Thesaurusföderationen: Ein Rahmenwerk für die flexible Integration von heterogenen, autonomen Thesauri

Zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften/Ingenieurwissenschaften

bei der Fakultät für Informatik
der Universität Fridericiana zu Karlsruhe eingereichte Dissertation

von

Ralf Nikolai
aus Hilden (Rheinland)

Tag der mündlichen Prüfung: 19.12.2002

Erster Gutachter: Prof. Dr. Peter C. Lockemann
Zweiter Gutachter: Prof. Dr. Rudi Studer

Danksagung

An dieser Stelle möchte ich all jenen meinen Dank aussprechen, die zum Gelingen dieser Arbeit beigetragen haben.

An erster Stelle möchte ich meine Doktorväter, Prof. Dr. P. Lockemann und Prof. Dr. R. Studer, nennen, die mit großem Engagement meine Arbeit aufmerksam und kritisch begleitet haben. Immer wieder erhielt ich neue Denkanstöße, die der Arbeit schließlich ihre Richtung gaben.

Diese Arbeit entstand im Forschungsbereich Datenbanksysteme (DBS) am Forschungszentrum Informatik (FZI). Allen meinen Kollegen und Studenten dort möchte ich für die fachlichen Diskussionen, die praktische wie auch die moralische Unterstützung und angenehme Arbeitsatmosphäre danken. Besondere Erwähnung gebührt hier Dr. Ralf Kramer, der als damaliger Abteilungsleiter den Startschuss für die Arbeit gab. Insbesondere er, Wassilios Kazakos, Dr. Ulrike Kölsch, Dr. Claudia Rolker, Andreas Schmidt, Dr. Rainer Schmidt und Peter Tomczyk haben durch ein stets offenes Ohr aber auch kontinuierliches Erfragen des Fortschritts zum Gelingen der Arbeit beigetragen.

Michael Gutbier möchte ich für sein großes Engangement während seiner Diplomarbeit aber auch darüber hinaus beim wiederholten Korrektur lesen der Arbeit ebenfalls ganz besonders danken. Unsere gemeinsame Liebe der deutschen Sprache trägt Früchte in einer sprachlich gelungenen Arbeit.

Auch außerhalb des universitären Umfeld gab es wesentliche „Erfolgsfaktoren“ dieser Arbeit.

Im zwischenzeitlich gegründeten Unternehmen disy Informationssysteme GmbH ermöglichten es die Kollegen, und hier ganz besonders zu nennen Sven Behrens, ausreichend Zeit für den erfolgreichen Abschluss der Arbeit zu finden.

Meine Eltern ermöglichten mir den Weg an die Universität und gaben mir stets das Gefühl, dass ich erreichen könne, was ich mir vorgenommen hatte. Meine Frau Dominique hat mich stets vorbehaltslos unterstützt und mir gerade in schwierigen Phasen den Rücken frei gehalten. Und unsere Kinder, Laura, Philipp, Lea Mara und Felix, haben ihren gestressten, häufig abwesenden (sowohl im Sinne von körperlich auch als geistig) Vater geduldig ertragen.

Inhaltsverzeichnis

Kapitel 1

Einleitung und Motivation

1.1 Motivation

Dem wachsenden Bedarf der „Informationsgesellschaft“ nach Informationen folgten in den letzten Jahren rasch wachsende Informationssysteme, die heterogene Informationen global verteilt und einfach zugreifbar vorhalten. Solche modernen Informationssysteme und datenintensiven Anwendungen können als eine wesentliche Komponente „verteilter Informationsumgebungen“ angesehen werden, die universellen Zugriff auf Informationen aus einer Vielzahl menschlicher Wissensgebiete ermöglichen. Charakteristische Eigenschaften derartiger großer Informationssysteme sind, dass sie auf großen, zum Teil autonomen Informationsquellen basieren, die häufig über offene Computernetze (lose) verbunden sind, eine große Anzahl von Benutzern unterstützen, eine Infrastruktur anbieten, die den einfachen Zugriff auf verschiedenen Dienste ermöglicht, und dass die Qualität dieser Dienste entscheidend für deren Erfolg ist. Von besonderer Bedeutung sind bei derartig großen zur Verfügung stehenden Datenmengen Dienste, die das gezielte Wiederauffinden von Informationen (Information Retrieval) ermöglichen.

Thesauri sind ein bewährtes Werkzeug, um diesen Prozess zu unterstützen. Sie bieten ein einheitliches und konsistentes Vokabular, das als Grundlage für semantisches Information Retrieval verwendet werden kann. Bei einem häufig fachübergreifenden Datenbestand, der auch mehrsprachig sein kann, sind traditionelle Fachthesauri, die in der Regel nur einsprachig vorliegen, aber nicht mehr ausreichend. Selbst in Dokumentenbeständen eines Fachinformationssystems finden sich oft Ausweitungen auf Begriffe angrenzender Fachgebiete. Es wird ein umfangreicheres und zugleich spezialisierteres Vokabular gefordert.

In Informationssystemen werden häufig jeweils an die besonderen Bedürfnisse der Benutzer angepasste Thesauri verwendet. Bei einer Integration der Informationssysteme wird auch eine Integration der Thesauri erforderlich, um den Benutzer beispielsweise dabei zu unterstützen, Informationen aus verschiedenen Informationsquellen zu erhalten. Die DG XIII der Europäischen Union hat bereits 1990 eine Liste von 1.000 häufig verwendeten Thesauri weltweit erstellt [Rad90]. Eine Verbindung dieser Thesauri wäre ein wichtiger Fortschritt bei der gemeinsamen Benutzung der Terminologie.

Da das Aufbauen eines neuen Thesaurus, aber auch die manuelle Integration existierender Thesauri immense Kosten verursacht (als Beispiel sei genannt, dass zur Erstellung einer initialen Version des Allgemeinen Umweltthesaurus GEMET mehrere Mannjahre benötigt wurden), sind neue Lösungen, die eine integrierte Sicht auf die Vokabulare mehrerer Thesauri unter Aufwendung finanziell vertretbarer Mittel ermöglichen, erforderlich. Zudem wird die klassische Form

der Integration von Thesauri der losen Kopplung von Informationssystemen nicht gerecht. Die erforderlichen technischen Voraussetzungen für das *logische* Zusammenbringen verteilter, heterogener Thesauri sind durch lokale und globale Vernetzung weitestgehend gegeben.

1.2 Zielbeschreibung

In dieser Arbeit soll ein Rahmenwerk für die lose Integration von heterogenen und autonomen Thesauri, *Thesaurusföderationen* genannt, erarbeitet werden. Das Konzept der Thesaurusföderationen soll den Anforderungen moderner Informationssysteme nach zugleich umfangreicheren und spezialisierteren Vokabularen unter Ausnutzung neuer technologischer Möglichkeiten gerecht werden. Der zu entwickelnde Integrations-Ansatz soll als Basis die mit großem Aufwand erstellten, bereits vorhandenen Thesauri (Komponententhesauri) verwenden und deren Vokabulare verknüpfen, so dass sie als ein Gesamtvokabular erscheinen.

Existierende Ansätze für einen integrierten Zugriff auf verschiedene Informationssysteme sowie der gleichzeitigen Verwendung verschiedener Terminologien basieren auf so genannten Multi-Thesaurus-Systemen. Ein wesentlicher Kritikpunkt an diesen Ansätzen ist der, dass jeweils nur Teilaspekte behandelt werden. Was fehlt, ist ein in ganzheitliches Rahmenwerk, das die Aspekte der Integration, der Behandlung von Konflikten und Unvollständigkeiten, der Verwendung im Information Retrieval und schließlich die Bewertung der Güte des integrierten Vokabulars betrachtet. Ein solches Rahmenwerk soll in dieser Arbeit erstmals erarbeitet werden. Dabei gilt es zu berücksichtigen, dass eine Überforderung des Benutzers durch die Komplexität des Gesamtvokabulars vermieden wird. U.a. soll das dynamische Ein-/Ausblenden von teilhabenden Thesauri unterstützt werden.

Die existierenden Ansätze der Multi-Thesaurus-Systeme berücksichtigen zudem nicht eine in verteilten Informationssystemen erstrebenswerte Autonomie der Thesauri und ihre häufig gegebene Heterogenität. Um diesen Anforderungen gerecht zu werden, soll sich unser Ansatz an den Konzepten föderierter Datenbanksysteme [SL90, Con97] orientieren, allerdings ohne die Einschränkung, ausschließlich von Datenbankverwaltungsystemen verwaltete Thesauri zu integrieren. Der Schwerpunkt soll hier auf der semantischen Integration liegen, die in föderierten Datenbanksystemen häufig nur ein Randthema ist. Neue Integrationsverfahren auf semantischer Ebene (Begriffsintegration), die im Gegensatz zu bekannten Ansätzen die Ergebnisse einer rechner-unterstützten Analyse der Inhalte und Güte der Thesauri berücksichtigen und entsprechend konfiguriert werden, sollen eine verbesserte semi-automatische Integration ermöglichen, ebenso erstmals eine Bewertung der Integrationsergebnisse. Diese Verfahren sollen die Reichhaltigkeit der Informationen in den Thesauri selbst ausnutzen [Rad90, S. 163] sowie auf weitere Wissensquellen zugreifen können, um den notwendigen menschlichen Einsatz zu minimieren.

Die Thesaurusföderation soll ihre Dienste als Mehrwertdienste anbieten und dazu auf die an der Föderation beteiligten heterogenen Komponententhesauri zugreifen, deren Autonomie erhalten bleibt.

Um den breiten Einsatz des entwickelten Ansatzes zu ermöglichen, soll das Konzept grundsätzlich fachgebietsunabhängig sein.

Auch wenn eine (semi-)automatische Integration unter Berücksichtigung der Autonomie einem durch manuelle Verfahren und Anpassung der beteiligten Thesauri entstandenem Super-Thesaurus unterlegen ist, ist dies möglicherweise die einzig praktikable Art und Weise, um ein flexibel skalierbares Multi-Thesaurus-System zu erstellen und zu pflegen.

1.3 Abgrenzung

Seit Anfang der 1990er Jahre sind Ontologien (vgl. Anhang D.2, S. 274) ein populäres Forschungsthema insbesondere im Bereich der Künstlichen Intelligenz [SFDB99]. Ontologien können als Thesauri betrachtet werden, die einerseits mehr Freiheitsgrade zulassen (hinsichtlich der zusätzlichen Verwendung von Attributen, um Begriffe zu beschreiben, sowie der prinzipiell uneingeschränkten Verwendung unterschiedlicher Relationstypen zwischen den Begriffen), andererseits aber einen höheren Grad an Formalisierung anstreben (durch formale Definitionen und Axiome über die Begriffe, Attribute und Relationen).

Die Anwendungsbereiche von Ontologien sind vielfältig, u.a. die Unterstützung der Integration von Informationen aus verschiedenen Quellen, die Unterstützung des Information Retrieval in verteilten Systemen sowie die Darstellung und Verarbeitung von fachspezifischem Wissen. In mancherlei Hinsicht existieren also Überschneidungen zwischen Ontologien und Thesauri.

Dennoch werden wir uns im Rahmen dieser Arbeit auf die ausschließliche Betrachtung der Integration von Thesauri beschränken. Dies hat folgende Gründe:

- Aufgrund des jungen Alters des Begriffes Ontologie in der Informatik wird der Begriff sehr vielfältig verwendet. Eine einheitliche Definition wurde bisher nur auf abstrakter Ebene gefunden. Konkret werden z.B. Datenbank-Schemata, objektorientierte Klassenhierarchien und Definitions-Thesauri als Ontologien bezeichnet [Sta00]. Der Begriff Thesaurus hingegen ist wesentlich konkreter definiert. Dies hat zur Folge, dass der allgemeingültige Umgang mit Thesauri wesentlich leichter fällt als ein allgemeingültiger Umgang mit Ontologien.

- Aufgrund der eingeschränkten Möglichkeiten für Relationstypen zwischen Begriffen kann die Semantik dieser Relationen wesentlich besser berücksichtigt werden, als bei den vielen Freiheiten, die eine Ontologie bietet. Das gleiche gilt für eine Analyse der Struktur und des Inhaltes von Thesauri respektive Ontologien. In dieser Hinsicht kann ein Thesaurus als ein Spezialfall einer Ontologie angesehen werden. Da es uns vermessen erscheint, allein im Rahmen dieser Arbeit das allgemeine Problem der Integration von Ontologien zu lösen, beschränken wir uns auf diesen einfacheren Spezialfall mit eingeschränkten Ausdrucksmöglichkeiten. Die Idee einer zukünftigen Erweiterbarkeit zur Integration von allgemeinen Ontologien steht bei unseren Betrachtungen also stets im Hintergrund.

- Schließlich ist die Entscheidung für uns auch eine Frage der Praxisrelevanz. Zum jetzigen Zeitpunkt gibt es eine Vielzahl von Thesauri, die in einer noch größeren Zahl von Information-Retrieval-Systemen und anderen Systemen eingesetzt werden. Die Anzahl der größeren Ontologien, deren Inhalte und Repräsentation deutlich über Thesauri hinausgehen und die zudem in Systemen eingesetzt werden, ist um Größenordnungen geringer. In der Zukunft mag sich dieses Verhältnis ändern, dennoch erscheint es uns für die nächsten Jahre, wenn nicht Jahrzehnte, als wichtig, auch mit den klassischen Thesauri umgehen zu können und Methoden und Werkzeuge zu haben, um diese zu integrieren. Selbst neueste Studien empfehlen Unternehmen, Taxonomien und Thesauri besser auszunutzen als direkt auf Ontologien „zu springen“ [Lin02]. Aus diesen Gründen werden auch aktuell neue Thesauri entwickelt, GEMET [CNR97] und die aktuellen Diskussionen zur Erstellung eines globalen Umweltthesaurus [FJ00] seien beispielhaft aufgeführt.

1.4 Aufbau der Arbeit

Die Arbeit ist wie folgt aufgebaut:

In Kapitel 2 wird die bereits in Abschnitt 1.2 dargestellte Zielbeschreibung detaillierter ausgeführt. Dies geht mit einer ausführlichen Problemanalyse und einer Zerlegung in Teilprobleme einher.

In Kapitel 3 wird der Stand der Forschung in Bezug auf eines der wesentlichen Probleme, die semantische Integration der Thesaurusstrukturen, die so genannte Begriffsintegration, diskutiert. Es werden Vor- und Nachteile der aus der Literatur bekannten Ansätze aufgezeigt. Die Methoden werden im Hinblick auf ihre Verwendbarkeit zur Lösung von Teilproblemen im vorgestellten Ansatz überprüft. Es wird eine Klassifikation von Multi-Thesaurus-Systemen erarbeitet, bei der die Idee der Thesaurusföderation eingeführt wird, um identifizierte Schwächen zu beseitigen.

Die Grundideen unseres Lösungsansatzes werden in Kapitel 4 vorgestellt. Neben dem generellen Aufbau werden einzelne Bausteine der Lösung erläutert. An diesen Bausteinen orientieren sich die nachfolgenden Kapitel. Für die eigentliche Begriffsintegration wird ein Phasenmodell als Vorgehensmodell eingeführt.

Grundlage des Lösungsansatzes sind formale Informationsmodelle für Thesauri und Thesaurusföderation. In Kapitel 5 werden daher Thesauri formal definiert. Dieser Formalismus wird die Basis zur Analyse und Integration von Thesauri. Das Informationsmodell für Thesauri wiederum wird Bestandteil des Informationsmodells für Thesaurusföderationen, das in Kapitel 6 anhand der Anforderungen an ein skalierbares und flexibles Multi-Thesaurus-System entwickelt wird. Als wesentliche Neuerung gegenüber klassischen Multi-Thesaurus-Systemen werden Invarianten und Konfliktmarkierungen eingeführt, die in gewissem Maße auch Widersprüche innerhalb der Thesaursföderation zulassen und eine situationsabhängige Auflösung dieser Widersprüche ermöglichen.

Die Akquise des Integrationswissens wurde als entscheidende Herausforderung bei der Integration von Thesauri identifiziert. Um diese Akquise innerhalb der verschiedenen Phasen des Vorgehensmodells adäquat zu unterstützen wird in Kapitel 7 eine Wissensakquisitionsarchtitektur entwickelt. Auf der Basis einer Blackboard-Architektur kann der entscheidende Vorteil der Trennung der Problemlösungstrategie von den Problemlösungsverfahren erzielt werden.

Die notwendigen Maßnahmen zur Vorbereitung eines Komponententhesaurus zur eigentlichen Integration werden in Kapitel 8 erläutert. Nachdem diese Vorbereitungsmaßnahmen ausgeführt worden sind, kann für die weiteren Phasen des Vorgehensmodells die Konformität des Komponententhesaurus mit dem Informationsmodell für Thesauri angenommen werden.

Die eingehende Analyse komplexer Komponentensysteme ist unverzichtbare Grundlage jedes Verfahrens, das zum Ziel hat, diese Komponentensysteme nicht nur oberflächlich und willkürlich zu verbinden, sondern eine bestmögliche Integration ihrer Elemente und Strukturen zu erreichen. Daher wird in Kapitel 9 ein Analyseverfahren für Thesauri erarbeitet und auf Thesauri, die integriert werden sollen, angewandt.

Mit dem in Kapitel 4 dargestellten Phasenmodell wird für die Problemlösungsstrategie bereits ein Rahmen vorgegeben. In Kapitel 10 wird dieses Vorgehensmodell für die komplexeste der Phasen, die Realisierungsphase, verfeinert. Die Realisierungsphase wiederum wird in Kapitel 11 detailliert betrachtet. Schwerpunkte bilden das Einbringen unterschiedlicher Lösungsverfahren sowie das Erzeugen und Festschreiben von Integrationswissen.

Nach dem Durchführen der Realisierungsphase sind die Komponententhesauri zu einer Thesaurusföderation integriert. Kriterien und Verfahren zur Bewertung der Güte dieser Föderation

werden in Kapitel 12 erarbeitet. Diese Kriterien dienen ebenfalls der Bewertung des gesamten Ansatzes.

Während die Kapitel 7 bis 12 Lösungen für die Entwicklung der Thesaurusföderation liefern, wird in Kapitel 13 der konkrete Einsatz einer solchen Föderation untersucht. Um diesen Einsatz zu unterstützen, wird eine Ausführungsmaschine entwickelt.

Das Kapitel 14 stellt schließlich die wesentlichen Ergebnisse der Arbeit zusammenfassend dar. Ausblickend werden mögliche Erweiterungen des Ansatzes vorgestellt.

In den Anhängen A bis C werden Einzelheiten der Implementierung dargestellt. Die im Rahmen dieser Arbeit verwendeten grundlegenden Begriffe aus den Bereichen Thesauri, semantische Netze, Ontologien, Graphentheorie und Linguistik werden, soweit sie für das Verständnis erforderlich sind, in einem Glossar in Anhang D definiert.

Kapitel 2

Problemanalyse

Einleitend motivieren wir die Anwendung von Thesauri als kontrollierte Indexierungs- und Recherchesprachen sowie Orientierungssysteme im Information Retrieval und stellen das grundlegende Thesaurusmodell vor. Anhand der Fallstudie eines föderierten Umweltinformationssystems werden in diesem Kapitel die Ausgangssituation und die Zielsetzung der Verwendung integrierter Recherche-Vokabulare geschildert. Von diesen Rahmenbedingungen ausgehend erfolgt eine detaillierte Analyse, die in einer Reihe von Forderungen an die Architektur, das Modell, die Erstellung und Handhabung adäquat integrierter Thesauri mündet.

2.1 Thesauri im Information Retrieval

2.1.1 Motivation

Das Ziel von Information-Retrieval-Systemen ist es, den Prozess des Wissenstransfers vom menschlichen Wissensproduzenten zum Informationsnachfragenden zu unterstützen (vgl. auch Anhang D.2, S. 272). An diesem Kommunikationsprozess sind Menschen zumindest in der Rolle als Wissensproduzent und als Informationsnachfragender beteiligt, häufig auch in der Rolle als Informationsanbieter, der für die Speicherung und das Verfügbar-Machen der Informationen verantwortlich ist. In allen drei Rollen bedienen sie sich der natürlichen Sprache[1]. Natürliche Sprache aber zeichnet sich besonders durch ihre Mehrdeutigkeit aus. Dies äußert sich auf der Ebene der Begriffe und Benennungen in einer Zuordnung von Begriffen zu ihren sprachlichen Repräsentanten (Benennungen), die nicht eindeutig ist (Synonymie, Homonymie und Polysemie). Es ergibt sich dadurch ein potenzielles Kommunikationsproblem, denn Information kann nur dann gefunden werden, wenn der folgende Transformationsprozess identische Benennungen des Informationsanbieters und des Informationsnachfragenden liefert (vgl. [Wer85]):

- Wissensproduzenten setzen bei der Erstellung von Dokumenten Begriffe in Benennungen um.
- Informationsanbieter interpretieren die Benennungen der Wissensproduzenten und transferieren sie in Begriffe, um die Dokumente zu verstehen.
- Bei einer Erschließung der Dokumente durch eine inhaltliche Indexierung werden die Begriffe von den Informationsanbietern wiederum in Benennungen umgesetzt.

