2. Auflage)

Hartley 1928
HARTLEY, R. V.: Transmission of Information. In: *Bell System Technical Journal* (1928), July, S. 535 ff.

Horridge u. a. 2004
HORRIDGE, M. ; KNUBLAUCH, H. ; RECTOR, A. ; STEVENS, R. ; WROE,

C.: *A Practical Guide to Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools*. 1.0. The University of Manchester, 2004

Hotho u. a. 2002
HOTHO, A. ; MAEDCHE, A. ; STAAB, S.: Ontology-based Text Document Clustering. In: *Künstliche Intelligenz* (2002), 04, S. 48 – 54

Jaynes 1979
JAYNES, E. T.: Where do we stand on maximum entropy? In: LEVINE, R. (Hrsg.) ; TRIBUS, M. (Hrsg.): *The Maximum Entropy Formalism*, MIT Press, 1979, S. 15 – 118

Jaynes 1981
JAYNES, E. T. ; ROSENKRANTZ, R. (Hrsg.): *Papers on Probablity, Statistics and Statistical Physics*. Dordrecht: Reidel, 1981

Jaynes 1982
JAYNES, E. T.: On The Rationale of Maximum-Entropy Methods. In: *Proceedings of the IEEE* Bd. 70, 1982, S. 939 – 952

Johnson u. Sinanovic 2001
JOHNSON, D. H. ; SINANOVIC, S.: *Symmetrizing the Kullback-Leibler Distance*. 2001. – Submitted to IEEE transactions on Information Theory 2001

Kern-Isberner 1998
KERN-ISBERNER, G.: Characterizing the principle of minimum cross-entropy within a conditional-logical framework. In: *Artificial Intelligence* 98 (1998), S. 169 – 208

Kern-Isberner 2001
KERN-ISBERNER, G.: *Conditionals in Nonmonotonic Reasoning and Belief Revision*. Springer, 2001 (LNAI 2087)

Kern-Isberner u. Fisseler 2004
KERN-ISBERNER, G. ; FISSELER, J.: Knowledge discovery by reversing inductive knowledge representation. In: *Proc. of 9th International Conference on the Principles of Knowledge Representation and Reasoning, KR-2004*, 2004

Kink 2005
KINK, S.: *Kreativität in der Musik durch Spezifizierung*, Universität Karlsruhe, Diplomarbeit, 2005

Kirchbach 2003
KIRCHBACH, H.P.: *Bericht der unabhängigen Kommision der sächsischen Staatsregierung zur Hochwasserkatastrophe im August 2002.*

`http://www.sachsen.de/de/bf/staatsregierung/ministerien/smi/smi/upload/hochwasserbericht_teil1.pdf`. Version: Februar 2003

Klir u. Wierman 1998
KLIR, G. J. ; WIERMAN, M. J.: *Uncertainty-Based Information*. Springer Verlag, 1998 (Studies in fuzziness and soft computing)

Koller u. Pfeffer 1998
KOLLER, D. ; PFEFFER, A.: Probabilistic frame-based systems. In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-99)*, 1998, S. 580 – 587

Kolmogorov 1965
KOLMOGOROV, A.N.: Three approaches to the quantitative definition of information. In: *Problems Inform. Transmission* 1 (1965), Nr. 1, S. 1 – 7

Kotz u. Johnson 1981
KOTZ, S. (Hrsg.) ; JOHNSON, N.L. (Hrsg.): *Encyclopedia of statistical science*. Bd. 4. John Wiley and Sons, 1981

Kraus u. a. 1990
KRAUS, S. ; LEHMANN, D. ; MAGIDOR, M.: Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. In: *Artificial Intelligence* 44 (1990), S. 167 – 207

Kudjoi 2004
KUDJOI, I.: Data clustering with kernel methods derived from Kullback-Leibler information / Helsinki University of Technology. 2004 (Mat-1.125). – Forschungsbericht

Kullback u. Leibler 1951
KULLBACK, S. ; LEIBLER, R.A.: On Information and Sufficiency. In: *Annal Math. Stat.* 22 (1951), S. 79 – 86

Kämpf u. a. 2002
KÄMPF, Ch. ; BRUDY-ZIPPELIUS, T. ; LIEBERT, J. ; IHRINGER, J.: Operable water quantity model system to support water quality management in a river sub-basin. In: PILLMANN, W. (Hrsg.) ; TOCHTERMANN, K. (Hrsg.): *Environmental Communication in the Information Society*, 2002, S. 561 – 568