[1] Selbst multimediale Dokumente beinhalten in der Regel einen sprachlichen Anteil, zumindest aber findet eine inhaltliche Erschließung in natürlicher Sprache statt.

- Der Informationsnachfragende hat für die Anfrage bestimmte Begriffe im Kopf, die er ebenfalls in Benennungen umsetzen muss.

Der Informationsnachfragenden wird nur dann zufrieden sein, wenn er die Benennungen des Dokumentes so interpretieren kann, dass er seine Begriffe wiederfindet.

Zur Kontrolle der Mehrdeutigkeiten werden Dokumentationssprachen entwickelt, die versuchen, die Nachteile der natürlichen Sprache auszugleichen. Während ältere Dokumentationssprachen wie Klassifikationen eher den Charakter einer künstlichen Sprache hatten, ist der Thesaurus eine natürlich-sprachliche Dokumentationssprache für ein bestimmtes Fachgebiet. Die Sprachkontrolle wird durch eine Menge ausgewählter Benennungen und ihrer Definitionen erreicht. Als Orientierungssystem repräsentiert ein Thesaurus Beziehungen zwischen den Begriffen, die einen Überblick über die begrifflichen Strukturen sowie das schnelle Auffinden einer beliebig allgemeinen oder spezifischen begrifflichen Einheit gewährleisten sollen [Wer85]. Somit kann ein Thesaurus in einem Information-Retrieval-System zugleich als Indexierungssprache zur inhaltlichen Erschließung der Dokumente, als Retrievalsprache zur Formulierung und Modifikation von Suchanfragen und als Orientierungssystem dienen. Der Thesaurus kann somit den oben beschriebenen Transformationsprozess in allen Phasen wesentlich erleichtern.

Im Lager der Informatiker und Informations- und Dokumentations-Wissenschaftler sind sowohl eine Reihe von überzeugten Thesaurus-Befürwortern zu finden als auch ebenso standfeste Thesaurus-Gegner. Hauptargumente der Thesaurus-Gegner sind, dass

- Thesauri in Zeiten der Volltext-Suche nicht mehr benötigt werden,
- die Kosten der Erstellung von Thesauri den Nutzen nicht rechtfertigen,
- eine Thesaurus-basierte Indexierung zu aufwendig ist und, falls sie manuell durchgeführt wird, den nicht-professionellen Benutzer aufgrund der Komplexität der Indexierungssprache überfordert.

Wir schließen uns der Meinung an, dass Indexierung wie auch das Abfassen von Kurzfassungen bei gegebener Recherchierbarkeit vollständiger Texte nicht mehr im bisherigen Maßstab erforderlich sind. Wichtig aber bleiben Thesauri bei der Rechercheunterstützung [Vie97, S. 5]. Es ist zudem zu beobachten, dass weitere Thesauri erstellt werden (interessant sind hier auch die anhand eines Dokumentenbestandes automatisch generierten Thesauri, vgl. z.B. [Vie97]), die in verschiedenen, auch neuen Anwendungsgebieten erfolgreich eingesetzt werden. Allein diese Tatsache ist eines der Hauptargumente der Thesaurus-Befürworter.

2.1.2 Definition

Eine genauere Definition für einen Thesaurus gibt die DIN-Norm 1463: „Ein Thesaurus ... ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Benennungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient. Er ist durch folgende Merkmale gekennzeichnet:

- Begriffe und Benennungen werden eindeutig aufeinander bezogen (terminologische Kontrolle)
- Beziehungen zwischen Begriffen (repräsentiert durch ihre Benennungen) werden dargestellt.“

Die Norm unterscheidet zwischen *Thesauri mit Vorzugsbenennungen* (*Deskriptoren*), die ausschließlich für die Indexierung und Suche zugelassen sind – andere äquivalente Begriffe (*Nicht-Deskriptoren*) verweisen auf die Deskriptoren und werden für die Indexierung und Suche nicht verwendet – und *Thesauri ohne Vorzugsbenennungen*, welche alle im Thesaurus befindlichen Benennungen für einen Begriff zur Indexierung zulassen. Im Folgenden wird unter Thesaurus immer ein Thesaurus mit Vorzugsbenennungen verstanden. Die Nicht-Deskriptoren sind hier zusätzliche Einstiegspunkte für den Benutzer.

Das *Vokabular* eines Thesaurus ist die Menge der Deskriptoren [DIN87]. Ein Thesaurus beinhaltet in der Regel das Vokabular *eines* Fachgebietes [Wer85].

Die unterschiedlichen Bedeutungen von Homonymen (Homographen) und Polysemen werden in Thesauri aufgelöst und durch näher bestimmende Zusätze kenntlich gemacht, z.B. Weide (Grünland) und Weide (Salix). Des Weiteren werden Allgemein-Deskriptoren, die zu allgemein sind und deshalb nicht selbst, sondern deren Unterbegriffe zur Indexierung verwendet werden sollen, durch einen Zusatz gekennzeichnet, z.B. Umweltschutz (benutze Unterbegriffe)[2].

Informationen über den beabsichtigten Gebrauch eines Deskriptors können in *Erläuterungen* gespeichert werden. Der Begriffsinhalt eines Deskriptors wird hauptsächlich durch die Beziehungen zwischen den Begriffen und Benennungen angegeben. Zusätzlich kann eine genaue Bestimmung des Begriffsinhaltes einem Deskriptor als *Definition* beigefügt werden. Ein Thesaurus, der für die überwiegende Zahl seiner Deskriptoren eine Definition enthält, ist somit eine Ausdrucksform für eine Ontologie.

Thesauri enthalten Relationen, die ein Netzwerk von Beziehungen zwischen Benennungen bzw. eine begriffliche Struktur darstellen [Wer85]. Beziehungen der ersten Art werden durch die *Äquivalenzrelation* und die *Benutze-Kombination-Relation* dargestellt. Durch die Äquivalenzrelation werden innerhalb des Thesaurus als bedeutungsgleich angesehene Benennungen zusammengefasst. Dazu werden für einen Deskriptor bedeutungsgleiche Nicht-Deskriptoren festgelegt. Die Benutze-Kombination-Relation (auch 1-zu-n-Äquivalenzrelation genannt) kann verwendet werden, um die Anzahl der Deskriptoren überschaubar zu halten, indem „Begriffe nicht durch die sie üblicherweise repräsentierten Benennungen, sondern durch Kombination bereits vorhandener Deskriptoren dargestellt werden“ [DIN87, S. 3]. Dazu wird von einem Nicht-Deskriptor auf eine Menge von Deskriptoren verwiesen.

Die *Hierarchierelation*, die weiter unterteilt werden kann in *Abstraktionsrelation* und *Bestandsrelation*, und die *Assoziationsrelation* werden zum Ausdrücken einer begrifflichen Struktur verwendet. Die Abstraktionsrelation (generische Relation) ist eine gerichtete Beziehung, die die Über- bzw. Unterordnung der Begriffe im Sinne des allgemeineren bzw. spezifischeren Begriffs darstellt. Die Bestandsrelation (partitive Relation) ist eine gerichtete Beziehung, die die Über- bzw. Unterordnung der Begriffe im Sinne des Ganzen bzw. eines Teiles darstellt.

Prinzipiell gibt es in einem Thesaurus keine Beschränkung der Anzahl der Ober- bzw. Unterbegriffe. Das ermöglicht innerhalb der Hierarchie potenziell an jeder Stelle [Wer85, S. 137]

- *Polyhierarchie*, d.h. jeder Begriff kann mehr als einen übergeordneten Begriff haben, und
- *Polydimensionalität*, d.h. jeder Begriff kann nach mehr als einem Unterteilungsgesichtspunkt in untergeordnete Deskriptoren (auch vermischt: Abstraktions- und Bestandsunterbegriffe) unterteilt werden.

Beziehungen zwischen Begriffen, die als wichtig erscheinen, aber weder eindeutig hierarchischer Natur noch äquivalent sind, werden durch die Assoziationsrelation ausgedrückt.

[2] Die Beispiele sind dem Thesaurus des Umweltbundesamtes [Bat94] entnommen.

In der Praxis verwendete Thesauri haben zum Teil Eigenschaften, die über die in der DIN-Norm vorgegebenen Eigenschaften hinausgehen. So besitzt etwa GEMET [CNR97] zwei zusätzliche Systeme, um die Deskriptoren zu ordnen:

- Ein Klassifikationsschema bestehend aus Super-Gruppen und Gruppen. Die Super-Gruppen (z.B. Natürliche und anthropologische Umwelt; Soziale Aspekte und Maßnahmen der Umweltpolitik) wurden aus Umweltmanagement-Gesichtspunkten angelegt und helfen bei der hierarchischen Strukturierung von GEMET. Die Gruppen (z.B. Biosphäre; Gesundheit und Ernährung) sind diesen Super-Gruppen zugeordnet und erlauben eine Kategorie- oder Disziplin-orientierte Perspektive. Die Deskriptoren können maximal einer Gruppe zugeordnet werden.

- Eine thematische Ordnung bestehend aus 40 (Umwelt-)Themen. Die Themen (z.B. Tourismus; Lärm und Vibrationen) sind orthogonal zu den Gruppen und übertragen den Thesaurus in eine Matrixstruktur. Ein Deskriptor kann bis zu vier Themen zugeordnet werden.

2.2 Fallstudie

2.2.1 Ausgangssituation Umweltinformationssysteme

Im Laufe der letzten Jahre wurden eine Reihe von Informationssystemen entwickelt, die Daten über den Zustand der Umwelt verwalten, zugänglich und auswertbar machen. Besonders starke Impulse erhielt die Entwicklung dieser Systeme nach dem Auftreten bedeutender Unfälle mit großflächigen oder langfristigen Schäden an Mensch und Umwelt (z.B. dem Reaktorunfall in Tschernobyl vom 26.04.1986, bei dem große Mengen radioaktiver Strahlung entwichen [Int96]) oder dem Entdecken von globalen Naturphänomenen, deren Entstehen auf anthropogenen Einfluss zurückgeführt wird (z.B. globale Klimaveränderungen, Ozonloch in der Atmosphäre). Bereits die Aufzählung dieser Beispiele gibt Hinweise darauf, dass diese Informationssysteme in vielfältiger Hinsicht heterogen sind:

Benutzergruppen: Bis vor wenigen Jahren war die mit Abstand bedeutendste Benutzergruppe von Umweltinformationssystemen Mitarbeiter in Behörden. Der Zugriff auf Umweltinformationen stellt für diese Benutzergruppe z.B. die Basis für gesetzgeberische Entscheidungen, die Überprüfung der Einhaltung umweltpolitischer Ziele, die Erforschung komplexer Zusammenhänge und das Erkennen von Veränderungen der natürlichen Umwelt dar. Bereits innerhalb dieser Gruppe ist eine große Heterogenität erkennbar: Die Benutzer kommen aus den unterschiedlichsten Fachgebieten, besitzen unterschiedlich stark ausgeprägtes Expertenwissen und verwenden die Umweltdaten mit unterschiedlichen Zielsetzungen.

Seit 1990 haben die Mitgliedsstaaten der Europäischen Union die Anforderungen der EU-Direktive des freien Zugangs zu von Behörden erhobenen oder verwalteten Umweltinformationen zu erfüllen (EU-Direktive 90/313/EC). Da beinahe zeitgleich mit dieser Direktive durch das World Wide Web [Con02] neue technologische Möglichkeiten entstanden, wurden und werden immer mehr Umweltinformationssysteme öffentlich zugänglich gemacht (s. z.B. [KNK+97, NKK+99, GRW97]). Damit erhalten neben dem Bürger mit vornehmlich privatem Interesse Benutzer aus Industrie und Handel, Nicht-Regierungsorganisationen und Universitäten Zugang. Die Heterogenität der Benutzergruppen ist somit noch einmal deutlich gewachsen.

Inhalte: Umweltinformationssysteme verwalten u.a. medienorientierte Daten (z.B. Boden-, Wasser-, Luftmesswerte), umweltbeeinflussende Daten (z.B. Daten über Verkehr, Abfall), Daten über den Zustand natürlicher Ressourcen (z.B. Artenvielfalt, Biomasse) und aggregierte Daten (z.B. Umweltberichte, thematische Karten) aus potenziell allen Wissenschaftsgebieten. Diese thematische Vielfalt resultiert aus dem Umwelteinfluss, den beinahe jede Aktivität ausübt.

Datenformate: Umweltinformationssysteme verwalten informationstragende Objekte in unterschiedlichen Formaten, z.B. Dokumente, Messreihen, Karten, Satellitenbilder, Simulationsmodelle und -ergebnisse.

Thesauri haben sich in solchen Umweltinformationssystemen als wertvolles Werkzeug bewährt, um das Wiederauffinden dieser heterogenen Informationen zu unterstützen. Die unterschiedlichen informationstragenden Objekte werden von den Informationsanbietern mit Begriffen des Thesaurus einheitlich indexiert. Die Informationssuchenden können sich des Begriffsnetzes in dem Thesaurus bedienen, um ihre Anfragen an das System möglichst präzise zu formulieren bzw. zu verfeinern. Das System wiederum kann bei der Anfragebearbeitung anhand der Beziehungen der Begriffe Anfrageerweiterungen durchführen, um die Ergebnismenge zu vergrößern.

Die Beantwortung von Fragestellungen aus dem Umweltbereich erfordert häufig Informationen nicht nur aus einem einzigen Informationssystem, sondern aus mehreren. Eine solche Fragestellung kann z.B. lauten, inwiefern eine bestimmte Anbaumethode unter bestimmten klimatischen und geologischen Bedingungen Einfluss auf die Produktivität, den Boden und das Grundwasser hat. Zur Beantwortung dieser Fragestellungen sind generelle Aussagen über die Entwicklungen der Belastung von Böden und Grundwasser wichtig, es ist die Feststellung so bewirtschafteter Gebiete erforderlich und schließlich werden Grundwasser- und Bodenmesswerte aus diesen Gebieten benötigt. Um die gesamten erforderlichen Informationen zu erhalten, war es bis Anfang der 90er Jahre erforderlich, diese bei den datenhaltenden Stellen zu bestellen. Aufgrund der Möglichkeiten der Web-Technologie wurde damit begonnen, diese Systeme öffentlich zugänglich zu machen. Auch wenn dies nun für einen Teil der Systeme gelungen ist, stellt sich die Situation immer noch als unbefriedigend dar. Werden Informationen aus verschiedenen Systemen benötigt, ist es erforderlich, jedes dieser Systeme separat zu befragen. Dabei wird der Benutzer mit unterschiedlichen Benutzerschnittstellen, unterschiedlichen Anfragemöglichkeiten und unterschiedlichen Thesauri konfrontiert.

Aktuelle Forschungsprojekte beschäftigen sich daher mit der Integration dieser Systeme, um einen einheitlichen Zugang zu mehreren Systemen gleichzeitig anbieten zu können [KKN+96, KKN+97, AB97, KR98, PCN00]. Da die Autonomie der beteiligten Systeme in der Regel erhalten bleiben soll, werden häufig Föderationsarchitekturen, deren grundlegender Aufbau sich an der I3-Referenzarchitektur orientiert [Wie96], verwendet und auf der Basis von Web-Technologien realisiert. Damit kann zwar die technische Integration dieser Systeme als gelöst angesehen werden, keineswegs aber die semantische Integration. Wesentlicher Bestandteil einer semantischen Integration für eine systemübergreifende Recherche ist eine Integration der in den verschiedenen Informationssystemen verwendeten Thesauri [KNR+97].

2.2.2 Auswahl der Thesauri

Die Herausforderungen einer solchen Integration der Thesauri zu einem so genannten Multi-Thesaurus-System sollen im Folgenden detailliert analysiert werden. Damit diese Analyse die in der realen Welt anzutreffenden Probleme erkennt, wurde von uns entschieden, sie anhand von konkreten Thesauri durchzuführen.

2.2.2.1 Auswahlkriterien

Kriterien für die Auswahl der Thesauri zur Verwendung in der Fallstudie waren:

Praktischer Einsatz in Umweltinformationssystemen: Wie bereits erläutert, sind die in Umweltinformationssystemen verwalteten Informationen in vielfältiger Hinsicht heterogen. Es existiert eine entsprechende Vielfalt an Thesauri, die bei der Integration von Umweltinformationssystemen zu berücksichtigen ist. Die für unsere Zwecke ausgewählten Thesauri sollen daher im praktischen Einsatz zur Retrieval-Unterstützung in Umweltinformationssystemen sein.

Inhaltliche Überlappungen: Die Begriffswelten der Thesauri sollen wesentliche Überschneidungen aufweisen, da erwartet wird, dass eine Integration fachlich verwandter Thesauri die komplexere Herausforderung darstellt als die Integration isolierter oder inhaltlich nur sehr lose verwandter Thesauri.

Heterogenität: Die Spannweite soll von minimalen Thesauri bis hin zu großen Systemen in verschiedenen Sprachen reichen.

Aktualität: Die Thesauri sollen den aktuellen Stand der Fachterminologie beinhalten und in aktuellen Systemen eingesetzt werden.

Freie Verfügbarkeit: Die Thesauri sollen (auf Anfrage) von den Organisationen zur Verfügung gestellt werden. Dies ermöglicht das Durchführen lokaler Experimente ebenso wie die Nachvollziehbarkeit der in dieser Arbeit vorgestellten Resultate.

2.2.2.2 GEMET, AGROVOC und GCMD Parameter Valids

Anhand der oben aufgeführten Kritierien wurden schließlich folgende Thesauri ausgewählt:

GEMET: Der General Multilingual Environmental Thesaurus (GEMET) [CNR97] wird u.a. im Catalogue of Data Sources (CDS) [ETC98] der Europäischen Umweltagentur (EUA) als Katalogsystem für Umweltdaten von europäischem Interesse eingesetzt.

AGROVOC: Der agrarwissenschaftliche Thesaurus AGROVOC [ZRSJ+92] wird im International Information System for the Agricultural Sciences and Technology (AGRIS) der Food and Agricultural Organisation (FAO) der Vereinten Nationen (UN) eingesetzt. AGRIS ist ein Informationssystem über Literatur aus dem gesamten agrarwissenschaftlichen Bereich (u.a. Agrartechnologie, Meereskunde und Fischerei, Lebensmittelwissenschaften und Umweltverschmutzung).

GCMD Parameter Valids: Die GCMD Parameter Valids (GCMD bedeutet Global Change Master Directory) [GCM99] werden u.a. innerhalb des INFEO-Systems [Cen00], das Zugang zu Erdbeobachtungsdaten liefert, als Thesaurus verwendet.

2.2.3 Zielsetzung

Zielsetzung ist es, dem Informationssuchenden eine übergreifende Recherche in allen integrierten Informationssystemen bei Verwendung der ihm vertrauten Recherchesprache, also des ihm vertrauten Thesaurus, zu ermöglichen. Dabei soll eine integrierte Sicht auf das Gesamtvokabular der beteiligten Thesauri, die *Komponententhesauri* genannt werden, hergestellt und angeboten

werden, ohne dass die Autonomie dieser Komponententhesauri in Frage gestellt wird. Das integrierte *Multi-Thesaurus-System* soll flexibel und erweiterbar sein, d.h. Thesauri sollen aus dem System herausgenommen und dem System hinzugefügt werden können.

Anfragen, die ein solches Multi-Thesaurus-System beantworten können soll, sind z.B.:

- Welche Begriffe gibt es, die etwas mit *environmental impact* zu tun haben?
- Wie ist der gegebene Begriff *environmental impact* im Multi-Thesaurus-System definiert? Durch welche Benennungen wird der gegebene Begriff *environmental impact* im Multi-Thesaurus-System repräsentiert?
- Welche spezifischeren Begriffe gibt es im System zu dem gegebenem Begriff *environmental impact*?
- Welcher Begriff oder welche Kombination von Begriffen soll verwendet werden, wenn mit dem Begriff *environmental impact* aus dem Multi-Thesaurus-System in einem Informationssystem i gesucht werden soll, in dem der Komponententhesaurus t zur Indexierung verwendet wird?

2.3 Analyse

Anhand einer detaillierten Analyse der Fallstudie und der in Abschnitt 2.2.3 aufgeführten Zielsetzung werden in diesem Abschnitt Anforderungen an eine adäquate Integration der Thesauri hergeleitet.

2.3.1 Informationsmodell für integrierte Thesauri

Im Gegensatz zu eigenständigen Thesauri gibt es für integrierte Thesauri bisher keinerlei Normen, die Vorgaben für ein Informationsmodell enthalten. Daher ist eine Analyse der Anforderungen an ein solches Modell für integrierte Thesauri erforderlich. Als sich aus der Fallstudie ergebende Randbedingung gilt, dass die Eigenständigkeit der Thesauri, die integriert werden, erhalten bleiben soll.

2.3.1.1 Thesauri

2.3.1.1.1 Komponententhesauri Die an dem Multi-Thesaurus-System teilnehmenden Thesauri werden Komponententhesauri genannt. Da den Komponententhesauri ihre Eigenständigkeit nicht genommen werden soll, dürfen direkt an den Komponententhesauri keine Änderungen vorgenommen werden (*Autonomieerhaltung*).

Es wird davon ausgegangen, dass die Komponententhesauri den Normen entsprechen. Das Informationsmodell für integrierte Thesauri muss daher ein Modell für Normen-konforme Thesauri als Bestandteil haben.