Kämpf u. a. 2003
KÄMPF, Ch. ; IHRINGER, J. ; DAEMI, A. ; CALMET, .: Agent-Based Expert Information Retrieval for Flood-Risk Management. In: GNAUCK, Albrecht (Hrsg.) ; HEINRICH, Ralph (Hrsg.): *The Information Society and Enlargement of the European Union - 17th Int. Conference on*

Informatics for Environmental Protection Bd. 1, metropolis, 2003, S. 274 – 281

Lassila u. Swick 1999
LASSILA, O. ; SWICK, R.: *Resource Description Framework (RDF) Model and Syntax Specification.* http://www.w3.org/TR/REC-rdf-syntax: W3C Recommendation, 1999

Lenat u. Guha 1990
LENAT, D. B. ; GUHA, R. V.: *Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project.* Addison-Wesley, 1990

Li u. Vitányi 1993
LI, M. ; VITÁNYI, P.: *An Introduction to Kolmogorov Complexity and Its Applications.* Springer Verlag, 1993

Li u. a. 2003
LI, Y. ; BANDAR, Z. A. ; MCLEAN, D.: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. In: *IEEE Transactions on knowledge and data engineering* 15 (2003), S. 871 – 881

Lieb u. Yngvason 1999
LIEB, E. H. ; YNGVASON, J.: The physics and mathematics of the Second Law of Thermodynamics. In: *Physics Reports* 310 (1999), S. 1 – 99

Lieb u. Yngvason 2000
LIEB, E. H. ; YNGVASON, J.: A Fresh Look at Entropy and the Second Law of Thermodynamics. In: *Physics Today* (2000), April, S. 32 – 37

Lieb u. Yngvason 2002
LIEB, E. H. ; YNGVASON, J.: The Mathematical Structure of the Second Law of Thermodynamics. In: *Current Developments in Mathematics* (2002), S. 89 – 130

MacKay 2003
MACKAY, D. J.: *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003

Mahalanobis 1936
MAHALANOBIS, P.C.: On the generalized distance in statistics. In: *Proc. Nat. Inst. Sci. India* 12 (1936), S. 49 – 55

Makinson 2003
MAKINSON, D.: Bridges between Classical and Nonmonotonic Logic. In: *Logic Journal of the IGPL* 11 (2003), Nr. 1, S. 69–96

Maxwell 1871
MAXWELL, J. C.: *Theory of Heat.* London : Longmans, 1871

McCarthy 1980
MCCARTHY, J.: Circumscription: A form of non-monotonic reasoning. In: *Artificial Intelligence* 13 (1980), S. 27 – 39

McCarthy 1995
MCCARTHY, J. L.: What has AI in Common with Philosophy. In: *Proceedings of IJCAI*, 1995, S. 2041 – 2044

McGuinness u. van Harmelen 2004
MCGUINNESS, D. L. ; HARMELEN, F. van: *OWL Web Ontology Language.* http://www.w3.org/TR/owl-features/: W3C Recommendation, 2004

Miller u. a. 1993
MILLER, G. A. ; BECKWITH, R. ; FELLBAUM, C. ; GROSS, D. ; MILLER, K.: *Introduction to WordNet: An On-line Lexical Database.* Cognitive Science Laboratory, Princeton University `www.cogsci.princeton.edu/~wn/`

Niedersächsisches 2003
NIEDERSÄCHSISCHES, Umweltministerium: *German Environmental Information Network (GEIN).* http://www.gein.de, 2003

Niles u. Pease 2003
NILES, I. ; PEASE, A.: Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03).* Las Vegas, Nevada, june 2003

Nilsson 1998
NILSSON, N. J.: *Artificial Intelligence: A New Synthesis.* Morgan Kaufmann, 1998

Oberle u. a. 2000
OBERLE, P. ; THEOBALD, S. ; EVDAKOV, O. ; NESTMANN, F.: GIS-supported flood modelling by the example of the river Neckar / Kassel Reports Hydraulic Engine. 2000 (9). – Forschungsbericht

Orăsan u. a. 2004
ORĂSAN, C. ; PEKAR, V. ; HASLER, L.: A comparison of summarisation methods based on term specificity estimation. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004).* Lisbon, Portugal, May 26 – 28 2004, S. 1037 – 1041. – Available at: http://clg.wlv.ac.uk/papers/orasan-04a.pdf