Über Normen hinausgehende Eigenschaften werden zwar während des Integrationsprozesses berücksichtigt (vgl. Abschnitt 2.3.2.2.3), spielen aber bei der Verwendung des integrierten Systems keine Rolle mehr und brauchen daher nicht im Modell für integrierte Thesauri berücksichtigt werden.

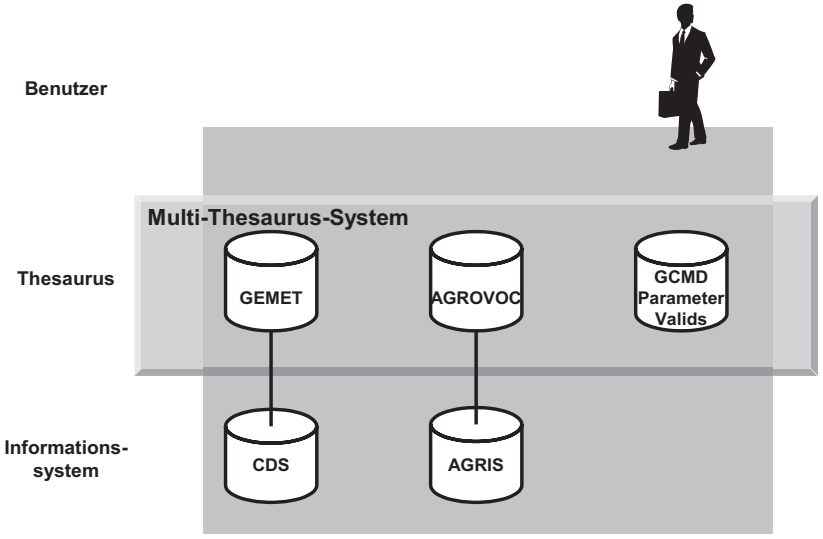

Abbildung 2.1: Verwendung eines Nichtindexierungsthesaurus

2.3.1.1.2 Indexierungs- und Nichtindexierungsthesauri Angenommen, ein Benutzer ist mit den GCMD Parameter Valids vertraut, da diese in verschiedenen Informationssystemen zur Indexierung verwendet werden. Nun möchte dieser Benutzer in den Systemen CDS und AGRIS (vgl. Abschnitt 2.2.2.2) recherchieren, die als föderiertes Informationssystem eine übergreifende Recherche einschließlich der integrierten Thesauri GEMET und AGROVOC anbieten. Tritt dieser Fall häufiger auf – z.B. weil ein Thesaurus weit verbreitet ist, der Zugriff auf die Informationssysteme, die diesen Thesaurus verwenden, aber nicht Bestandteil des föderierten Informationssystems ist – soll solch ein Thesaurus ebenfalls in das Multi-Thesaurus-System integriert werden können. Dann muss zwischen Indexierungsthesauri, also Thesauri, die innerhalb des föderierten Informationssystems zur Indexierung informationstragender Objekte verwendet werden (im Beispiel: GEMET und AGROVOC), und Nichtindexierungsthesauri (im Beispiel: GCMD Parameter Valids) unterschieden werden (vgl. Abbildung 2.1).

2.3.1.1.3 Auswahl von Teilmengen Bei dem Zugriff auf ein föderiertes Informationssystem sind nicht immer Informationen aus allen Komponentensystemen erforderlich. Beispielsweise hat ein Benutzer zwar Interesse an Daten zur landwirtschaftlichen Nutzung von Flächen, nicht aber an Satellitenbildern. In der Fallstudie ist also der Zugriff auf den CDS und AGRIS erforderlich, nicht aber auf INFEO. Entsprechend soll es auch möglich sein, von dem Multi-Thesaurus-System nur die relevanten Thesauri präsentiert zu bekommen (GEMET und AGROVOC). Nicht-relevante Thesauri sollen ausgeblendet werden können (GCMD Parameter Valids) (vgl. Abbildung 2.2).

2.3.1.2 Inter-Thesaurus-Relationen

2.3.1.2.1 Relationstypen Bereits die DIN-Norm zur Erstellung mehrsprachiger Thesauri [DIN93a] gibt wichtige Hinweise darauf, welche Arten von Beziehungen zwischen Begriffen auftreten können. Zwar beschränkt sich die DIN-Norm auf verschiedene Arten von Äquivalenzen (genaue Äquivalenz, unscharfe Äquivalenz, Teil-Äquivalenz, 1:n-Äquivalenz) zwischen Benennungen in verschiedenen Sprachen. Aber schon in der DIN-Norm werden zur Darstellung Beziehungen zwischen Begriffen *eines* Thesaurus vorgeschlagen. Die DIN-Norm als Ausgangsbasis

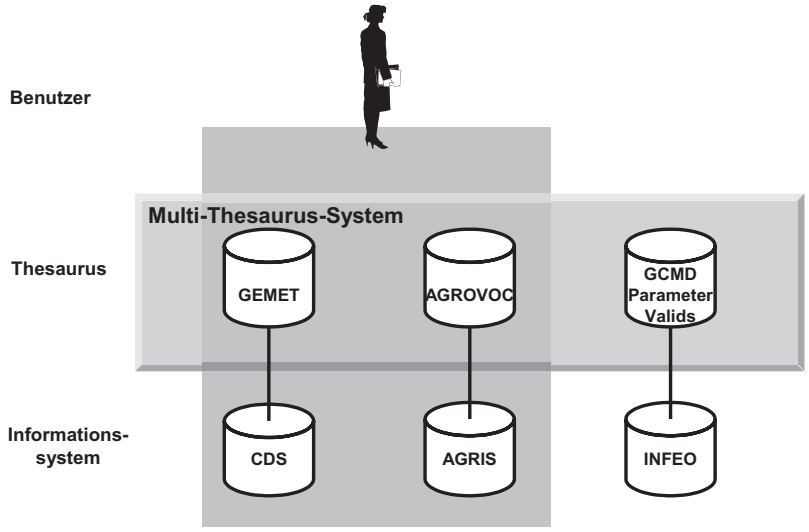

Abbildung 2.2: Auswahl einer relevanten Teilmenge von Thesauri

nehmend, wurden die Thesauri AGROVOC, GEMET und GCMD Parameter Valids auf weitere Beziehungstypen hin analysiert. Dabei wurden folgende potenzielle Beziehungstypen identifiziert, die hier mit Beispielen belegt werden:

Äquivalenzbeziehungen: Beziehungen zwischen Begriffen, die innerhalb des Multi-Thesaurus-Systems als bedeutungsgleich angesehen werden. Unterschieden werden weiterhin:

1:1-Äquivalenzbeziehungen (genaue Äquivalenz): Ein Begriff aus einem Thesaurus ist bedeutungsgleich mit genau einem Begriff aus einem anderen Thesaurus. Beispiele hierfür sind *fish farming* aus GEMET und *fish culture* aus AGROVOC sowie eine große Anzahl lexikalisch identischer Benennungen (ggf. nach einer Normierung), die in beiden Thesauri denselben Begriff benennen (u.a. *aquaculture*, *genetics* und *drainage system* bzw. *drainage systems*).

1:n-Äquivalenzbeziehungen (Begriffszerlegung): Ein Begriff ist bedeutungsgleich mit der *konjunktiven* Verknüpfung von zwei oder mehr Begriffen aus einem anderen Thesaurus. Der Begriff *fish factories* aus AGROVOC entspricht der Kombination der Begriffe *industrial plant* und *fishing industry* aus GEMET.

Hierarchiebeziehungen: Hierarchiebeziehungen sind gerichtete Beziehungen zwischen über- und untergeordneten Begriffen aus verschiedenen Thesauri. Eine genauere Betrachtung unterscheidet weiterhin:

Generische Beziehungen: Die generische Hierarchiebeziehung drückt die Beziehung zwischen einem übergeordneten Begriff mit einer umfassenderen (generischeren) Bedeutung als der untergeordnete Begriff aus (Teil-Äquivalenzen). Als Beispiel seien *marine sciences* aus AGROVOC als übergeordneter Begriff und *marine biology* aus GEMET als untergeordneter Begriff genannt.

Partitive Beziehungen: Bei der partitiven Hierarchiebeziehung ist der übergeordnete Begriff ein *Ganzes*, das den untergeordneten Begriff als einen *Bestandteil* enthält. Der Begriff *bicycle* aus GEMET enthält als Bestandteile die Begriffe *brakes*, *seats* und *wheels* aus AGROVOC.

Instanzbeziehung: In einer Instanzbeziehung stehen zwei Begriffe, wenn der untergeordnete Begriff eine Instanz des übergeordneten Begriffs ist. Ein Beispiel für eine solche Beziehung ist der Begriff *north america* aus AGROVOC, der eine eindeutig identifizierbare Ausprägung des Begriffes *continent* aus GEMET ist.

Weitere Beziehungen: Es existieren weitere Beziehungen zwischen Begriffen aus verschiedenen Thesauri, die als wichtig erscheinen, aber weder eindeutig hierarchischer Natur noch äquivalent sind. Beispielhaft seien *meteorological phenomenon* aus GEMET und *atmospheric disturbances* aus AGROVOC aufgeführt.

Diese Vielfalt von potenziellen Beziehungen zwischen Begriffen aus verschiedenen Thesauri sollen innerhalb eines Multi-Thesaurus-System durch Inter-Thesaurus-Relationen ausgedrückt werden können.

2.3.1.2.2 Abgelehnte Beziehungen Während des Integrationsprozesses können algorithmisch erkannte Inter-Thesaurus-Relationen vorgeschlagen werden, die von einem menschlichen Experten abgelehnt werden. Als Beispiel sei die Benennung *lime* genannt, die in beiden Thesauri vorkommt. Diese Benennung aber repräsentiert in GEMET einen Baustoff (Kalk), in AGROVOC hingegen eine Zitrusfrucht (Limone). Eine anhand der lexikalischen Gleichheit vorgeschlagene Äquivalenzbeziehung wird daher von einem Experten abgelehnt werden.

Um ein erneutes Vorschlagen während des weiteren Integrationsprozesses oder während der Integration von aktualisierten Versionen der Thesauri zu vermeiden, sollen solche abgelehnten Beziehungen vermerkt werden. Des Weiteren ist eine Begründung der Ablehnung (im obigen Beispiel: Homonym) sowohl bei der Integration weiterer Komponententhesauri hilfreich als auch für den Benutzer, der somit z.B. auf die *Homonymproblematik* aufmerksam gemacht werden kann (die potenzielle Mehrdeutigkeit ist bei dem Beispiel *lime* weder in AGROVOC noch in GEMET gekennzeichnet).

2.3.1.3 Begriffe

2.3.1.3.1 Föderierte Begriffe Ein Begriff, der Bestandteil eines aus autonomen Komponententhesauri bestehenden Multi-Thesaurus-Systems ist, wird föderierter Begriff genannt. Er soll die in Abschnitt 2.3.1.2.1 geforderten Relationen eingehen können, möglichst ohne dabei die in Abschnitt 2.3.1.4 geforderten Invarianten zu verletzen. Stehen zwei Begriffe aus verschiedenen Thesauri in einer 1:1-Äquivalenzbeziehung, sollen als Vorzugsbezeichner die entsprechenden Deskriptoren der Thesauri gewählt werden. Eine weitere Einschränkung auf nur einen dieser Deskriptoren (entsprechend der Einschränkung auf einen Deskriptor bei Thesauri mit Vorzugsbezeichnern) erscheint nicht sinnvoll, da

- in einem Indexierungsthesaurus (vgl. Abschnitt 2.3.1.1.2) jeder Deskriptor zur Indexierung im föderierten Informationssystem verwendet werden kann,
- bei Nicht-Indexierungsthesauri gerade der Gebrauch des bekannten Vokabulars, also der Deskriptoren aus dem Nicht-Indexierungsthesaurus, gefordert wird und
- die Bedeutung eines föderierten Begriffes vom Umfang und Inhalt eines Begriffes aus *einem* Thesaurus abweichen kann und dies durch das Aufführen der gleichberechtigten Deskriptoren verdeutlicht werden kann.

2.3.1.3.2 Ergänzende Begriffe Ziel der Integration von Thesauri ist das Herstellen von Verbindungen und nicht das Ergänzen neuer Inhalte. Es können jedoch Fälle auftreten, in denen das Hinzufügen von Begriffen die Qualität des Integrationsergebnisses verbessert:

Darstellen von Schwesterbeziehungen: Es werden Begriffe identifiziert, die Abstraktionsunterbegriffe unter einem gemeinsamen Oberbegriff sein könnten. Dieser gemeinsame Oberbegriff existiert aber in keinem Komponententhesaurus. Statt die beiden Begriffe durch eine Assoziationsbeziehung miteinander zu verbinden, deren Semantik diesen Zusammenhang nur unzureichend ausdrücken würde, ist es sinnvoll, diesen gemeinsamen Oberbegriff einzuführen. Als Beispiel seien *administrative organisation* in GEMET und *international organisation* in AGROVOC genannt sowie der zu ergänzende Begriff *organisation*. Nur durch die jetzt möglichen Abstraktionsbeziehungen wird deutlich, dass es sich bei beiden Begriffen um Organisationen handelt.

Ausgleich von Abstraktionsniveauunterschieden: Durch die Verbindung von Abstraktionsstrukturen können Abstraktionsniveauunterschiede zwischen den Schwesterknoten entstehen oder aber sehr unterschiedliche semantische Aspekte dargestellt sein. In einem solchen Fall ist das Einführen von Zwischenbegriffen wünschenswert, um die Aufteilung eines Begriffes in seine Abstraktionsunterbegriffe möglichst homogen zu halten. Das in Abbildung 2.3 dargestellte Beispiel illustriert dies: *products* hat als Abstraktionsunterbegriffe *agricultural products* und *chemical products* in GEMET. In AGROVOC existiert der Begriff *byproducts* u.a. mit den Unterbegriffen *meat byproducts* und *brewery byproducts*. Wenn nun *byproducts* als direkter Abstraktionsunterbegriff von *products* aufgeführt wird, ist es aus Gründen eines einheitlichen Abstraktionsniveaus der Schwesterknoten sinnvoll, einen Begriff *primary products* ebenfalls als direkten Abstraktionsunterbegriff zu ergänzen. Dieser wird dann Abstraktionsoberbegriff von *agricultural products* und von *chemical products*.

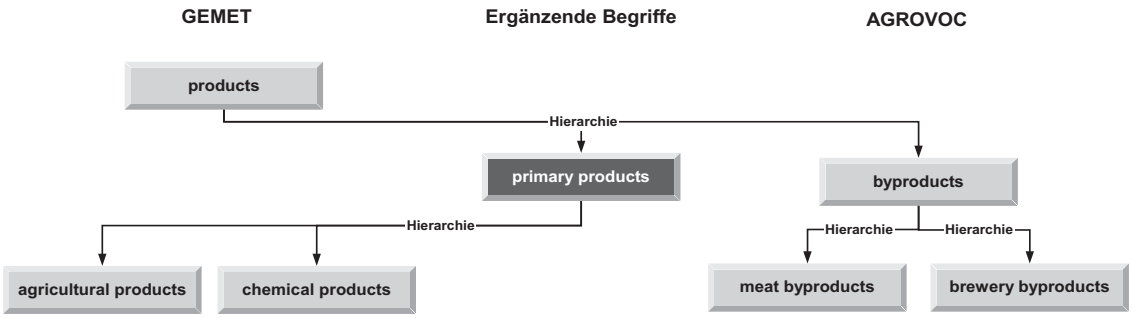

Abbildung 2.3: Einfügen eines ergänzenden Begriffs zum Ausgleich von Abstraktionsniveauunterschieden von Schwesterknoten

Die Beispiele verdeutlichen die Notwendigkeit, dass es möglich sein soll, während der Integration auch Begriffe, die in keinem der Komponententhesauri enthalten sind, ergänzen zu können.

2.3.1.4 Invarianten

Durch das Ausdrücken der in Abschnitt 2.3.1.2 aufgeführten Beziehungen zwischen Begriffen aus verschiedenen Thesauri können Verstöße gegen die Invarianten eines einzelnen Thesaurus entstehen (die im Rahmen dieser Arbeit geltenden Invarianten werden in Abschnitt 5.2.3 hergeleitet).

Beispielsweise wird für einen Komponententhesaurus die Redundanzfreiheit der Hierarchierelation gefordert. Selbst wenn diese Forderung von allen beteiligten Komponententhesauri erfüllt wird, kann das Etablieren von Beziehungen zwischen Begriffen verschiedener Komponententhesauri dazu führen, dass innerhalb des Multi-Thesaurus-Systems diese Forderung nicht mehr erfüllt ist. Abbildung 2.4 stellt zwei Beispiele für einen solchen Verstoß da, von denen das linke erläutert wird: Das Einführen einer Äquivalenzbeziehung zwischen den den Begriffen *government* und einer Hierarchiebeziehung zwischen den Begriffen *form of government* aus GEMET und *regional government* aus AGROVOC führt dazu, dass die Hierarchiebeziehung zwischen *government* und *regional government* in AGROVOC redundant wird.

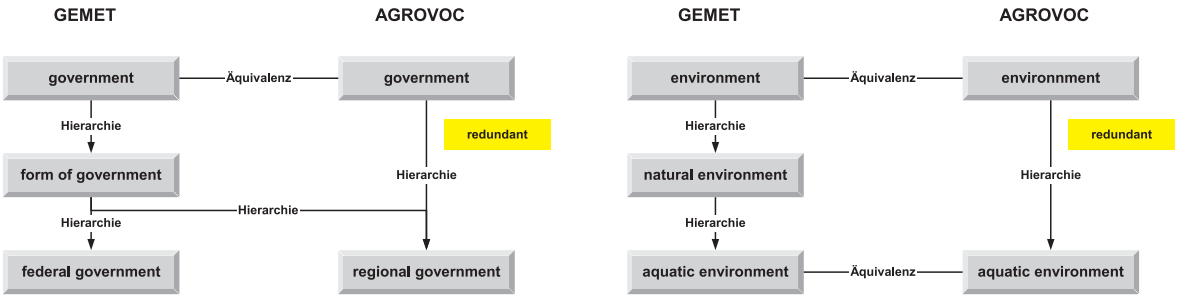

Abbildung 2.4: Beispiele für Hierarchierelationen, deren Redundanz durch das Ausdrücken von Inter-Thesaurus-Beziehungen verursacht wird

Für ein Multi-Thesaurus-System gilt es ebenso wie für einen einzelnen Thesaurus, den gültigen Entwurfsraum explizit festzulegen und diesen auf möglichst sinnvolle und widerspruchsfreie Entwürfe einzuschränken. Um dies zu erreichen, gilt es zu entscheiden, welche Invarianten eines einzelnen Thesaurus übertragen werden können, welche in abgeschwächter Form gelten und welche Konflikte hingenommen werden.

2.3.1.5 Konflikte

Wenn innerhalb eines Multi-Thesaurus-Systems Verstöße gegen Invarianten hingenommen werden sollen, soll eine Markierung dieser so genannten Konflikte möglich sein. Dazu wird gefordert, dass innerhalb des Modells eines Multi-Thesaurus-Systems Konflikte unterschiedlichen Typs zusammen mit den Konfliktursachen ausgedrückt werden können (s. auch Abschnitt 2.3.2.4).

2.3.2 Begriffsintegration

Zentrale Aufgabe der Thesaurusintegration ist es, die Beziehungen zwischen den Begriffen der Komponententhesauri möglichst vollständig und korrekt aufzufinden und zu etablieren, also die eigentliche Begriffsintegration zu betreiben (vgl. Abschnitte 2.3.1.2 und 2.3.2.6). Dabei sollen die spezifischen Charakteristika der Komponententhesauri angemessen berücksichtigt, gegebenenfalls ergänzende Begriffe eingeführt und schließlich alle Konflikte markiert und entsprechend behandelt werden. Da bei der Begriffsintegration die Semantik der Begriffe und Relationen eines Thesaurus wesentlich ist, müssen Verfahren, die dies unterstützen, so angelegt sein, dass sie es menschlichen Experten ermöglichen, ihr semantisches Wissen einzubringen.

In den folgenden Abschnitten werden die Teilaufgaben der Begriffsintegration näher analysiert.

2.3.2.1 Analyse der Komponententhesauri

Die eingehende Analyse komplexer Komponentensysteme ist unverzichtbare Grundlage jedes Verfahrens, das zum Ziel hat, diese Komponentensysteme nicht nur oberflächlich und willkürlich zu verbinden, sondern eine bestmögliche Integration ihrer Elemente und Strukturen zu erreichen.

Thesauri können – selbst innerhalb der in Abschnitt 2.1 eingeführten und später in Kapitel 5 formalisierten Definitionen – sehr unterschiedlich ausgeprägt sein. Grund dafür ist, dass sämtliche Definitionen die Inhalte in Form von Begriffen, Benennungen und Relationsausprägungen in keiner Weise festlegen. Damit werden den Thesaurusentwicklern viele Freiheiten für einen auf ein bestimmtes Gebiet und eine bestimmte Anwendung maßgeschneiderten Thesaurus gelassen. So kann z.B. ein Chemiethesaurus vorwiegend Formeln als Benennungen enthalten, während ein anderer Thesaurus allgemeine politische Benennungen enthält und wiederum ein anderer die bei dem Umgang mit Altstandorten gebräuchliche, sehr spezielle Fachterminologie. Breite und Tiefe der Vokabulare zweier Thesauri selbst aus demselben Fachgebiet können sehr unterschiedlich sein. Thesauri können bereits präkoordinierte Begriffe enthalten (d.h. Komposita und adjektivische Wortgruppen als Benennungen für verknüpfte begriffliche Einheiten) oder aber die Kombination von Begriffen im Wesentlichen der Postkoordination überlassen (d.h. die Begriffskombination geschieht bei Bedarf durch den Benutzer). Unterschiedlich können auch die Relationen ausgeprägt sein. So enthält z.B. AGROVOC im Verhältnis zu Hierarchierelationen sehr viele Assoziationsrelationen, bei GEMET ist dieses Verhältnis deutlich geringer.

Diese Charakteristika haben einen entscheidenden Einfluss auf das Vorgehen bei der Integration. Im oben aufgeführten Beispiel soll den AGROVOC-Assoziationsrelationen eine größere Beachtung zukommen, da sie bei den AGROVOC-Entwicklern offensichtlich eine wichtige Rolle gespielt haben. Insgesamt ist eine eingehende Analyse der Komponententhesauri zur Feststellung der Vorbedingungen der Integration ebenso wie zu einer angemessenen Integration erforderlich. Bestandteil einer solchen Analyse soll auch eine Kompatibilitätsanalyse der Komponententhesauri sein, damit es möglich wird, die Integrationserwartungen vor der eigentlichen Integration zu spezifizieren. Eine Bewertung der (vorläufigen) Integrationsergebnisse kann dann anhand dieser Erwartungen geschehen.

Da eine manuelle Analyse der Komponententhesauri, die Inhalt, Struktur und Güte beschreiben soll, aufgrund der Komplexität der Thesauri (AGROVOC beispielsweise besteht aus ca. 16.000 Begriffen, 15.000 Hierarchiebeziehungen und 26.000 Assoziationsbeziehungen) sehr aufwendig ist, soll die Analyse durch Analysemodelle und -werkzeuge unterstützt werden.

2.3.2.2 Auffinden und Klassifizieren von Inter-Thesaurus-Relationen

Zwischen den Begriffen in den verschiedenen Thesauri sollen unter Berücksichtigung des gesamten Begriffsnetzes sowie der Ergebnisse der in Abschnitt 2.3.2.1 beschriebenen Analyse Beziehungen gefunden werden (vgl. Abschnitt 2.3.1.2).

Angenommen wird, dass die Thesauri mindestens Überschneidungen in den von ihnen abgedeckten Fachgebieten besitzen[3]. Als Konsequenz der für jeweils ein bestimmtes Informationssystem maßgeschneiderten Thesauri muss aber auch bei (Teil-) Thesauri aus demselben Fachgebiet mit wesentlichen Unterschieden gerechnet werden, die eine Reihe von Auswirkungen auf die Begriffsintegration haben, die es angemessen zu berücksichtigen gilt. Falls die Integrationsverfahren es

[3] Ansonsten würde eine Integration ausschließlich über gemeinsame Dach-Begriffe stattfinden können. Dieser Fall wird als für die Praxis der Integration von Thesauri für das Information Retrieval als wenig relevant eingeschätzt und daher im Rahmen dieser Arbeit nicht weiter betrachtet.

erfordern, sind in einer Vorbereitungsphase Anpassungen und Anreicherungen erforderlich. Die Integrationsverfahren müssen so konfiguriert werden können, dass sie trotz all dieser Unterschiede möglichst vollständige und korrekte Ergebnisse liefern. Insbesondere gilt es, das in den Thesauri selbst gespeicherte Wissen möglichst vollständig zur Integration zu nutzen.

Wir differenzieren Unterschiede in der Begriffsbildung, der Begriffsdarstellung, dem Thesaurusmodell und der Verwendung des Thesaurus zum Indexieren und zum Retrieval.

Bei der Begriffsintegration sind diese Unterschiede entsprechend zu berücksichtigen.

2.3.2.2.1 Berücksichtigung unterschiedlicher Begriffsbildungen Unter *Begriffsbildung* (conceptualisation) wird in diesem Zusammenhang die Auswahl, Differenzierung und Einordnung von Begriffen eines Thesaurus verstanden. Das Ergebnis der Begriffsbildung ist somit die Menge von Begriffen eines Thesaurus sowie die Menge von Relationen über diesen Begriffen.

Unterschiedliche Begriffsauswahl: Abhängig von dem zu indexierenden Informationsbestand wird die Menge der Begriffe eines Thesaurus gewählt. Teilbereiche können in einem Thesaurus daher gar nicht oder überproportional stark vertreten sein. Die Integrationsdichte wird somit ebenfalls sehr unterschiedlich sein. Als Beispiel sei der Bereich *environmental impact* (Umweltbelastung bzw. Umwelteinfluss) genannt, der in AGROVOC nur durch eine geringe Anzahl von wenig spezifischen Begriffen repräsentiert ist (*environmental impacts*, *environmental impact assessment* und *resource depletion* mit den Unterbegriffen *overfishing* und *water depletion*), während GEMET eine Reihe spezifischer Begriffe, z.B. die Auswirkungen verschiedener Aktivitäten wie *environmental impact of tourism* oder Einflussquellen wie *accidental release of organisms*, *introduction of plant species* und *poaching* (Wilderei), enthält.

Unterschiedliche Begriffsaufteilungen: Selbst gleiche Begriffe sind in den verschiedenen Komponententhesauri in der Regel unterschiedlich in Unterbegriffe aufgeteilt (vgl. auch [VJBCS98]). So können verschiedene Typen von Beziehungen zwischen den jeweiligen Unterbegriffen existieren.Der Begriff *water (substance)* in GEMET entspricht etwa *water* in AGROVOC. In GEMET ist dieser Begriff dann aufgeteilt in u.a. folgende Abstraktionsunterbegriffe *consumptive water*, *drinking water*, *mineral water* und *water for agricultural use*. In AGROVOC sind u.a. die Abstraktionsunterbegriffe *distilled water*, *drinking water* und *irrigation water* enthalten. Während es für die Begriffe *mineral water* und *distilled water* keine Beziehungen zwischen den jeweiligen Unterbegriffen gibt, ist *drinking water* äquivalent und zwischen *water for agricultural use* und *irrigation water* existiert eine Abstraktionsbeziehung.

Unterschiedliche Abstraktionsebenen: Durch eine unterschiedliche Ausdifferenzierung von Begriffen können die Thesauri Begriffe von unterschiedlichen Abstraktionsebenen beinhalten (vgl. auch [VJBCS98]). Evtl. können Abstraktionsbeziehungen zwischen den Begriffen bestehen. Im obigen Beispiel enthält GEMET ausschließlich den allgemeineren Begriff *water for agricultural use*, während AGROVOC ausschließlich den spezielleren Begriff *irrigation water* enthält. Die Begriffe können mit einer Abstraktionsbeziehung verbunden werden.

Unterschiedliche Granularität (Teil-Äquivalenz): Die Intensionen (Inhalte) der Begriffe entsprechen sich zwar oft teilweise, nicht aber vollständig. So enthält der Begriff *reptiles* in AGROVOC den Unterbegriff *testudinata*, der wiederum die Unterbegriffe *turtles*, *freshwater tortoises* und *terrestrial tortoises* besitzt. In GEMET hingegen enthält der Begriff

reptiles in dieser Hinsicht nur den Abtraktionsunterbegriff *tortoises*. Dieser entspricht somit zum Teil dem Begriff *testudinata*, der aber zusätzlich *turtles* umfasst. Außerdem gibt eine Teil-Äquivalenz zwischen *tortoises* und *freshwater tortoises* sowie *terrestrial tortoises*. Als umgangssprachliches Beispiel seien *Jugendlicher* und *Teenager* genannt.

Unterschiedliche Begriffseinordnungen (Serialisierungen): Identische Begriffe werden in den Hierarchien unterschiedlich eingeordnet. Z.B. wird *mineral water* in GEMET unter *water (substance)* eingeordnet, während derselbe Begriff in AGROVOC unter *beverages* eingeordnet und damit unter einem ganz anderen Gesichtspunkt betrachtet wird. Der Extremfall einer direkt gegensätzlichen Einordnung ist ebenfalls möglich. Dies ist in GEMET und AGROVOC z.B. für die Begriffe *economic development* und *economic growth* der Fall.

2.3.2.2.2 Berücksichtigung unterschiedlicher Begriffsdarstellungen Die während der Begriffsbildung ausgewählten Begriffe sind abstrakt und müssen durch konkrete Daten repräsentiert und definiert werden. Auch bei dieser Begriffsdarstellung unterscheiden sich die Thesauri.

Benennungen: Auf der Ebene der sprachlichen Mittel zur Darstellung der Begriffe in Form von ein- oder mehrwortigen Benennungen treten in den unterschiedlichen Thesauri Synonyme, Homonyme, Schreibvarianten und die unterschiedliche Verwendung von Singular- bzw. Pluralformen auf:

Synonyme: In den verschiedenen Thesauri werden gleiche Begriffe durch verschiedene Benennungen repräsentiert. Als Beispiel seien *high water* und *flood*, *fodder plant* und *feed crop* jeweils in GEMET und AGROVOC genannt. Häufig ist dies der Fall, wenn ein Fremdwort in dem einen Thesaurus, hingegen ein natives Wort in dem anderen Thesaurus verwendet wird. Beispiele sind *bovids* und *bovidae* sowie *dendometry* und *forest mensuration* oder *riverside vegetation* und *riparrian vegetation* jeweils aus GEMET und AGROVOC.

Homonyme/Polyseme: In den verschiedenen Thesauri werden gleiche Benennungen verwandt, um verschiedene Begriffe darzustellen. Das Beispiel *lime* wurde bereits in Abschnitt 2.3.1.2.2 genannt. Ein weiteres Beispiel ist *wood*, das in AGROVOC einen Begriff im Sinne von Holz darstellt und in GEMET im Sinne von Bewaldung.

Schreibvarianten: Benennungen weichen aufgrund von Varianten in der Schreibweise voneinander ab. Beispiele sind *by-product* und *byproduct*, *germ plasm* und *germplasm*, *competition (biological)* und *biological competition* (die erste Schreibvariante stammt jeweils aus GEMET, die zweite aus AGROVOC).

Singular-/Pluralformen: Benennungen können in den Thesauri im Singular oder im Plural enthalten sein. In der englischen Sprache wird beispielsweise für Benennungen mit einer gebräuchlichen Pluralform in der Regel diese im Thesaurus aufgeführt. Dies ist für AGROVOC der Fall, in GEMET aber nicht. So enthält AGROVOC allein die Benennung *men* während GEMET ausschließlich *man* enthält.

Groß-/Kleinschreibung: Die Groß-/Kleinschreibung der Benennungen ist in der Regel die gleiche wie bei der Verwendung in natürlichsprachigen Texten. Im Englischen gilt generell, dass die Benennungen klein geschrieben werden, Ausnahmen sind Eigennamen und Abkürzungen. Die GEMET-Benennungen gehorchen diesen Regeln, in AGROVOC jedoch wird jede Benennung komplett großgeschrieben (*PEST CONTROL*). Ebenso werden die Topterme der GCMD Parameter Valids großgeschrieben,

Benennungen auf anderen Ebenen mit großen Anfangsbuchstaben (*Feeding Habitat*), es sei denn es werden Begriffe zusammengefasst *Dust/ash*[4].

Definitionen: Für eine vollständige Begriffsdefinition ist auch die Angabe mehrerer gleichbedeutender Benennungen (Synonyme) häufig nicht ausreichend. Daher werden die Begriffe innerhalb der Thesauri zusätzlich durch textuelle Definitionen des Begriffsinhaltes, Erläuterungen (scope notes) zum Gebrauch und der Bedeutung sowie durch die Relationen mit anderen Begriffen definiert. Innerhalb der verschiedenen Thesauri werden die Begriffe sehr unterschiedlich definiert.

Unterschiedlicher Detaillierungsgrad: Innerhalb der GCMD Parameter Valids werden Begriffe stets durch *eine einzige* Benennung definiert, die in eine polyhierarchische Struktur von Begriffen eingeordnet ist. Synonyme oder weitere textuelle Definitionen gibt es nicht. Dass mit *stress* keine psychische Anspannung gemeint ist, kann nur anhand der Einordnung unter *tectonics* erkannt werden. Zur Erkennung der tatsächlichen Bedeutung ist zumindest geologisches Grundwissen erforderlich. Nur wenn verschiedene Begriffe innerhalb der GCMD Parameter Valids zu einem Begriff zusammengefasst werden, werden die entsprechenden Benennungen als Komposita bzw. Mehrwortbenennungen aufgeführt (z.B. *migratory rates/routes*).

AGROVOC hingegen gibt in der Regel zusätzlich ein oder mehrere Synonyme an (für *stress* etwa *pysical stress factor*, *physiological stress resistance*, *distress*, *stray voltage effects*, *abiotic stress* und *biotic stress*), anhand derer die Bedeutung innerhalb des Thesaurus näher ersichtlich wird. AGROVOC erlaubt ebenfalls eine polyhierarchische Einordnung und gibt zusätzlich in der Regel mehrere verwandte Begriffe an (für *stress* etwa *dysregulation*, *starvation*, *stimuli*, *water deprivation* und *shock*), die eine weitere Einordnung ermöglichen.

GEMET erlaubt prinzipiell die gleichen Möglichkeiten wie ARGOVOC, macht aber von Synonymen und verwandten Begriffen deutlich seltener Gebrauch. Dafür werden für einen Großteil der Begriffe Definitionen in natürlicher Sprache gegeben (für *stress* lautet diese Definition *A stimulus or succession of stimulu of such magnitude as to tend to disrupt the homeostasis of the organism. (Source: McGraw-Hill)*).

Von eigenständigen Erläuterungen zur Bedeutung und zur Verwendung machen weder AGROVOC noch GEMET noch GCMD Parameter Valids Gebrauch. Stattdessen werden Polyseme und Homonyme in AGROVOC und GEMET durch Zusätze in den Benennungen aufgelöst (Beispiele: *tanker (truck)* und *tanker (ship)* in GEMET und *cancer (genus)* und *cancer (disease)* in AGROVOC). Diese Zusätze können aber auch weitere semantische Informationen enthalten, z.B. dass es sich um ein Symbol handelt (*fe (symbol)* in AGROVOC), einen Hinweis auf das Herkunftsland der Benennung (*H-FCKW (D)* in GEMET) oder die Wortart (*public (n.)* in GEMET).

In welcher Intensivität und auf welche Art und Weise von den Möglichkeiten einer Definition über die verschiedenen Beziehungstypen Gebrauch gemacht wird, ist in den einzelnen Thesauri ebenfalls sehr unterschiedlich. AGROVOC verwendet in der Regel ausschließlich Abstraktionsunterbegriffe (z.B. *arid climate* und *humid climate* als Unterbegriffe von *climate*) oder ausschließlich Bestandsunterbegriffe (z.B. *atmosphere* als Bestandsunterbegriff von *earth*). In GEMET hingegen werden häufig auch gemischt Abstraktions- und Bestandsunterbegriffe aufgeführt (z.B. *microclimate* und *climatic change* als Unterbegriffe von *climate*).

[4]Im Rahmen dieser Arbeit werden aus Gründen der Anschaulichkeit Benennungen aus allen Thesauri in der üblichen Form der Groß-/Kleinschreibung dargestellt.

Mehrsprachige Äquivalenzen: In einem multilingualen Thesaurus können weitere Hinweise auf die Bedeutung eines Begriffs durch die Betrachtung von äquivalenten Benennungen in verschiedenen Sprachen gewonnen werden. Dies gilt in der Fallstudie für GEMET und AGROVOC, die jeweils in mehreren Sprachen (bei gleicher Struktur innerhalb der verschiedenen Sprachen) vorliegen.

Unterschiedlicher Formalisierungsgrad: In Ontologien wird versucht, die Definitionen möglichst formal zu halten (z.B. basierend auf deskriptiven Sprachen). Dies ist aufgrund des hohen Aufwands für vollständige und konsistente formale Definitionen für Thesauri nicht der Fall. Dennoch können die Definitionen gewissen Regeln gehorchen (vgl. z.B. [DIN93b]). In GEMET – in der Fallstudie der einzige Thesaurus mit Definitionen – sind diese Definitionen weder für alle Begriffe vorhanden noch – wo vorhanden – homogen. Die Spannbreite wird an zwei Beispielen ersichtlich: Als Definition für *antrophic activitity* ist *Human activity (Source: Random House)* angegeben. *climate* hingegen wird äußerst ausführlich durch *The long-term prevalent weather conditions of an area. / The average weather condition in a region of the world. Many aspects of the Earth's geography affect the climate. Equatorial, or low, latitudes are hotter than the polar latitudes because of the angle at which the rays of sunlight arrive at the Earth's surface. The difference in temperature at the equator and at the poles has an influence on the global circulation of huge masses of air. Cool air at the poles sinks and spreads along the surface of the Earth towards the equator. Cool air forces its way under the lower density warmer air in the lower regions, pushing the lighter air up and toward the poles, where it will cool and descend (Source: Collins / Wright)* definiert.

Fehlende Begründungen: Der Grund für die Etablierung einer bestimmten Beziehung ist anhand der Thesauri allein nicht erkennbar. Das gilt auch für AGROVOC, GEMET und die GCMD Parameter Valids.

2.3.2.2.3 Berücksichtigung unterschiedlicher Thesaurusmodelle Obwohl DIN- und ISO-Normen [DIN87, DIN93a, ISO90] das Modell eines Thesaurus sehr genau definieren, weichen die in der Praxis verwendeten Thesaurusmodelle an verschiedenen Stellen voneinander ab. Alleine bei Betrachtung der Thesauri AGROVOC, GEMET und GCMD Parameter Valids können folgende Unterschiede festgestellt werden:

Differenzierung der Hierarchierelationen: Eine Differenzierung der Hierarchierelation in Abstraktionsrelation, Bestandsrelation und Instanzrelation, wie sie DIN 1463 Teil 1 [DIN87] optional vorsieht, findet in allen drei Thesauri nicht statt. Die mögliche Schlussfolgerung, dass jeweils nur ein Typ von Hierarchiebeziehung verwendet wird, kann durch Gegenbeispiele widerlegt werden (z.B. die Bestandsrelation zwischen *earth* und *atmosphere* und die Abstraktionsrelation zwischen *climate* und *microclimate* in AGROVOC).

Angaben von Definitionen und Erläuterungen: Obwohl von der DIN 1463 Teil 1 [DIN87] audrücklich zumindest in solchen Fällen gefordert, in denen „Zweifel an den einheitlichen Interpretationen eines Deskriptors bestehen“, erlaubt nur GEMET die Angabe von Definitionen. Erläuterungen sind prinzipiell in GEMET und AGROVOC erlaubt, wobei sie in AGROVOC nicht selten die Form von Definitionen annehmen (z.B. *Avoid confusion with Benin in Nigeria* zu *benin* (früherer Teil Französisch Westafrikas) als reine Erläuterung, hingegen *Cultivation of trees and shrubs, individually or in small groups, for ornament or instruction rather than use or profit* als Erläuterung mit definierendem Charakter für *arboriculture* (Baumkultur)).

Übersetzungen: In den multilingualen Thesauri AGROVOC und GEMET sind Deskriptoren für alle Thesaurussprachen angegeben, in den monoligualen GCMD Parameter Valids einsprachig auf Englisch.

Spezifische Erweiterungen: Zum Teil gehen Thesaurusmodelle über die Definitionen in den Normen hinaus. Häufig besitzen sie z.B. zusätzliche Ordnungssysteme für die Begriffe. In der Fallstudie ist das der Fall für GEMET, der eine disziplin-orientierte Ordnung durch so genannte Gruppen und Super-Gruppen sowie eine weitere thematische Ordnung durch so genannte (Umwelt-) Themen etabliert (vgl. Abschnitt 2.1).

Weitere Unterschiede sind möglich, da in den Normen keine definitiven Vorgaben gemacht werden (z.B. monohierarchischer Thesaurus vs. polyhierarchischer Thesaurus, Thesaurus mit bzw. ohne Vorzugsbezeichner, Flexionsformen als Synonyme) und auch keinerlei Invarianten explizit definiert werden (z.B. werden Zyklen in der Hierarchierelation von den Normen nur implizit ausgeschlossen). Inhaltlich können Thesauri eher Corpus-zentriert (angepasst an den Dokumentenbestand) oder eher benutzerzentriert (angepasst an eine bestimmte Benutzergruppe) sein. Sie können den Charakter eines möglichst vollständigen Fachwörterbuches oder einer Ad-hoc-Sammlung von Begriffen haben.

2.3.2.2.4 Berücksichtigung unterschiedlicher Verwendung bei Indexierung und Retrieval Welche Begriffe wie zum Indexieren und Retrieval verwendet werden, hängt von verschiedenen Faktoren wie Indexierungsrichtlinien (s. z.B. [Nat98]) oder der sozio-kulturellen Umgebung ab. Idealerweise werden auch diese potenziell sehr unterschiedlichen Vorgehensweisen in die Begriffsintegration einfließen. Erforderlich dazu ist eine Analyse der indexierten Dokumente, anhand derer Rückschlüsse auf das Indexierungsverhalten gezogen werden können.