Paris u. Vencovská 1990
PARIS, J. ; VENCOVSKÁ, A.: A note on the inevitablitiy of maximum entropy. In: *Approximate Reasoning* 4 (1990), S. 183 – 224

Paris u. Vencovská 1997
PARIS, J. ; VENCOVSKÁ, A.: In defense of the maximum entropy inference process. In: *Approximate Reasoning* 17 (1997), S. 77 – 103

Pearl 1991
Kapitel 8. In: PEARL, J.: *Philosophy and AI.* MIT Press, 1991, S. 157 – 187

Pepper u. Moore 2001
PEPPER, S. ; MOORE, G.: *XML Topic Maps (XTM) 1.0.* http://www.topicmaps.org/xtm/index.html: ISO, 2001

Petta u. Trappl 2001
PETTA, P. ; TRAPPL, R.: Emotions and agents. In: *Multi-Agent systems and applications* (2001), S. 301–316. ISBN 3–540–42312–5

Pierce 1980
PIERCE, J.R.: *An Introduction to Information Theory - Symbols, Signals and Noise.* Second. Dover, 1980

Plastino u. Plastino 1999
PLASTINO, A. ; PLASTINO, A. R.: Tsallis entropy and Jaynes' Information Theory formalism. In: *Braz. J. Phys.* 29 (1999), Nr. 1, S. 50 – 60

Plate u. Merz 2001
PLATE, E. J. ; MERZ, B.: *Naturkatastrophen: Ursachen - Auswirkungen - Vorsorge.* Schweizerbart, 2001

Rath 2002
RATH, H.: GPS des Web. In: *iX* 6 (2002), S. 115 – 122

Reiter 1980
REITER, R.: A logic for Default Reasoning. In: *Artificial Intelligence* 13 (1980), S. 81 – 132

Resnik 1995
RESNIK, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: *Proceedings of IJCAI*, 1995, S. 448–453

Russel u. Norvig 2003
RUSSEL, S. ; NORVIG, P.: *Artificial Intelligence - A modern approach.* Second. Prentice Hall, 2003

Rényi 1970
RÉNYI, A.: *Probability theory.* Springer Verlag, 1970

Rényi 1976a
RÉNYI, A.: On measures of entropy and information. In: *Selected papers of Alfred Renyi* Bd. 2. Hungarian Academy of Sciences, 1976, S. 565 – 580

Rényi 1976b
RÉNYI, A.: Some fundamental questions of information theory. In: *Selected papers of Alfred Renyi* Bd. 2. Hungarian Academy of Sciences, 1976, S. 526 – 552

Rényi 1982
RÉNYI, A.: *Tagebuch über die Informationstheorie.* Birkhäuser Verlag, 1982

Rödder u. Kern-Isberner 1997
RÖDDER, W. ; KERN-ISBERNER, G.: Léa Sombé und entropie-optimale Wissensverarbeitung mit der Expertensystem-Shell SPIRIT. In: *OR-Spektrum* 19 (1997), S. 41 – 46

Scheffer u. Schachtschnabel 2002
SCHEFFER, F. ; SCHACHTSCHNABEL, P.: *Lehrbuch der Bodenkunde.* Spektrum Akademischer Verlag, 2002 (15. Auflage)

Shannon 1948
SHANNON, C. E.: A mathematical theory of communication. In: *Bell System Technical Journal* 27 (1948), July and October, S. 379 – 423, 623 – 656

Shore u. Johnson 1981
SHORE, J. E. ; JOHNSON, R. W.: Properties of cross-entropy minimization. In: *IEEE Transactions on Information Theory* 27 (1981), Nr. 4, S. 472 – 482

Smith 2004
SMITH, B.: Beyond Concepts: Ontology as Reality Representation. In: *Proceedings of FOIS 2004*

Sowa 2000
SOWA, J. F.: *Knowledge Representation.* Brooks/Cole, 2000

Staab u. Studer 2004
STAAB, S. (Hrsg.) ; STUDER, R.i (Hrsg.): *Handbook on Ontologies.* Springer, 2004

Studer u. a. 1998
STUDER, R. ; BENJAMIN, V. R. ; FENSEL, D.: Knowledge Engineering: Principles and Methods. In: *IEEE Transactions on Data and Knowledge Engineering* 25 (1998), Nr. 1 – 2, S. 161 – 197