2.3.2.3 Einführen von Ergänzenden Begriffen

Wie bereits in Abschnitt 2.3.1.3.2 dargestellt, kann es sinnvoll sein, Begriffe einzuführen, die in keinem Komponententhesaurus enthalten sind, so genannte Ergänzende Begriffe. Die daraus resultierende Anforderung an die Begriffsintegration lautet, entsprechende Stellen für die Einführung solcher Ergänzender Begriffe aufzudecken und idealerweise auch Benennungen für diese Begriffe vorzuschlagen. Dazu ist nach 2.3.1.3.2 das Erkennen von Schwesterbeziehungen und von Abstraktionsniveauunterschieden zwischen Schwesterknoten erforderlich sowie das Auffinden entsprechender Begriffe.

2.3.2.4 Konflikterkennung und -behandlung

Wenn die Autonomie der Komponententhesauri respektiert werden soll, dürfen bei der Begriffsintegration keinerlei Eingriffe in die Strukturen der Komponententhesauri stattfinden. Das Etablieren von Beziehungen zwischen Begriffen aus verschiedenen Thesauri kann dann aber dazu führen, dass gegen die Invarianten des Multi-Thesaurus-Systems verstoßen wird (vgl. Abschnitte 2.3.1.4 und 2.3.1.5). Damit der Benutzer dennoch einen kohärenten Gesamteindruck erhält, sollen solche Konflikte erkannt und markiert werden. Dies erfordert

- eine Klassifikation möglicher Konflikte,
- Mechanismen für das Auffinden solcher Konflikte (Konfliktanalyse),

- das Feststellen der Ursachen eines Konfliktes und schließlich

- die Kennzeichnung des Konfliktes mit seinen Ursachen.

Eine Markierung von Konflikten ist Voraussetzung für eine angemessene Behandlung. Die Konfliktbehandlung soll entscheiden, wie mit einem entsprechend markierten Konflikt in einer bestimmten Situation umgegangen wird.

Während die Konflikterkennung und -markierung integraler Bestandteil der Begriffsintegration ist, muss abhängig von der Konfliktbehandlungsstrategie entschieden werden, ob diese als Bestandteil der Begriffsintegration oder aber der Anfragebearbeitung (vgl. Abschnitt 2.3.4) angesehen wird.

2.3.2.5 Vorgehensmodell zur Integration

Die vorangegangenen Abschnitte verdeutlichen bereits die Komplexität einer Begriffsintegration von Begriffsnetzen mit mehreren tausend Begriffen und einem Mehrfachen an Bezeichnern und Beziehungen. Zusätzlich gilt es zu berücksichtigen, dass es alleine im Umweltbereich und nahe verwandten Bereichen eine Vielzahl weiterer Thesauri gibt (z.B. der allgemeine Umweltthesaurus des Umweltbundesamtes [Bat94], die sehr umfangreichen Thesauri CAB Thesaurus [CAB99] und NASA Thesaurus [NAS98], der sehr spezielle Bioethics Thesaurus [Ken99]) und in einer Liste der Generaldirektion 13 (DG XIII) der Europäischen Union 1.000 weltweit häufig verwendete Thesauri aus den verschiedenen Bereichen aufgeführt werden [Rad90]. Auch wenn eine Integration all dieser Thesauri nicht erstrebenswert ist, kann doch davon ausgegangen werden, dass es wie in der in Abschnitt 2.2 dargestellten Fallstudie in der Zukunft Thesaurusintegrationen wiederholt und in den unterschiedlichsten Anwendungsbereichen geben wird. Aus diesen Gründen soll ein Vorgehensmodell zur Begriffsintegration entwickelt werden. Ein solches Vorgehensmodell reduziert den Aufwand einer Begriffsintegration, indem eine allgemeine Methodik angegeben wird, die auf die Aufgabe „Begriffsintegration“ angewandt werden kann. Dazu muss das Vorgehensmodell spezifisch genug sein, um den Begriffsintegratoren ausreichend Anleitung und Hilfe zu geben, und flexibel genug, um unter den verschiedenen Randbedingungen angewendet werden zu können.

2.3.2.6 Referenzarchitektur für die Begriffsintegration

Der Prozess der Begriffsintegration ist in einer Systemarchitektur, die wir auch als *Wissensakquisitionsarchitektur* bezeichnen, da sie die Gewinnung des Integrationswissens unterstützt, entsprechend zu berücksichtigen. Diese Wissensakquisitionsarchitektur soll Referenzcharakter für die Integration von Thesauri und Ontologien besitzen.

Das benötigte Integrationswissen gilt es aufgrund der Komplexität der Komponententhesauri und des erforderlichen gründlichen Verständnisses der Semantik *semi-automatisch*, d.h. mit der Unterstützung menschlicher Experten, zu erwerben. Dieser Wissensakquisitionsvorgang ist aufgrund der erforderlichen Interaktionen mit Experten *vor der Bereitstellung* eines integrierten Zugangs durchzuführen und soll die Einbindung unterschiedlicher Verfahren, Wissensquellen und Qualitätssicherungsmechanismen unterstützen. All diese Aspekte des Wissensakquisitionsvorganges sollen von der Wissensakquisitionsarchitektur angemessen unterstützt werden.

2.3.2.7 Benutzeragent

Ein Multi-Thesaurus-System bestehend aus AGROVOC, GEMET und GCMD Parameter Valids ist bereits ein komplexes Wissensrepäsentationssystem mit weit über 10.000 Begriffen und einem mehrfachen an Beziehungen. Es ist zu erwarten, dass mit der zunehmenden Informationsvernetzung [LKK+97, Der99, Sch99a] und somit größeren integrierten Informationssystemen auch die Komplexität der Multi-Thesaurus-Systeme weiter zunehmen wird. Um den menschlichen Integrationsexperten bei der Verwendung dieser komplexen Systeme zu unterstützen, sollen „intelligente“ Darstellungsformen für Multi-Thesaurus-Systeme angeboten werden, die u.a. folgenden Anforderungen genügen:

- Eine Überfrachtung des menschlichen Experten durch die Komplexität des Multi-Thesaurus-Systems soll vermieden werden. Dazu sollen eine Fokussierung auf relevante Ausschnitte, unterschiedliche Ansichten und eine leichte Handhabbarkeit beitragen.
- Die Orientierung im Begriffsnetz soll durch ausreichende Interaktionsmöglichkeiten unterstützt werden.
- Die vollständigen Informationen des Multi-Thesaurus-Systems müssen dargestellt werden können. Dazu gehören außer den Begriffen (vgl. Abschnitt 2.3.1.3) mit ihren Definitionen und Relationen unterschiedlichen Typs (vgl. Abschnitt 2.3.1.2) auch die Zugehörigkeit der Begriffe zu den Komponententhesauri sowie Konflikte (vgl. Abschnitt 2.3.1.5).

Wir betrachten im Rahmen dieser Arbeit ausschließlich Benutzeragenten zur Unterstützung des menschlichen Experten während des Integrationsprozesses. Je nach Anwendungskontext werden für eine Thesaurusföderation verschiedene Benutzeragenten für den Endbenutzer existieren, die auf definierten Schnittstellen aufsetzen können.

2.3.3 Bewertung der Güte eines Multi-Thesaurus-Systems

Nach dem Erstellen eines Multi-Thesaurus-Systems sind möglichst objektive Aussagen über die Güte dieses Systems erforderlich. Anhand dieser Aussagen soll festgestellt werden können,

- ob die *Integrationserwartungen* durch das vorliegende Ergebnis erfüllt werden (vgl. Abschnitt 2.3.2.1) und
- inwiefern die Integration *vollständig* und *korrekt* ist.

Diese Feststellungen werden zum einen dazu benötigt, um zu entscheiden, ob weitere Integrationsschritte erforderlich sind. Zum anderen sind sie erforderlich, um die *Effizienz* und *Effektivität* des *Multi-Thesaurus-Systems* sowie der *Integrationsverfahren* – auch vergleichend mit anderen Systemen und Verfahren – beurteilen zu können. Sie sind somit unentbehrliche Grundlage, um Verbesserungen an vorhandenen Systemen und Verfahren vornehmen und die Auswirkungen solcher Modifikationen einschätzen zu können.

2.3.4 Ausführungsmaschine

Um im laufenden Betrieb eine integrierte Sicht auf das Gesamtvokabular anbieten zu können, ist innerhalb der Architektur eines föderierten Informationssystems eine *Ausführungsmaschine* zum integrierten Zugang zu den Komponententhesauri erforderlich. Die Ausführungsmaschine muss

Anfragen entgegennehmen, bearbeiten, ggf. optimieren und schließlich die Ergebnisse zurückliefern. Anfragen können z.B. lauten, alle Benennungen, die die Zeichenkette „water monitoring“ enthalten, zu suchen, zu einem gegebenen föderierten Begriff alle Abstraktionsunterbegriffe zu ermitteln oder eine Menge von Begriffen aus Nichtindexierungsthesauri auf Begriffe von Indexierungsthesauri abzubilden. Daher ist sowohl Zugang zu den Komponententhesauri als auch Integrationswissen erforderlich. Es gilt zu berücksichtigen, dass die *Autonomie* der Komponententhesauri erhalten bleibt und sowohl die *Entfernung* (z.B. werden CDS mit GEMET auf einem Server des Niedersächsischen Umweltministeriums in Hannover, Deutschland und INFEO mit GCMD Parameter Valids auf einem Server des Joint Research Centers (JRC) in Ispra, Italien betrieben) als auch die *Heterogenität* der Thesauri (unterschiedliche Datenmodelle, Datenschema und Zugriffsschnittstellen) überwunden werden müssen.

An die eigentliche Anfragebearbeitung werden folgende Anforderungen gestellt:

- Die globale Anfrage muss in Teilanfragen zerlegt werden, die an die Anfragebearbeitung der Komponentensysteme (in der Fallstudie die Thesauri AGROVOC, GEMET und GCMD Parameter Valids sowie eine Begriffsintegrationskomponente) weitergeleitet werden.
- Aus den lokalen Antworten muss ggf. mit Hilfe von Post-Processing-Anfragen die globale Antwort zusammengesetzt werden. U.U. sind markierte Konflikte angemessen zu behandeln.
- Die Anfragebearbeitung des Multi-Thesaurus-Systems ist in die Anfragebearbeitung des übergeordneten Informationssystems einzubetten.

Um die Integration einer größeren Anzahl von Komponententhesauri zu ermöglichen sowie bei gleichzeitigem Zugriff durch viele Benutzer noch das gewünschte Verhalten zu zeigen, wird ausreichende *Skalierbarkeit* gefordert. Wesentliche Anforderung an eine Anfrageoptimierung ist es daher, für eine gegebene Anfrage einen möglichst effizienten Ausführungsplan zu finden, so dass die Ergebnisse in möglichst kurzer Zeit berechnet werden können.

Weitere Optimierungsziele (z.B. minimale Kosten) sind möglich, sollen im Rahmen dieser Arbeit aber nicht betrachtet werden.

2.4 Fokus der Arbeit

Zusammenfassend kann festgehalten werden, dass die zentrale Herausforderung die Verbesserung der *Skalierbarkeit* ist. Dies betrifft sowohl die mögliche Integration von mehreren Komponententhesauri als auch den eigentlichen Integrationsprozess von komplexen Wissensrepräsentationssystemen. Der Schwerpunkt unserer Forschung soll somit die Entwicklung von Methoden und Werkzeugen

- zur Erleichterung des Prozesses der Thesaurusintegration und
- zur Unterstützung des Benutzers beim Umgang mit dem integrierten System.selbst

sein. Eine weitere bedeutende Herausforderung ist die zu erreichende *Flexibilität*. Dies gilt sowohl hinsichtlich der sehr heterogenen zu integrierenden Thesauri als auch hinsichtlich einer möglichen Auswahl einer Teilmenge der integrierten Thesauri für die Anfragebearbeitung.

Kapitel 3

Stand der Forschung

In den einführenden Kapiteln wurden die Anforderungen an Modelle, Erstellung und Handhabung lose integrierter Thesauri formuliert. Als zentrale Anforderungen haben sich dabei Skalierbarkeit und Flexibiliät herauskristallisiert. Wir haben daher als Fokus der Arbeit die Verbesserung der Skalierbarkeit und Flexibilität von Multi-Thesaurus-Systemen identifiziert. In diesem Kapitel soll durch eine Betrachtung des aktuellen Stands der Forschung zum einen die Relevanz der Fragestellung nochmals nachgewiesen werden, zum anderen werden Forschungsarbeiten vorgestellt, die wichtige Grundlagen für unsere Arbeit bilden.

3.1 Informationsmodelle für integrierte Thesauri

Als grundlegende Anforderung bei der Integration von Thesauri wurde in Abschnitt 2.3.1 ein adäquates Informationsmodell identifiziert. In der Literatur werden bereits eine Reihe von Arbeiten vorgestellt, die das Verbinden von Komponententhesauri zum Ziel haben. Die dort verwendeten verschiedenen Informationsmodelle für Multi-Thesaurus-Systeme werden von uns klassifiziert. Des Weiteren werden Ontologie-Integrationsprojekte betrachtet, die aus Sicht der verwendeten Modelle für Multi-Ontologie-Systeme interessant erscheinen. Sowohl die Modelle für Multi-Thesaurus-Systeme als auch für Multi-Ontologie-Systeme werden hinsichtlich der Anforderungen aus Abschnitt 2.3.1 bewertet.

3.1.1 Klassifikation von Modellen für Multi-Thesaurus-Systeme

Bei zunehmendem Grad der Integration werden von uns die aus der Literatur bekannten Modelle für Multi-Thesaurus-Systeme in Multi-Thesaurus-Umgebungen, Thesaurus-Wechsel-Systeme und Thesaurusverbünde klassifiziert [NTK98].

3.1.1.1 Multi-Thesaurus-Umgebungen

Eine Multi-Thesaurus-Umgebung enthält mehrere Thesauri, deren Benennungen *nicht* semantisch in Beziehungen gesetzt sind (vgl. Abb. 3.1). Einige dieser Systeme haben einen gemeinsamen Index für alle Thesauri oder Suchfunktionen über alle Thesauri. Beispiele sind das in [NM85] vorgestellte so genannte „Vokabular-Wechsel-System“ und die Arbeit von Stern und Rischette über die Erzeugung eines „Super-Thesaurus“ aus mehreren Agrar-Thesauri [SR90].

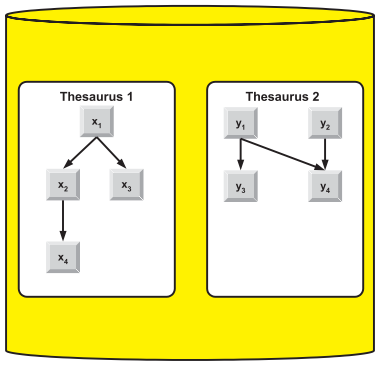

Abbildung 3.1: Multi-Thesaurus-Umgebung

3.1.1.2 Thesaurus-Wechsel-Systeme

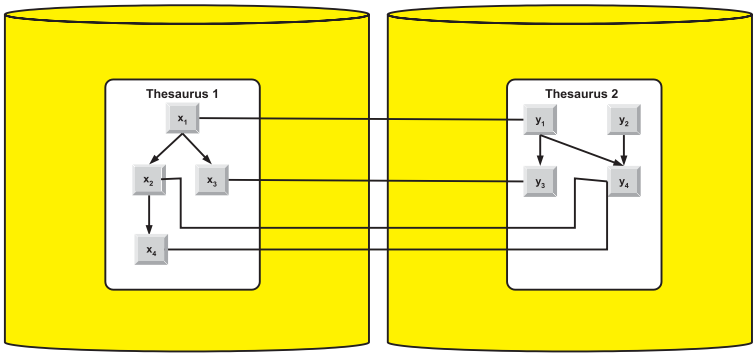

Abbildung 3.2: Thesaurus-Wechsel-System (in dieser und der folgenden Abbildung repräsentieren Verbindungen mit Pfeilspitzen Hierarchiebeziehungen, Verbindungen ohne Pfeilspitzen Äquivalenzbeziehungen)

Thesaurus-Wechsel-Systeme erlauben es, Begriffe von einem Thesaurus auf Begriffe eines anderen Thesaurus abzubilden (vgl. Abb. 3.2). Die Abbildung kann eine partielle Abbildung sein, die nur exakte Abbildungen zwischen den Begriffen der verschiedenen Thesauri enthält, oder eine vollständige Abbildung, die jeden Begriff auf den nächstbesten Begriff eines anderen Thesaurus abbildet. Sylvester und Klingbiel präsentieren ein solches System zum „Thema-Wechseln“ (engl. subject switching), das zur Re-Indexierung von zuvor mit dem NASA-Thesaurus indexierten Dokumenten mit dem DTIC-Thesaurus verwendet wird.

Sollen in einem Thesaurus-Wechsel-System mehr als zwei Thesauri verbunden werden, kann statt einer Abbildung zwischen jedem Paar von Thesauri eine indirekte Abbildung über eine Interlingua stattfinden (vgl. Abb. 3.3 sowie Anhang D.2, S. 272). In [Nev70] wird eine Methode vorgeschlagen, wie die Interlingua und die erforderlichen Abbildungen aus den Komponententhesauri gebildet werden können.

Eine Interlingua kann, wenn sie nicht nur die abzubildenden Konzepte, sondern auch die Beziehungen zwischen diesen enthält, wiederum selbst ein Thesaurus sein, der dann *Meta-Thesaurus* genannt wird. Bekannte Beispiele für solche Meta-Thesauri sind „Meta“ des Unified Medical Language Systems [Squ93] und der European Educational Thesaurus [Rou92]. Eine Möglichkeit, einen Meta-Thesaurus aus den Komponententhesauri zu erhalten, besteht darin, die Komponententhesauri zu vereinigen (s. Thesaurusvereinigung in Abschnitt 3.1.1.3).

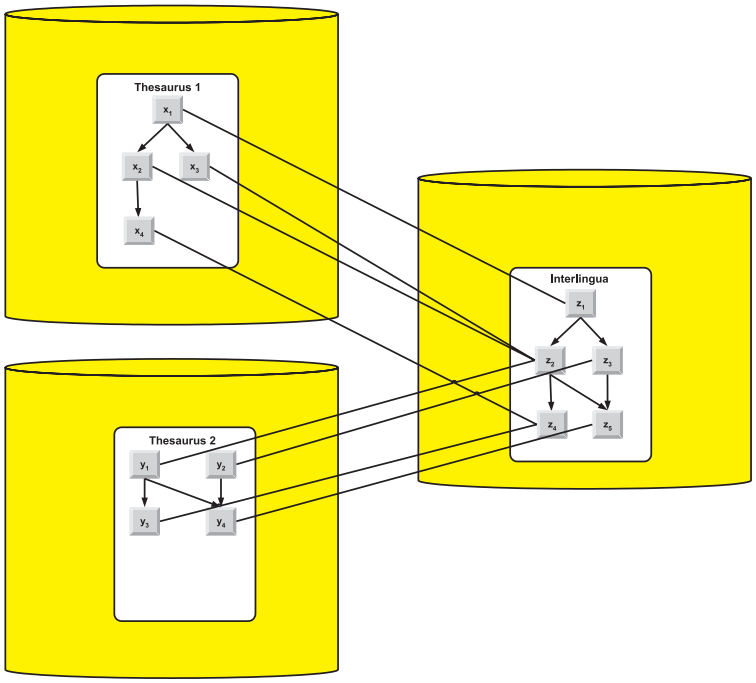

Abbildung 3.3: Thesaurus-Wechsel-System mit einer Interlingua

3.1.1.3 Thesaurusverbünde

Thesaurusverbünde vereinigen mehrere Thesauri zu einem neuen Thesaurus oder zumindest zu einem Thesaurus-ähnlichen Gebilde. Die aus der Literatur bekannten Ansätze können weiter klassifiziert werden in Thesauruskopplungen und Thesaurusvereinigungen.

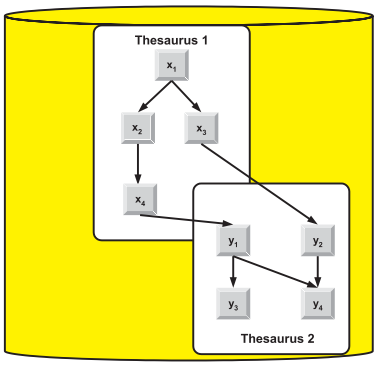

Abbildung 3.4: Thesauruskopplung

Thesauruskopplung (Micro-/Macrothesaurus): Ein Spezialfall unter den Thesaurusverbünden ist das Koppeln von Spezialthesauri (Thesauri mit fachlich sehr spezialisiertem Vokabular, auch Subthesauri oder Mikrothesauri genannt) an einen allgemeineren Dachthesaurus (oder Makrothesaurus). Die Kopplung geschieht derart, dass die Topterme eines Spezialthesaurus zugleich die spezialisiertesten Begriffe in der Hierarchie des allgemeinen Thesaurus sind (vgl. Abb. 3.4). Fragen der Inkonsistenz treten in diesem speziellen Fall nicht auf. Auch sind die Einzelthesauri unverändert im Zusammenschluss enthalten. Als

Beispiel sei Roulin genannt, der das Ankoppeln von Mikrothesauri an den European Educational Thesaurus beschreibt [Rou92].

Thesaurusvereinigung: Thesaurusvereinigungen bilden aus mehreren Komponententhesauri einen neuen, in sich konsistenten Thesaurus und realisieren somit den größtmöglichen Grad der Integration (vgl. Abb. 3.5). Dabei vereinigen sie – wie Thesaurusföderationen – die Begriffe und Beziehungen der Komponententhesauri in sich und fügen weitere Beziehungen hinzu. Allerdings verzichten sie auf einen Beibehalt der Autonomie der Komponententhesauri, die innerhalb der Thesaurusvereinigung soweit modifiziert werden, dass die neue Thesaurusvereinigung keine Inkonsistenzen und Konflikte enthält. Erforderlichenfalls werden Begriffe und Beziehungen entfernt. Das Modell einer Thesaurusvereinigung ist identisch mit dem der integrierten Komponententhesauri.

Thesaurusvereinigungen können Ausgangspunkt für einen neuen Thesaurus sein, der sich unabhängig von den Komponententhesauri weiterentwickelt. GEMET, der General European Multilingual Environment Thesaurus [CNR97], ist ein Beispiel für eine solche Thesaurusvereinigung. Eine andere Aufgabe für Thesaurusvereinigungen ist die eines Meta-Thesaurus für Thesaurus-Wechsel-Systeme. Der oben erwähnte Thesaurus „Meta“ etwa wurde so konstruiert [Rad87, RM87, Rad90]. Weitere Arbeiten zur Integration von Medizin-Thesauri werden z.B. in [SB92] vorgestellt. Auch der in [SC97] vorgestellte Ansatz liefert als Ergebnis eine Thesaurusvereinigung.

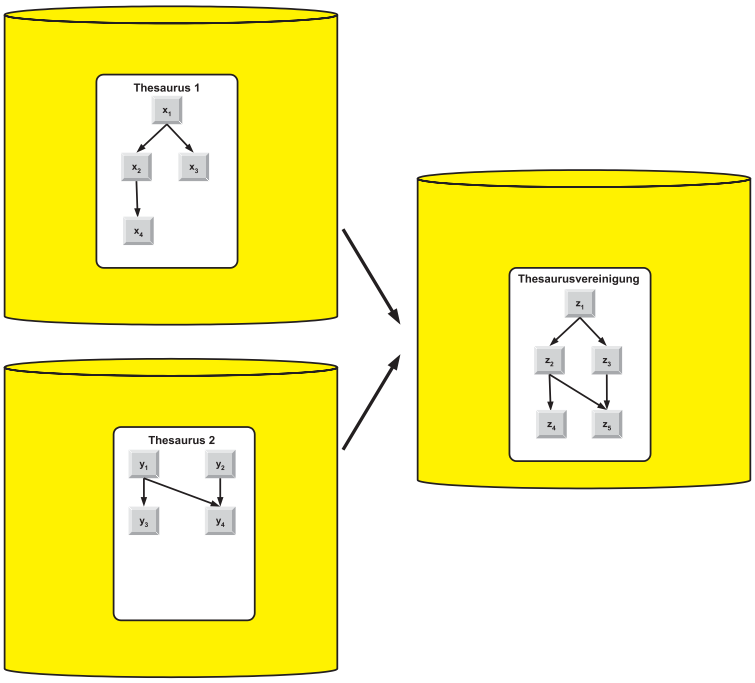

Abbildung 3.5: Thesaurusvereinigung

3.1.2 Modelle für Multi-Ontologie-Systeme

Da schon der Begriff *Ontologie* sehr unterschiedlich definiert wird, ist in der Literatur auch eine sehr große Spannweite von Multi-Ontologie-Systemen anzutreffen. Daher konzentrieren wir uns auf Projekte, die als Zielsetzung eine Integration bei weitgehender Autonomieerhaltung der

Komponentensysteme (keine oder minimale Änderungen) haben und somit eine unserer zentralen Anforderungen erfüllen.

3.1.2.1 Die ONIONS-Methodologie

Die in [GPG99, GPS98, PAG99] vorgestellte ONIONS-Methodologie (hervorgegangen aus dem GALEN-Projekt [Rec99]) geht zum Teil über eine Integration mit der Zielsetzung einer Unterstützung des Information-Retrieval-Prozesses hinaus (formale Schlussfolgerungen und Berechnungen auf Basis einer integrierten Ontologie). Die Integration findet mittels einer neuen, formalen Ontologie (repräsentiert mittels deskriptiver Sprachen) statt, die anhand von generischen Ontologien und den zu integrierenden Ontologien erstellt wird. Die Einträge der Komponentenontologien können nun mit Äquivalenzbeziehungen auf Einträge der neuen Ontologie abgebildet werden.

3.1.2.2 Der OBSERVER-Ansatz

Innerhalb des OBSERVER-Ansatzes [Men98, MIKS00] wird auf den Aufbau einer Interlingua verzichtet. Stattdessen werden direkt Beziehungen zwischen den Begriffen der verschiedenen Ontologien hergestellt. Dabei sind außer Äquivalenzbeziehungen (synonym relationship) auch generische Hierarchiebeziehungen (hyponym/hypernym relationship), Überlappungsbeziehungen (overlap relationship) und abgelehnte Beziehungen (disjoint relationship) möglich [MIKS00, S. 17f]. Überlappungsbeziehungen erlauben es auszudrücken, zu welchen Anteilen sich Begriffe überlappen, die weder Synonym sind noch in einer generischen Beziehung stehen. Als Beispiel wird $< Publikation, 90\% > overlap < Buchbeitrag, 20\% >$ aufgeführt. Dabei dient die Betrachtung der Extensionen zur Feststellung des Überlappungsgrades.

Obwohl der Fokus des OBSERVER-Ansatzes die Anfragebearbeitung bei der Verwendung von Multi-Ontologie-Systemen ist, findet eine Betrachtung von Konfliktfällen nicht statt.

3.1.2.3 Scalable Knowledge Composition (SKC)

Im Rahmen des Projektes „Scalable Knowledge Composition“ (SKC) [MWJ99, JMN+99, JW99] an der Universität Stanford wird die Integration von Ontologien über eine Artikulationsontologie (articulation ontology) genannte Interlingua hergestellt. Die Begriffe in der Artikulationsontologie können in beliebig zu definierenden Beziehungen untereinander und zu Begriffen der Komponentenontologien stehen. Die durch die Beziehungen zur Artikulationsontologie ausgedrückten Beziehungen zwischen Komponentenontologien werden als Artikulationspunkte bezeichnet. Zusätzlich ist es möglich, abgelehnte Beziehungen auszudrücken.

Eine Erkennung und Behandlung von Konflikten findet ausschließlich durch die Möglichkeit der Einführung Ergänzender Begriffe in die Artikulationsontologie statt.

3.1.2.4 Chimera

In [MFRW00] wird mit Chimera ein weiter Ansatz zum Verbinden von Ontologien entwickelt. Zur Repräsentation der Ontologien wird das OKBC-Modell, das ein Frame-basiertes System ist, verwendet [CFF+98a, CFF+98b]. Da das zentrale Ziel der Entwickler von OKBC die maximale Allgemeingültigkeit und Interoperabilitat zwischen Wissensrepräsentationssystemen war, bleibt das Modell sehr allgemein und lässt eine Reihe von Freiräumen offen.

Integrationsergebnis im Chimera-Ansatz ist wie bei ONIONS eine neue Ontologie. Einfache Konflikte in zugrunde liegenden Ontologien (z.B. Zyklen in Hierarchien) werden erkannt und in einem Vorverarbeitungsschritt aufgelöst. Das Zusammenfassen von äquivalenten Begriffen steht im Vordergrund, zusätzlich werden hierarchische Beziehungen zwischen Begriffen gesucht. Die maschinellen Vorschläge für Integrationsstellen beruhen im Wesentlichen auf einer Untersuchung der Herkunft eines Begriffes und auf lexikalischen Vergleichen der Benennungen. Falls durch die Integration Konflikte in Form von logischen Inkonsistenzen entstehen, werden diese durch Modifikationen der neuen Ontologie aufgelöst. Die Art und Anzahl der Konsistenzprüfungen ist jedoch auf wenige einfache Prüfungen, wie fehlende Definitionen oder redundante Ober-Begriffe (dort Klassen genannt), eingeschränkt.

Besonders hinzuweisen ist darauf, dass Chimera ein interaktives Werkzeug zum Fusionieren von Ontologien ist. Zwar zeigt Chimera dem menschlichen Integrator mögliche Integrations- und Problemstellen – die Entscheidung jedoch, *was* an diesen Stellen zu tun ist, wird vollständig dem Menschen überlassen.

3.1.2.5 PROMPT

Mit PROMPT wird ein weiterer Frame-basierter Ansatz zum automatischen Fusionieren von Ontologien vorgestellt [NM00]. Die Slots in diesem Modell bezeichnen aber im Wesentlichen Bestandsunterbegriffe, Beziehungen zwischen den Frames sind Abstraktionsbeziehungen. Abgesehen von der in jeder Ontologie üblichen Instanzenebene, die bei Thesauri keine Rolle spielt, gehen nur die Facets, durch die als tertiäre Relationen zusätzliche Constraints wie Kardinalitäten und Wertebereiche ausgedrückt werden können, über die Ausdrucksmöglichkeiten innerhalb eines Thesaurus hinaus.

Wie bei ONIONS und Chimera wird mit PROMPT aus den Ausgangsontologien eine neue, fusionierte Ontologie erzeugt. Dazu wird ein einfacher Algorithmus entwickelt. Konflikte wie Namenskonflikte und nicht-aufgelöste Verweise, die entstehen können, werden markiert und durch den menschlichen Integrationsexperten aufgelöst.

3.1.3 Bewertung der Modelle

Die Ergebnisse einer Bewertung der vorgestellten Typen von Modellen für Multi-Thesaurus- und Multi-Ontologie-Systeme hinsichtlich der Anforderungen aus Abschnitt 2.3.1 werden in Tabelle 3.1 dargestellt.

Es wird ersichtlich, dass keines der bekannten Modelle für Multi-Thesaurus-Systeme alle Anforderungen erfüllt. Die Kopplung der Komponententhesauri ist entweder zu lose (Extrembeispiel ist die Multi-Thesaurus-Umgebung) oder zu eng (Extrembeispiel ist die Thesaurusvereinigung). Ersteres führt dazu, dass Zusammenhänge nicht dargestellt werden können, letzteres zu einem Verlust der Autonomie. Thesauruskopplungen sind zu sehr auf den Spezialfall der Kopplung eines allgemeinen Dachthesaurus mit einem Spezialthesaurus zugeschnitten, als dass sie auf unsere Fallstudie übertragen werden könnten. Am weitesten erfüllt werden können die Anforderungen von einem Thesaurus-Wechsel-System, das eine Interlingua verwendet. Allerdings kann – falls überhaupt – nur eine unbereinigte Gesamtansicht des Vokabulars (Komponententhesauri und Interlingua) präsentiert werden, da eine Konfliktbehandlung fehlt. Eine Interlingua wird im Wesentlichen zum Einführen von Ergänzenden Begriffen und der Repräsentation einer abweichenden Struktur des integrierten Vokabulars von den Komponententhesauri verwendet.

Auch Multi-Ontologie-Ansätze können nicht alle Anforderungen erfüllen. Ein Teil dieser Ansätze

	Multi-Thesaurus-Umgebung	Thesaurus-Wechsel-System	Thesaurus-kopplung	Thesaurus-vereinigung	Ontologie-komposition
Thesauri					
Autonomieerhaltung der Komponententhesauri	ja	ja	ja	nein	ja
Unterscheidung Indexierungs-/ Nicht-Indexierungs-Thesaurus	nein	ja	nein	nein	nein
Relationen					
1:1-Äquivalenzrelation	nein	ja	nein	n.a.	ja
1:n-Äquivalenzrelation	nein	ja	nein	n.a.	ja
Hierarchierelation	nein	nein (*)	ja	n.a.	ja
Assoziationsrelation	nein	nein (*)	nein	n.a.	ja
Begriffe					
Föderierte Begriffe	nein	ja	ja	nein	ja
Ergänzende Begriffe	nein	nein (**)	nein	ja	ja
Homonymauflösung	nein	n.a.	nein	ja	nein
Invarianten und Konflikte					
Invarianten	Thesaurus-Invarianten gelten	nicht in bekannten Systemen	nicht in bekannten Systemen	Thesaurus-Invarianten gelten	nicht in bekannten Systemen
Konfliktmarkierung	n.a.	nein	nein	nein	nein
Konfliktbereinigung	n.a.	nein	nein	ja	nein

Erläuterung der Abkürzungen:
n.a. = nicht anwendbar
(*) = über die Beziehungen innerhalb einer Interlingua möglich
(**) = bei Verwendung einer Interlingua möglich

Tabelle 3.1: Bewertung der Modelle für Multi-Thesaurus-Systeme

kann im Hinblick auf das zugrundeliegende Modell einem Typ von Multi-Thesaurus-System zugeordnet werden (das ONIONS-Modell als Thesaurus-Wechsel-System mit Interlingua, Chimera und PROMPT als Thesaurusvereinigung) und erfüllt somit keine weiteren Anforderungen. Die weiteren vorgestellten Ansätze bezeichnen wir als *Ontologiekompositionen*. Diese erlauben zwar das Ausdrücken vielfältiger Beziehungen zwischen Begriffen verschiedener autonomer Komponentenontologien, ohne aber Invarianten explizit zu machen oder Verstöße gegen diese zu erkennen und zu behandeln (OBSERVER-Ansatz, SKC-Ansatz). Welche Typen für Beziehungen zwischen den Ontologien tatsächlich verwendet werden, bleibt jeweils den Integratoren überlassen. Diese Freiheit führt zwar zu einer großen allgemeinen Anwendbarkeit, aber aufgrund der potenziell sehr unterschiedlichen Semantik sind dann auch integrationsunterstützende Werkzeuge jeweils maßzuschneidern. Des Weiteren erlauben die Ontologiekompositionen nicht die Unterscheidung zwischen Indexierungs- und Nichtindexierungsontologien. Dies kann darauf zurückgeführt werden, dass als Einsatzzweck der Ontologien – im Gegensatz zu Thesauri – die Unterstützung des Information Retrieval nicht im Vordergrund steht.

Die bei OBSERVER verwendete Überlappungsbeziehung ist in solchen Fällen sinnvoll, in denen der Überlappungsgrad festgestellt werden kann und es keine andere Möglichkeit gibt, solch eine Überlappung der Bedeutungen auszudrücken. Aufgrund der Schwierigkeiten der Feststellung des Überlappungsgrades, der für Thesaurus-Benutzer ungewohnten Semantik und der Möglichkeit, solche Beziehungen bei für das Information Retrieval vertretbarem Semantikverlust durch andere Beziehungen (z.B. Assoziationsbeziehung) auszudrücken, erscheint diese Beziehung in Multi-Thesaurus-Systemen nicht sinnvoll. Erst bei der Notwendigkeit der exakten Bestimmung des Informationsverlustes bei der Anfragebearbeitung kommen die Vorteile von Überlappungsbeziehungen zum Tragen.

3.2 Begriffsintegration

Die Begriffsintegration ist die zentrale, aber auch komplexeste Aufgabe der Thesaurusintegration. Anhand der dabei verwendeten Datenbasis klassifizieren wir die möglichen Ansätze zur Begriffsintegration. Entsprechend den in Abschnitt 2.3.2 identifizierten Teilaufgaben der Begriffsintegration analysieren wir anschließend den diesbezüglichen Stand der Forschung.

3.2.1 Datengrundlagenorientierte Klassifikation der Ansätze

Wir unterscheiden die aus der Literatur bekannten Ansätze der Begriffsintegration in dokumentenbestandsorientierte, thesaurusbasierte und anfrageorientierte Ansätze. Diese Klassifikation wird nicht anhand der eingesetzten Verfahren getroffen, die durchaus jeweils in den unterschiedlichen Ansätzen eingesetzt werden können, sondern anhand der zentralen Datengrundlage, die zum Auffinden und Klassifizieren von Inter-Thesaurus-Relationen dient. Eine beliebige Kombination der Ansätze ist möglich.

3.2.1.1 Dokumentenbestandsbasierte Ansätze

Bei den *dokumentenbestandsbasierten Ansätzen* wird davon ausgegangen, dass eine Menge gleicher Dokumente mit verschiedenen Thesauri indiziert worden ist. Informationen innerhalb der Thesauri spielen nur eine untergeordnete Rolle, vom Bedeutungsinhalt und -umfang der Begriffe wird abstrahiert. Stattdessen werden Beziehungen zwischen den Begriffen verschiedener

Thesauri identifiziert, indem die Verwendung der Begriffe zur Indexierung des gemeinsamen Dokumentenbestandes in den unterschiedlichen Informationssystemen analysiert wird.

Ein Beispiel für solch einen dokumentenbestandsbasierten Ansatz wird von Amba, Narasimhamurthi, O'Kane und Turner in [Amb92, ANOT96] vorgestellt. Es wird ein Cluster-Algorithmus zur Gruppierung der Benennungen entsprechend ihres gemeinsamen Auftretens eingesetzt, um 1:n-Abbildungen der Begriffe eines Nicht-Indexierungsthesaurus auf die Begriffe eines Indexierungsthesaurus (Thesaurus-Wechsel-System) herzustellen.

Vorteile der dokumentenbestandsbasierten Ansätze sind eine leichte Formalisierung und Automatisierung. Analysen der Semantik der Begriffe und ihrer komplexen Zusammenhänge sind ebensowenig notwendig wie Modelle über die Art der Thesauruserstellung und -benutzung. Allerdings können so ausschließlich Äquivalenz- und Benutze-Kombination-Beziehungen entdeckt werden, weitere Typen von Beziehungen erfordern verstärkt eine Berücksichtigung der Semantik und des jeweiligen Thesauruskontextes. Wesentlicher Nachteil aber ist, dass die Voraussetzung, eine genügend große Menge von Dokumenten, die mit beiden Thesauri indexiert ist, zur Verfügung zu haben, in der Regel nicht erfüllt ist. Dies gilt insbesondere dann, wenn mehr als zwei Thesauri integriert werden sollen. Selbst bei ausreichend großer mit allen Thesauri indexierter Dokumentenmenge, werden bei Benennungen, die selten oder gar nicht gleichzeitig mit Benennungen eines anderen Thesaurus verwendet worden sind, keine oder nur schlecht abgesicherte Verbindungen identifiziert.

Aufgrund der in der Regel nicht erfüllten oben beschriebenen Voraussetzung werden dokumentenbestandsbasierte Ansätze im Rahmen dieser Arbeit nicht weiter betrachtet.

3.2.1.2 Thesaurusbasierte Ansätze

Theaurusbasierte Ansätze verwenden als wesentliche Datengrundlage zur Integration die Thesauri selber und nicht die mit den Thesauri indexierten Dokumentenbestände. Es findet eine mehr oder weniger umfangreiche Analyse der unterschiedlichen Informationen (z.B. der Benennungen, Beziehungen, Klassifikationen, Definitionen, Erläuterungen) in den Thesauri statt, um Inter-Thesaurus-Relationen, Konflikte und Ergänzende Begriffe zu erkennen. Da weitere Voraussetzungen zur Integration entfallen, ist die überwiegende Anzahl der in der Literatur anzutreffenden Ansätze thesaurusbasiert (z.B. [Ait81, Rad87, MR88, SB92, SC97, WSN98, AF99]) bzw. ontologiebasiert (z.B. [RPR$^+$98, GPS98, PAG99, MWJ99, JMN$^+$99]).

Der wesentliche Nachteil rein thesaurusbasierter Ansätze ist die Nicht-Berücksichtigung der tatsächlichen Verwendung der Thesaurusbegriffe. Um unabhängig von der bisherigen Verwendung allgemeingültige Integrationsmöglichkeiten zu finden, ist daher eine möglichst umfangreiche Analyse der Begriffe in ihren jeweiligen Thesaurus-internen Kontexten erforderlich. Da hierzu semantisches Wissen benötigt wird, von dem nicht vorausgesetzt werden kann, dass es allein aufgrund der Thesaurusinformationen abgeleitet werden kann, ist die Einbeziehung entsprechender Experten unabdingbar. Eine vollständige Automatisierung erscheint nicht möglich, stattdessen wird ein semi-automatisches Vorgehen bei weitestgehender Entlastung des menschlichen Experten angestrebt.

Da thesaurusbasierte Ansätze in vielen Fällen die einzig praktikable Lösung sind, wird im Rahmen dieser Arbeit der Schwerpunkt auf diesen Ansätzen liegen.