Sure u. a. 2002
SURE, Y. ; ERDMANN, M. ; ANGELE, J. ; STAAB, S. ; STUDER, R. ; WENKE, D.: OntoEdit: Collaborative Ontology Development for the Semantic Web. In: *Proceedings of the First International Semantic Web Conference on The Semantic Web* Bd. 2342, Springer, 2002, S. 221 – 235

Temperley 2002
TEMPERLEY, D.: *The Cognition of Basic Musical Structures.* MIT Press, 2002

Thong u. Liu 2004
THONG, N. (Hrsg.) ; LIU, J. (Hrsg.): *Intelligent Technologies for Information Analysis.* Springer, 2004

Tolman 1979
TOLMAN, R. C.: *The principles of statistical mechanics.* Dover, 1979

Touretzky 1986
TOURETZKY, D.S.: *The Mathematics of Inheritance Systems: Research Notes in Artificial Intelligence.* Los Altos, CA : Morgan Kaufmann, 1986

Tsallis u. a. 1988
TSALLIS, C. ; BORGES, E. P. ; BALDOVIN, F.: Mixing and equilibration: Protagonists in the scene of nonextensive statistical mechanic. In: *Journal of Stat. Phys.* 52 (1988), Nr. 479

Uffink 1995
UFFINK, J.: Can the maximum entropy principle be explained as a consistency requirement? In: *Studies in History and Philosophy of Modern Physics* 26 (1995), S. 223 – 261

Uffink 2001
UFFINK, J.: Bluff your way in the Second Law of Thermodynamics. In: *Studies in History and Philosophy of Modern Physics* 32 (2001), S. 305 – 394

Unit 1989
United Nations Centre for Regional Development (Veranst.): *Challenges of the IDNDR. Report and Summary of Proceedings of the International Symposium on Challenges of the IDNDR.* 1989 (UNCRD Meeting Report Series 32)

Uschold 2003
USCHOLD, M.: Where is the Semantics in the Semantic Web? In: *AI Magazine* 24 (2003), September, Nr. 3, S. 25 – 36

Uschold u. King 1995
USCHOLD, M. ; KING, M.: Towards a Methodology for Building Ontologies. In: SKUCE, D. (Hrsg.): *IJCAI 95 Workshop on Basic Ontological Issues in Knowledge Sharing*. Montreal, Canada, 1995, S. 6.1 – 6.10

Uschold u. Grüninger 1996
USCHOLD, M. F. ; GRÜNINGER, M.: Ontologies: Principles, Methods and Applications. In: *Knowledge Engineering Review* 11 (1996), Nr. 2, S. 93 – 155

Viitaniemi u. Eronen 2003
VIITANIEMI, T. ; ERONEN, A.: A probabilistic model for the transcription of single-voice melodies / Tampere University of Technology. 2003 (FINSIG). – Forschungsbericht

Vossen 1998
VOSSEN, P. (Hrsg.): *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, 1998

Weber u. a. 2002
WEBER, K. (Hrsg.) ; NAGENBORG, M. (Hrsg.) ; SPINNER, H. F. (Hrsg.): *Studien zur Wissensordnung*. Bd. 2: *Wissensarten, Wissensordnungen, Wissensregime*. Leske + Buderich, 2002

Wehrl 1978
WEHRL, A.: General properties of entropy. In: *Reviews of Modern Physics* 50 (1978), Nr. 2

Weiss 2000
WEISS, G. (Hrsg.): *Multiagent Systems - A Modern Approach to Distributed Artificial Intelligence*. MIT Press, 2000

Weisstein 2005
WEISSTEIN, E. W.: *Minkowski Metric*. From MathWorld–A Wolfram Web Resource. `http://mathworld.wolfram.com/MinkowskiMetric.html`. Version: 2005

Witten u. a. 1999
WITTEN, I. H. ; MOFFAT, A. ; BELL, T. C.: *Managing Gigabytes - Compressing and Indexing Documents and Images*. Second Edition. Morgan Kaufmann Publishers, 1999

Yang 1999
YANG, Y.: An Evaluation of Statistical Approaches to Text Categorization. In: *Information Retrieval* 1 (1999), Nr. 1/2, 69 – 90. `citeseer.ist.psu.edu/yang97evaluation.html`

Yoffe 2004
YOFFE, B.: *Kreativität in der Musik.* 2004. – Private Besprechung