3.2.1.3 Anfragebasierte Ansätze

Obwohl uns aus der Literatur keine reinen anfragebasierten Ansätze bekannt sind, stellen wir diese Möglichkeit an dieser Stelle der Vollständigkeit halber dar. Wesentliche Grundlage eines anfragebasierten Ansatzes zum Entdecken von Beziehungen zwischen Begriffen verschiedener Thesauri ist die Auswertung von Benutzeranfragen sowie von den vom System gelieferten Ergebnisdokumenten. Während bei dokumentenbestandsbasierten und bei thesaurusbasierten Ansätzen die wesentlichen Aufgaben der Integration der Thesauri vor der Benutzung des Multi-Thesaurus-Systems durchgeführt werden, findet die Integration bei anfragebasierten Ansätzen erst während der Benutzung des Multi-Thesaurus-Systems statt. Das System lernt Zusammenhänge z.B. anhand einer Auswertung innerhalb einer Anfrage benutzter Begriffe aus verschiedenen Thesauri oder einer Analyse, welche Begriffe eines Thesaurus in einem Dokument vorkommen, das mit einer Anfrage aus Begriffen eines anderen Thesaurus gefunden wurde.

Eine mögliche Realisierung eines anfragebasierten Ansatzes besteht im Einsatz von Techniken des *fallbasierten Schließens* [AP94]. In die Fallbasis können ausgewählte bisherige Anfragen und Ergebnisse sowie deren Bewertung aufgenommen werden. Neue Anfragen können dann mit ähnlichen alten Anfragen verglichen werden. Mit diesem Ansatz kommen allerdings die generellen Schwierigkeiten des fallbasierten Schließens zum Tragen, zum einen die Ähnlichkeit der Anfragen zu bestimmen, und zum anderen die Schlüsse für eine ähnliche Lösung zu ziehen.

3.2.2 Verfahren zum Auffinden von Ergänzenden Begriffen

Prinzipiell unterschieden werden kann zwischen Begriffsintegrationsverfahren, die das Auffinden und Bewerten von Inter-Thesaurus-Beziehungen zum Ziel haben und solchen, die Ergänzende Begriffe auffinden sollen.

Das automatische Herleiten zumindest der Positionen innerhalb von Begriffsnetzen für Ergänzende Begriffe wird unserer Kenntnis nach bisher ausschließlich durch die Formale Begriffsanalyse [Wil98] unterstützt. Die hier notwendige Voraussetzung einer vollständigen und eindeutigen Attributierung der Begriffe durch die sie charakterisiernden Merkmale ist bei der Integration von Thesaurusbegriffen aber nicht gegeben. Diese Voraussetzung für die verschiedenen Komponententhesauri herzustellen, wäre mit unvertretbar hohem Aufwand – der im Wesentlichen intellektueller Art ist und daher von einem menschlichen Experten erbracht werden müsste – verbunden. Daher können die Verfahren der Formalen Begriffsanalyse nicht auf die Begriffsintegration übertragen werden.

Wenn das Auffinden Ergänzender Begriffe bisher überhaupt eine Rolle bei den verschiedenen Ansätzen der Begriffsintegration spielt, dann indem die Position und auch die Benennung manuell erkannt werden. Dies ist z.B. innerhalb des SKC-Ansatzes [MWJ99, MWK99, JMN+99, JW99] der Fall. Hier können zur Auflösung von manuell erkannten Konflikten vom Integrationsexperten Ergänzende Begriffe eingebracht werden.

3.2.3 Verfahren zum Auffinden und Klassifizieren von Inter-Thesaurus-Relationen

Der Prozess des Auffindens und Klassifizierens von Inter-Thesaurus-Relationen beinhaltet das Erkennen von Beziehungen unterschiedlicher Typen zwischen Begriffen verschiedener Thesauri unter möglichst vollständiger Berücksichtigung ihres jeweiligen Kontextes. Innerhalb dieses Prozesses werden in den bekannten Ansätzen eine Reihe unterschiedlicher Verfahren eingesetzt.

3.2.3.1 Linguistische Verfahren

Die Linguistik bietet auf Ebene der Phoneme, Morpheme, Wörter, Phrasen, Sätze und Texte Verfahren zur Analyse der Sprache an (vgl. Anhang D.3). In den uns bekannten Ansätzen der Begriffsintegration werden diese Verfahren auf die Benennungen der Thesauri angewandt und bieten häufig die Grundlage der Integration. Aufgrund der Beschaffenheit der Benennungen (Ein- und Mehrwortbenennungen, die überwiegend Nomen, aber auch adjektivische Phrasen sind) spielen als zu untersuchende Einheiten ausschließlich Morpheme, Wörter und Phrasen eine Rolle. Eine Analyse von Sätzen und Texten könnte, wenn sie auf die Definitionen und Erläuterungen zu den Begriffen angewendet würde, zwar zu einem besseren maschinellem „Verständnis" führen, wird aber aufgrund der Komplexität der Verfahren in den uns bekannten Ansätzen nicht durchgeführt.

3.2.3.1.1 Vergleich von Zeichenketten Die einfachsten, aber in der Computerlinguistik unverzichtbaren Verfahren sind jene, die auf dem Vergleich von Zeichenketten beruhen.

Identifikation identischer Zeichenfolgen: Zwei Zeichenfolgen werden Zeichen für Zeichen auf Identität verglichen. Bei der Begriffsintegration werden mit dem Vergleich – evtl. vorher normierter – Zeichenfolgen identische Benennungen in verschiedenen Komponententhesauri gefunden. Von diesen wird häufig geschlossen, dass es sich auch um identische Begriffe handelt.

Beispielsweise wird in [MR88, SC97] eine Äquivalenzannahme ausschließlich durch eine Identifikation identischer Zeichenfolgen aufgestellt. Eine Bestätigung der Äquivalenzannahmen kann durch weitere Verfahren erfolgen.

Identifikation ähnlicher Zeichenfolgen: Statt exakte Übereinstimmungen beim Zeichenfolgenvergleich zu fordern, werden bei der Identifikation ähnlicher Zeichenfolgen Unterschiede bis zu einem gewissen Maße zugelassen, um ähnliche Benennungen zu identifizieren, von denen geschlossen wird, dass sie identische/ähnliche Begriffe bezeichnen.

Als ähnliche Zeichenfolgen werden in [Rad87] Zeichenfolgen identifiziert, die identische Teilzeichenfolgen einer minimalen Länge besitzen.

Im Information Retrieval verwendete Ähnlichkeitsmaße zweier Zeichenfolgen sind beispielsweise der *Hamming-Abstand* (für Zeichenketten gleicher Länge ist der Abstand als die Anzahl der Stellen mit unterschiedlichen Zeichen definiert) oder *Levenstein-Abstand* (auch *edit distance*; kleinste Anzahl von Einfüge-, Lösch- und Ersetzungs-Operationen, um zwei Zeichenfolgen identisch zu bekommen) [BYRN99]. Sie haben das grundsätzliche Ziel, Benennungen erfolgreich zu vergleichen, auch wenn syntaktische Fehler vorliegen. Bei der Thesaurusintegration können sie verwendet werden, um Schreibvarianten zu behandeln (z.B. color und colour, by-product und byproduct). Uns ist aber kein Ansatz der Thesaurusintegration bekannt, der lexikalische Abstandsmaße mit dieser Zielsetzung verwendet.

3.2.3.1.2 Morphologische Analysen Verfahren der morphologischen Analyse untersuchen Formvarianten, Vorkommen und Funktion der Morpheme [Bußß90]. Es kann weiter zwischen Verfahren der Flexionsanalyse und Verfahren der Wortbildungsanalyse unterschieden werden. Erstere betrachten die Zusammenhänge zwischen den durch Flexion (Deklination (Nomen), Konjugation (Verb) oder Komparation (Adjektiv)) entstandenen Wörtern und den Wörtern

in Grundform. Letztere betrachten die Zusammenhänge bei der Bildung komplexer Wörter (Derivation: Bildung komplexer Einwortbenennungen, Komposition: Bildung von Komposita) aus sprachlichen Einheiten.

Wortstammbildung: Verfahren der Wortstammbildung sind Wortbildungsanalyseverfahren, die ein Wort (in Grundform) auf seinen Wortstamm zurückführen. Zur Wortstammbildung englischer Wörter wird in der Regel Porters Algorithmus [Por80] eingesetzt. Beispiel ist das in [SC97] vorgestellte Verfahren der Thesaurusintegration, das die Wortstammbildung zur Normierung der Bezeichner verwendet. Zwar kann bei lexikalischen Vergleichen der Wortstämme eine bessere Trefferquote erzielt werden, doch es wird auch die Homonymproblematik vergrößert, indem Wörter unterschiedlicher Bedeutung auf den gleichen Stamm zurückgeführt werden (Beispiel: minister und ministry werden beide auf den Wortstamm minister zurückgeführt). Exaktere Vergleiche ermöglicht die Lemmatisierung.

Lemmatisierung: Durch die Lemmatisierung kann ein Wort auf seine Grundform zurückgeführt werden. Bei der Begriffsintegration ist dies z.B. erforderlich, wenn die Benennungen des einen Thesaurus in Pluralform sind, während die Benennungen des anderen Thesaurus in Singularform sind. Zusätzlich können auch Verben in die Infinitivform zurückgeführt werden. Die Lemmatisierung wird z.B. in [MR88] verwendet, um die Vergleichbarkeit der Benennungen zu erhöhen.

Wortart-Konversion: Die Wortart-Konversion überführt ein Wort aus einer Wortart (z.B. Adjektiv *ecological*) in eine andere Wortart (z.B. Nomen *ecology*). Für englische Adjektive werden Regeln für eine solche Wortart-Konversion von Adjektiven zu Nomen z.B. in [PP69] beschrieben. Diese Möglichkeit wird in keinem uns bekannten Ansatz verwendet, kann aber zur besseren Vergleichbarkeit von Benennungen, insbesondere Mehrwortbenennungen, vergleichbar der Identifikation ähnlicher Zeichenfolgen, beitragen.

Analyse von Mehrwortbenennungen: Eine einfache *Analyse von Mehrwortbenennungen zum Auffinden von Abstraktionsbeziehungen* erfolgt in [MR88]. Wenn eine Mehrwortbenennung A aus allen Teilwörtern der Benennung B besteht und genau ein zusätzliches Wort enthält, wird angenommen, dass der durch A bezeichnete Begriff ein Abstraktionsunterbegriff von B ist (z.B. *programming language* ist Abstraktionsunterbegriff von *language*). Das zusätzliche Wort wird als Modifikator betrachtet, der den Begriffsinhalt näher spezifiziert und somit den Begriffsumfang einschränkt.

Wortpermutationen und Wortentfernungen werden durchgeführt, um unterschiedliche Schreibweisen von Mehrwortbenennungen zu vereinheitlichen. Es können sehr häufig vorkommende Wörter wie *the*, *of* oder *from* entfernt und die Reihenfolge der restlichen Wörter vertauscht werden. Wenn gleichzeitig eine Lemmatisierung durchgeführt wird, kann z.B. *Theory of Systems* auf *System Theory* zurückgeführt werden und somit die Grundlage für eine bessere Erfolgsquote bei der Identifikation identischer Zeichenfolgen geschaffen werden. Dieser Ansatz wird z.B. in [SC97] verfolgt.

Insgesamt lässt sich feststellen, dass bisher nur sehr einfache Analysen der Mehrwortbegriffe durchgeführt werden und das Potential der linguistischen Mittel nicht ausgeschöpft wird (z.B. Mehrwortbildungsregeln, Wortklassenerkennung). Zu dem in [SC97] beschriebenen Ansatz sei angemerkt, dass eine Permutation der Wortreihenfolge bei dem Entfernen von häufig gebrauchten Wörtern nicht immer die richtige Entscheidung ist, da die Bedeutung hierdurch grundlegend verändert werden kann. Ein leicht einsichtiges Beispiel ist die Konjunktion *and*: Aus *urban and rural planning* darf nicht das unsinnige *rural planning urban*

werden. Stattdessen sind solche Fälle gesondert zu berücksichtigen. Weiteres Beispiel ist der Artikel *the*: Bei *rights of the individual* reicht es, das Wort zu löschen, da die Präpositionsentfernung von *of* bereits die Wörter permutiert. Bei *chewing the cud* hingegen soll eine Wortpermutation bei der Entfernung von *the* stattfinden.

3.2.3.2 Strukturbasierte Verfahren

Verfahren, die die Struktur auswerten, können generell unterschieden werden in solche, die versuchen, identische (oder ähnliche) Graphstrukturen in verschiedenen Thesauri aufzufinden, um daraus Vorschläge für Inter-Thesaurus-Beziehungen herzuleiten, und in solche, die die Graphstruktur verwenden, um semantische Abstandsmaße zwischen Begriffen innerhalb einer Struktur zu berechnen. Die erste Art von Verfahren bezeichnen wir als Verfahren zum Auffinden von Isomorphismen, die zweite Art als semantische Abstandsmaße.

3.2.3.2.1 Auffinden von Isomorphismen Abbildungen, die die Knoten und Kanten eines Graphen umkehrbar eindeutig aufeinander abbilden und dabei seine Struktur erhalten, werden als *Isomorphismen* bezeichnet. Da die Begriffe eines Thesaurus als Knoten und die Beziehungen als (gerichtete) Kanten eines Graphen betrachtet werden können, können durch das Auffinden von Isomorphismen bei Teilgraphen Vorschläge für identische Teilstrukturen generiert werden. Die Graphentheorie hat eine Reihe von Algorithmen zur Erkennung von Isomorphien entwickelt (z.B. Kanonische Numerierungen [CG95], Tree-Editing- oder Tree-Alignment-Algorithmen [JWZ95]).

Aufgrund eines wenig klaren Entwurfsraumes der Ismorphie-Erkennungs-Algorithmen und fehlender Anpassungsmöglichkeiten werden solche Algorithmen aber unseres Wissens nach lediglich in dem Ansatz von [BK99] verwendet. Hier wird allein die Abstraktionsrelation betrachtet. Es wird ein eigens entwickelter heuristischer Algorithmus zum Auffinden der Isomorphien angewandt. Für die empirischen Untersuchungen wurde der ideale Fall hergestellt, zwei strukturell und bezüglich der dargestellten Diskurswelt identische „Ontologien“ zu haben, die sich ausschließlich in ihren Bezeichnern unterscheiden. Bei dieser idealisierten Voraussetzung wurde festgestellt, dass es häufig strukturelle Übereinstimmungen (Automorphismen) zwischen *unterschiedlichen* Teilgraphen gab, so dass eine Betrachtung der Graphstrukturen ohne Berücksichtigung der Knoteninhalte, also der Benennungen, wenig hilfreich ist.

Generell kann angemerkt werden, dass nicht von der Annahme ausgegangen werden kann, dass eine Kante, die in einem Komponententhesaurus vorhanden ist, auch in einem anderen Komponententhesaurus vorhanden sein muss. Verfahren zum Auffinden von Isomorphismen sind gegenüber fehlenden oder anderen Kanten aber zu wenig tolerant, als dass sie wirkungsvoll eingesetzt werden können. Nur bei Strukturen, die sehr etabliert und allgemein anerkannt sind (z.B. wissenschaftliche Taxonomien) und daher voraussichtlich in den unterschiedlichen Thesauri identisch abgebildet sind, kann ein größerer Nutzen erwartet werden. In diesem Fall sind aber häufig bereits übereinstimmende Benennungen vorhanden und die strukturelle Gleichheit spielt eine nachgeordnete Rolle.

3.2.3.2.2 Semantische Abstandsmaße Die Ähnlichkeit zweier Begriffe wird durch *Abstandsmaße*, die durch Abstandsfunktionen berechnet werden, ausgedrückt. Diese Abstandsfunktionen können auf lexikalischen Vergleichen, auf einer Auswertung des gemeinsamen Vorkommens innerhalb eines Dokumentenbestandes oder aber auf einer Auswertung der Abstraktionshierarchien basieren. Im ersten Fall wird eine Ähnlichkeit der Bezeichner gemessen (syn-

taktische Ähnlichkeit, s. Abschnitt 3.2.3.1), im zweiten und im dritten Fall die Ähnlichkeit von Begriffen (semantische Ähnlichkeit).

Bei der Begriffsintegration ist eine große semantische Ähnlichkeit zweier Begriffe ein Indikator dafür, dass eine Inter-Thesaurus-Beziehung etabliert werden sollte. Dies gilt insbesondere, wenn zusätzlich lexikalische Ähnlichkeiten festgestellt wurden. Der Berechnung semantischer Abstandsmaße liegt in der Regel eine Auswertung der Abstraktionshierarchie zugrunde. In einigen Ansätzen werden weitere Informationen ausgewertet.

Mili und Rada [MR88] definieren den semantischen Abstand zweier Begriffe als die Länge des kürzesten Pfades von Abstraktionsbeziehungen zwischen den beiden Begriffen. Darauf basierend wird ein Abstandsmaß für zwei Begriffsmengen definiert. Dieses Abstandsmaß wird aber zum Vergleichen von Anfragen mit Dokumenten verwendet und nicht zur Begriffsintegration. Ähnliche Ansätze, die ebenfalls auf einem Spreading-Activation-Modell basieren, werden z.B. in [Pai91, CLBD93] vorgestellt.

Spanoudakis definiert in [SC96] die Abstraktionsdistanz zweier Objekte als Summe der Wichtigkeiten ihrer nicht-gemeinsamen Oberklassen. Dabei ist die Wichtigkeit der Kehrwert der Tiefe innerhalb der Abstraktionshierarchie. Nicht-gemeinsame Oberklassen auf den oberen Ebenen fallen somit stärker ins Gewicht als auf den unteren Ebenen. Zu einem Gesamtabstand werden zusätzlich ein Klassifikationsabstand und Attributierungsabstand berechnet, die hier nicht relevant sind.

Sinitichiakis greift die in [SC96] vorgestellte Abstandsfunktion auf und überträgt sie auf die Abstandsbestimmung bei der Begriffsintegration [SC97]. Die Abstraktionsdistanz wird direkt übertragen, als Klassifikationsabstand wird die Anzahl nicht-gemeinsamer Gruppen (die hier Facetten genannt werden) definiert und es erfolgt eine Normierung. Da die Begriffsintegration von den Toptermen die Hierarchie absteigend durchgeführt wird (Top-Down), ist garantiert, dass alle Oberbegriffe eines zu untersuchenden Begriffs bereits integriert sind (durch die Etablierung von Äquivalenzbeziehungen) und somit der Abstand berechnet werden kann. Die Integrationsvorschläge in Form von Äquivalenzbeziehungen, die aufgrund lexikalischer Vergleiche gefunden werden, können so zusätzlich bewertet werden.

Während der Begriffsintegration können semantische Abstandsmaße wichtige zusätzliche Informationen darüber liefern, ob zwei Begriffe äquivalent sind. Die Berechnung fordert allerdings, dass entweder die Integration bereits abgeschlossen ist (Abstandsmaße von Mili und Rada [MR88, Pai91, CLBD93], die auf einer Spreading-Activation-Theorie basieren) oder zumindest die übergeordneten Begriffshierarchien bereits integriert sind (Abstandsmaß in [SC97], basierend auf analoger Ähnlichkeit und dem Kontrastmodell [Tve77]). Im ersten Fall kann das Maß somit während der Begriffsintegration nicht verwendet werden. Im zweiten Fall ist ein strenger Top-Down-Ansatz notwendig, innerhalb dessen über die in [SC97] gefundenen Äquivalenzen hinausgehend auch Abstraktionsober-/-unterbegriffe bereits etabliert und Konflikte bereinigt sind. Es wurde bisher nicht untersucht, inwiefern diese Einschränkungen gelockert werden können. Auch die Voraussetzung der identischen Gruppen in den Komponententhesauri wird in der weit überwiegenden Anzahl realer Fälle nicht gegeben sein. Es ist zu klären, inwieweit Vorarbeit zu leisten ist, um diese Voraussetzung herzustellen.

3.2.3.3 Attributbasierte Verfahren

Manche Thesauri bieten die Möglichkeit, Begriffe nicht ausschließlich über deren Bezeichnungen und Definitionen zu repräsentieren, sondern zusätzlich diese Begriffe charakterisierende Attribute angeben zu können (z.B. ein Fisch lebt im Wasser, kann Schwimmen, hat Kiemen, hat Schuppen).

Ein einfaches Verfahren zur Auswertung von Attributen wird in [Rad90, S. 163] vorgestellt. Anhand der Schnittmengenbildung entlang der Abstraktionspfade werden Vorschläge für Abstraktionsbeziehungen innerhalb einer Thesaurusvereinigung generiert.

Darüberhinausgehend stellt die Formale Begriffsanalyse [Wil98] umfangreiche theoretische Grundlagen für Begriffshierarchien bereit, die auf den Attributen der Begriffe basieren. Die Anwendung auf die Integration solcher Begriffshierarchien mit unterschiedlichen Attributen wurde aber bisher nicht betrachtet.

3.2.3.4 Verwendung externer Wissensquellen

In den meisten der uns bekannten Ansätze der Begriffsintegration wird keine externe Wissensquelle verwendet (z.B. in [SC97, Rou92, Squ93, Rad87, MR88]) oder aber es wird ausschließlich auf einen menschlichen Entscheider als externe Wissensquelle zugegriffen (z.B. in [MWJ99]).

Dennoch kann externen Wissensquellen eine wichtige Rolle zugeschrieben werden, wenn es darum geht, unterschiedliche Beziehungstypen zwischen den Begriffen der Komponententhesauri zu entdecken. Solche externen Wissensquellen können entweder direkt Informationen über Beziehungen zwischen Begriffen enthalten oder aber es besteht die Möglichkeit, diese Informationen aus den Wissensquellen herzuleiten.

Das bedeutenste Beispiel für den ersten Fall ist WordNet [Mil98, Fel98], eine frei verfügbare, umfangreiche elektronische lexikalische Datenbasis der englischen Sprache (s. auch Anhang D.2, S. 274), die an der Princeton Universität entwickelt wird. WordNet fasst in einem Kontext identische Benennungen zu so genannten Synsets (Synonymmengen) zusammen, die Begriffe repräsentieren. Zwischen diesen Begriffen sind u.a. Abstraktions- und Bestandsrelationen etabliert. Lexikalische Grundeinheit sind Wörter. Mehrwortbenennungen und Phrasen sind in geringem Umfang enthalten. Aufgrund des großen Umfangs (WordNet Version 1.6 enthält z.B. über 80.000 Substantive, die zu über 60.000 Begriffen zusammengefasst sind), bietet WordNet eine erfolgsversprechende Basis, wenn Beziehungen zwischen allgemeinen Begriffen (solche, die in einem allgemeinen Wörterbuch zu finden sind, WordNet deckt hingegen keine fachspezifischen Vokabulare ab) gefunden und klassifiziert werden sollen. Auch zur Bestimmung eines semantischen Abstandsmaßes können die Beziehungen zwischen den Begriffen in WordNet verwendet werden (vgl. Abschnitt 3.2.3.2.2). In [LSM95] trägt das semantische Abstandsmaß in WordNet zur Disambiguierung (Zweifelsfreiheit) der Wortbedeutung bei.

Für den zweiten Fall, der Herleitung von Beziehungen zwischen Begriffen, können unterschiedliche Quellen als Ausgangsbasis dienen. In der Regel werden Dokumentenbestände (Text-Korpora) eines bestimmten Fachgebietes analysiert. Als Beispiel sei der Ansatz von Viegener [Vie97] zur automatischen Generierung von Thesauri aus solchen Dokumentenbeständen genannt. Ebenfalls zur Generierung eines einfachen Thesaurus (Berücksichtigung nur einer Relationsart) werden in [Jan99] die textuellen Definitionen eines Wörterbuches ausgewertet. Dazu werden Begriffe und Beziehungen zwischen diesen Begriffen über die Definitionen extrahiert sowie diese Beziehungen mit einem auf PageRank [PB98] basierenden Verfahren gewichtet.

3.2.3.5 Bewertung

Die Koexistenz der vielfältigen Verfahren belegt, dass bisher nicht ein einzelnes Verfahren optimale Ergebnisse erzielen kann. Vielmehr sind linguistische Verfahren Grundlage jeder Integration. Um jedoch die Gefahr, verschiedene Begriffe mit mehrdeutigen Benennungen als identische Begriffe zu identifizieren, die mit der größeren Breite des Gebietes wächst [Rec99, S. 3], ist

eine ausschließliche linguistische Analyse der Benennungen nicht ausreichend. In der natürlichen Sprache wird die Bedeutung eines Begriffes durch seinen Kontext disambiguiert. Diese Art der Disambiguierung muss daher auch zum Erkennen potenzieller Integrationsstellen angewendet werden. Da der Kontext eines Begriffs innerhalb eines Thesaurus durch seine Beziehungen zu anderen Begriffen, durch Erläuterungen und Definitionen ausgedrückt wird, gilt es, ein Zusammenwirken von linguistischen Verfahren mit solchen, die diesen Kontext auswerten, zu unterstützen. Architekturen, die ein Zusammenwirken verschiedenartiger Verfahren zur möglichst umfangreichen Auswertung der Informationen in den Thesauri selber sowie das Einbeziehen externer Wissensquellen unterstützen, existieren aber bisher nicht.

Die bekannten Verfahren haben überwiegend das Ziel, äquivalente Begriffe zu finden (z.B. [SC97]). Das Auffinden anderer Relationstypen wird hingegen weitgehend vernachlässigt. Konfigurationsmöglichkeiten, um die Verfahren den Randbedingungen entsprechend anzupassen, wurden bisher völlig außer Acht gelassen. Eine Anpassung findet ausschließlich implizit an die „zufällig“ vorgegebenen Komponententhesauri statt.

3.2.4 Konflikterkennung und -behandlung

In der überwiegenden Anzahl der bekannten Ansätze wird eine Konflikterkennung manuell durchgeführt, die Kriterien für einen Konflikt sind dabei nicht explizit festgelegt. Als Beispiel sei der SKC-Ansatz [MWJ99, MWK99, JMN+99, JW99]) genannt.

Die Konfliktbehandlung innerhalb des SKC-Ansatzes [MWJ99, MWK99, JMN+99, JW99] erlaubt ausschließlich, bei auftretenden Konflikten Beziehungen nicht herzustellen oder aber Ergänzende Begriffe in die Artikulationsontologie einzuführen. Da die Autonomie der Komponentenontologien erhalten bleibt, werden widersprüchliche oder redundante Beziehungen (implizit) akzeptiert. Eine Markierung solcher Konflikte ist nicht vorgesehen.

In [BK99] wird Konfliktfreiheit gefordert, allerdings werden nur zwei Konflikttypen betrachtet: Die Injektivität einer Abbildung wird verletzt, wenn zwei unterschiedliche Begriffe des einen Thesaurus[1] auf denselben Begriff des anderen Thesaurus abgebildet werden. Es wird nicht erläutert, wie in einem Fall, bei dem ein Begriff des einen Thesaurus zwei oder mehr Begriffe des anderen Thesaurus umfasst, der Konflikt aufgelöst wird. Der andere Konflikttyp ist der Verstoß des Erhalts der Subsumtionsbeziehungen, d.h., Begriffe, zwischen denen es eine Inter-Thesaurus-Äquivalenzbeziehung gibt, müssen entweder in beiden Thesauri in einer (indirekten) Abstraktionsbeziehung gleicher Richtung stehen oder es darf in keinem Thesaurus eine Abstraktionsbeziehung zwischen den Begriffen geben. Diese Forderung geht über die Forderung der Zyklenfreiheit deutlich hinaus. Sie mag aus der sehr theoretischen Sichtweise in [BK99] gerechtfertigt sein, da dort gefordert wird, dass auch bei der Betrachtung der integrierten Thesauri die Bedeutung der Komponententhesauri *unverändert* sein muss. Es wird somit aber ausgeschlossen, dass durch die Integration Verbesserungen an der Gesamtstruktur entstehen. Eine solche Restriktion erscheint nicht sinnvoll. Stattdessen sollte geprüft werden, ob die durch die Integration implizierte Abstraktionsbeziehung im Kontext des Komponententhesaurus sinnvoll ist.

Auf technischer Ebene nennt bereits Rada in [MR88] zwei Konflikttypen, die als

- Verletzung der paarweisen Disjunktheit der Relationen
- und Verletzung der Redundanzfreiheit der Hierarchierelation

[1] In [BK99] wird zwar von Ontologien gesprochen, da keinerlei ontologische Besonderheiten berücksichtigt werden, kann die Arbeit aber auf Thesauri übertragen werden.

angesehen werden können. Der erste Fall bedeutet, dass ein Begriff nicht gleichzeitig in zwei verschiedenen Beziehungen mit ein und demselben anderen Begriff stehen darf. Dabei wird auch die Richtung der Beziehung berücksichtigt, so dass ein Begriff nicht gleichzeitig Ober- und Unterbegriff eines anderen Begriffs sein darf. Durch die Konfliktauflösung wird eine Beziehung entfernt bzw. bei widersprüchlichen Ober-/Unterbegriffsbeziehungen werden diese durch eine Assoziationsbeziehung ersetzt. Der zweite Fall bedeutet, dass es von einem Begriff nicht gleichzeitig einen direkten und einen indirekten Verweis auf ein und denselben Unterbegriff geben darf. In diesem Fall wird der direkte Verweis entfernt.

Formale Kriterien für die Zyklenfreiheit der Abstraktionsrelation und die paarweise Disjunktheit der Relationen werden in [SC97] aufgeführt. Es wird somit nur der erste Konflikttyp aus [MR88] wieder aufgegriffen und so erweitert, dass nicht nur direkte Zyklen, sondern auch indirekte Zyklen betrachtet werden. Im Gegensatz zur rein algorithmischen Konfliktauflösung in [MR88] wird in [SC97] das Vorgehen bei der Konfliktauflösung dem menschlichen Integrator überlassen.

Abgesehen von der Unvollständigkeit der Berücksichtigung der verschiedenen Konflikttypen, die schon daraus ersichtlich ist, dass in den unterschiedlichen Ansätzen unterschiedliche Typen behandelt werden, ist anzumerken, dass eine explizite Aufführung aller Invarianten für die Komponententhesauri und das Multi-Thesaurus-System sowie eine angemessene Konflikterkennung und Konfliktbehandlung in keinem der uns bekannten Ansätze berücksichtigt wurde.

3.2.5 Vorgehensmodelle

3.2.5.1 Mehrstufige Verfahren

Einfache Vorgehensmodelle, die wir als *mehrstufige Verfahren* bezeichnen, wenden ein Verfahren, ggf. leicht variiert, auf unterschiedliche Datengrundlagen an. Ein Beispiel ist das bei der Integration von SNOMED und EMTREE in [Rad90, S. 160] vorgestellte mehrstufige Verfahren:

- Ein lexikalischer Vergleich der Deskriptoren prüft in der ersten Stufe auf exakte Gleichheit.
- In der zweiten Stufe betrachtet der exakte lexikalische Vergleich zusätzlich die Nicht-Deskriptoren.
- In der dritten Stufe wird erst die Wortreihenfolge von Komposita vertauscht und anschließend der exakte lexikalische Vergleich durchgeführt.

Ein vierter Schritt wird in später durchgeführten Experimenten ergänzt. Dieser Schritt erlaubt es einem menschlichen Experten, jede vorgeschlagene Beziehung zu bewerten (Akzeptanz/Ablehnung) und zu klassifizieren (Äquivalenzbeziehung, Assoziationsbeziehung). Bei diesem Schritt wird im Gegensatz zu den anderen drei Stufen kein Algorithmus zum Auffinden von Beziehungen durchgeführt, sondern es findet eine Bewertung und Klassifizierung der vorgeschlagenen Beziehungen statt. Da es sich hier um eine generell andere Aufgabe handelt, unterteilen weiterreichende Ansätze das Vorgehensmodell gemäß diesen Aufgaben in Phasen.

3.2.5.2 Phasenmodelle

In der Informatik hat es sich bewährt, komplexe Prozesse oder Vorgehen in *Phasen* zu unterteilen. In jeder solchen Phase wird eine Teilaufgabe bearbeitet. Beispiele sind Software-Prozess-Modelle aus dem Bereich des Software-Engineering und im Datenbankbereich Phasenmodelle zur Schema- oder Sichtenintegration [BLN86, GLN92, FP93].

In [BK99] werden zwar unterschiedliche Phasen eines Integrationsprozesses dargestellt, der vorgestellte Ansatz bezieht sich aber nur auf eine Analysephase.

Die expliziteste Aufteilung in Phasen wird nach unserer Kenntnis nach in [SC97] vorgenommen. Das dort verwendete Phasenmodell lehnt sich an den Prozess der Sichtenintegration an und sieht folgende Phasen vor:

Prä-Integration: Transfer der zu integrierenden Thesauri in ein gemeinsames Datenmodell und -schema.

Analyse: Auffinden äquivalenter Begriffe in den verschiedenen Thesauri. Dazu werden einfache linguistische Verfahren durchgeführt und Synonyme betrachtet. Durch Berechnen eines semantischen Abstandsmaßes findet eine Bewertung dieser Vorschläge statt, die zu einem Verwerfen oder einer Weiterverwendung führt.

Bestätigung: Auffinden von Konflikten und Konfliktauflösung.

Integration: Erstellen des integrierten Thesaurus durch Erzeugen der Thesaurusvereinigung.

Retrukturierung: Falls erforderlich, Restrukturierung der Thesaurusvereinigung, um sicherzustellen, dass alle Begriffe aller Komponententhesauri enthalten sind (Vollständigkeit) und jeder Begriff durch maximal einen Deskriptor repräsentiert wird (Minimalität).

Ausschließlich die Phasen Analyse und Bestätigung werden in [SC97] näher betrachtet.

Als genereller Mangel sowohl des in [SC97] beschriebenen Vorgehensmodells als auch aller anderen uns bekannten Vorgehensmodelle kann festgestellt werden, dass eine grundsätzliche Teilaufgabe vollständig unberücksichtigt bleibt: die Analyse der Komponententhesauri (vgl. Abschnitt 2.3.2.1). Auch eine Bewertung des Integrationsergebnisses wird in den bekannten Ansätzen – wenn überhaupt – nur sehr oberflächlich behandelt. Mangelnde Felxibilität resultiert aus den starren Integrationsabläufen, die in [SC97] wie auch den anderen Ansätzen eigens und nur für die untersuchten Thesauri angelegt sind.

3.2.5.3 Transformation in ausdrucksstärkere Modelle

Die Semantik der Begriffe eines Thesaurus wird zwar durch Benennungen in Form von Deskriptoren und Nicht-Deskriptoren, durch die Beziehungen zu anderen Begriffen, durch Erläuterungen und Definitionen und eine mögliche Zuordnung der Begriffe zu thematischen Gruppen auf vielfältige Art und Weise ausgedrückt. Dennoch unterstützen diese Möglichkeiten der semantischen Repräsentation die maschinelle Verarbeitung bei einigen Fragestellungen nicht ausreichend:

- Es ist nicht ersichtlich, aus welchem Grund eine bestimmte Relation etabliert wurde. Explizite und überprüfbare Kriterien zur Einordnung von Begriffen fehlen. So stellt z.B. Rector in [Rec99, S. 10] fest, dass Abstraktionsrelationen weder explizit begründet noch – auch resultierend aus dieser Schwäche – konsistent etabliert werden.
- In vielen Thesauri ist nicht ersichtlich, ob eine hierarchische Beziehung eine Abstraktions- oder eine Bestandsrelation ist.
- Aufgrund mangelnder Formalität der Definitionen würde eine Analyse dieser Definitionen die aufwendige Verarbeitung natürlicher Sprache erfordern.

Modelle für Terminologien werden in [RMC97, RMCEM97][2] in drei Klassen eingeteilt:

Systeme der 1. Generation: Einfache hierarchische Strukturen.

Systeme der 2. Generation: Komposite Systeme, die es erlauben, aus elementaren semantischen Einheiten komplexere semantische Einheiten zu konstruieren. Zusätzliche Vermerke sind möglich.

Systeme der 3. Generation: Formale komposite Systeme, die es erlauben, anhand einer gebenen Menge von Symbolen und formalen Manipulationsregeln (Kompositionsbedingungen) aus einfachen semantischen Einheiten komplexe semantische Einheiten zu konstruieren und in einer kanonischen Form zu repräsentieren (Normalisierung). Die Systeme unterstützen zudem eine automatische Klassifikation. Zugrunde liegende Formalismen können z.B. Konzeptuelle Graphen (Conceptual Graphs oder Conceptual Structures) [Sow84, Sow00, Wil97] oder deskriptive Sprachen sein.

Diese Klassifikation beinhaltet einen zunehmend formaleren Grad der Repräsentation der Semantik. Klassische Thesauri erfüllen die Anforderungen der zweiten Generation nur teilweise und müssen daher zwischen der ersten und zweiten Generation eingeordnet werden. Da eine möglichst formal repräsentierte Semantik die maschinelle Verarbeitung der Begriffsbedeutung (im Sinne von Begriffsinhalt) erleichtert, beinhalten manche Vorgehensmodelle in einer Integrations-Vorbereitungsphase einen Transfer von den Systemen erster oder zweiter Generation in Systeme zweiter oder dritter Generation. In dieser Hinsicht am weitesten fortgeschritten ist die ONIONS-Methodologie [GPG99, GPS98, PAG99], die einen Transfer einer Ontologie in eine formale Ontologie (basierend auf einer deskriptiven Sprache) vorsieht und diese formalen Ontologien zur eigentlichen Begriffsintegration verwendet.

Die generelle Frage bei der Betrachtung solcher Systeme lautet, ob diese sehr feingranulare und formale Semantik für einen Thesaurus, der zum Zwecke des Information Retrieval eingesetzt werden soll, Sinn macht oder aber die Anwendung im Wesentlichen auf formale und Computerbasierte Lexika beschränkt bleibt.

Durch eine formal repräsentierte Semantik kann die Qualität hinsichtlich Konsistenz, Vollständigkeit, Exaktheit der Definitionen der Begriffsinhalte, Reorganisations- und Erweiterungsmöglichkeiten eines Thesaurus (oder allgemeiner einer Terminologie) erheblich gesteigert werden. Allerdings konnte selbst innerhalb eines engen und etablierten Fachbereiches (chirurgische Eingriffe in der Medizin), im dem über Bedeutung und Gebrauch der Begriffe vergleichsweise großer Konsens herrscht, nicht immer Einigkeit über eine normalisierte Definition der Begriffe erzielt werden [RRMCZ98]. Bei Thesauri, die von *verschiedenen, autonomen* Erstellern für einen *größeren Anwendungsbereich* (z.B. Umwelt) konstruiert werden, kann mit deutlich unterschiedlichen Definitionen gerechnet werden. Für semantische Vergleiche ist es dann wiederum erforderlich, eine gewisse Unschärfe zuzulassen, so dass die gewonnen Präzision eines sehr formalen Systems in Frage gestellt wird. Zudem unterstützen die bisherigen Systeme solche unscharfen Vergleiche nicht [RRMCZ98]. Weitere Schwierigkeit ist die Vielzahl an Möglichkeiten, die durch eine sehr feingranulare Repräsentation der Semantik entsteht. Einerseits wird dadurch der Effekt verschiedener Definitionen für einen Begriff noch verstärkt, andererseits besteht die Gefahr, dass die Benutzer solcher Systeme von den Möglichkeiten überwältigt werden [RRMCZ98]. Zuletzt sei natürlich auf den großen Aufwand hingewiesen, um aus einem klassischen Thesaurus ein System der zweiten bzw. dritten Generation zu gewinnen. Das Ergebnis wird voraussichtlich auch nicht

[2] s. auch `http://gift.irmkant.rm.cnr.it/secgen.htm`

mit dem Ausgangsthesaurus übereinstimmen, da weitere Beziehungen sowie inkonsistent etablierte Beziehungen, Begriffsüberschneidungen etc. aufgedeckt werden. Der Ausgangsthesaurus würde somit bereits für die Integration nicht mehr unverändert vorliegen.

Unser Fazit ist, dass eine derartige Transformierung den enormen Aufwand zum Zwecke der Begriffsintegration nicht rechtfertigt. Stattdessen sehen wir einzelne Ergebnisse dieser Forschungen als relevant für einen einfacheren Ansatz an, der grundsätzlich eine größere Unschärfe zulässt und den notwendigen intellektuellen Aufwand eines menschlichen Experten darauf beschränkt, dass Integrationsvorschläge, die im Wesentlichen anhand der bereits im Thesaurus vorhandenen Informationen hergeleitet werden können, bewertet werden. Wenn aber die Erstellung eines neuen Thesaurus das Ziel ist, erscheint es uns durchaus untersuchenswert, ob die Konstruktion eines Systems der zweiten oder dritten Generation – aus dem dann ein System der ersten Generation hergeleitet werden kann – die Arbeit erleichtert.

3.3 Bewertung der Güte eines Multi-Thesaurus-Systems

In der überwiegenden Zahl der uns aus der Literatur bekannten Ansätze findet eine Bewertung des Integrationsergebnisses nicht statt. In [SC97] wird dies damit begründet, dass zuvor das Werkzeug zum Zusammenführen der Thesauri entwickelt werden soll. Diese Aussage halten wir gültig für viele weitere Ansätze, aber in der Konsequenz für falsch. Bereits vor der Erstellung des Multi-Thesaurus-Systems ist es erforderlich, Kriterien für eine Bewertung der Güte des Ergebnisses zu haben, damit bei der Entwicklung von Systemen Zwischenstände eingeschätzt und einzelne Verfahren bewertet werden können. Erst mit dieser Bewertung ist eine zielgerichtete Weiterentwicklung möglich.

Bei dokumentenbestandsbasierten Ansätzen zum Auffinden und Klassifizieren von Inter-Thesaurus-Relationen (vgl. Abschnitt 3.2.1.1) bietet es sich an, die Bewertung der Güte eines Multi-Thesaurus-Systems direkt mit einem Effektivitätsmaß für das Information Retrieval vorzunehmen. In [ANOT96] etwa wird als Effektivitätsmaß für die Güte der Abbildung ein gewichtetes harmonisches Mittel aus Recall und Precision [Sal73] verwendet. Eine Validierung findet statt, indem die Menge der gemeinsamen Dokumente in eine Trainings-Menge und eine Validierungs-Menge unterteilt wird.

Die Voraussetzung des Zugriffs auf vollkommen bekannte Dokumentenbestände ist in der Regel aber nicht gegeben. Daher sind Kriterien für die Bewertung von Multi-Thesaurus-Systemen noch zu entwickeln.

3.4 Resümee

In diesem Kapitel wurde untersucht, ob bereits Ansätze existieren, die die von uns aufgestellten Anforderungen erfüllen. Zusammenfassend sind die Ergebnisse in Tabelle 3.2 dargestellt. Die untersuchten Ansätze sind dazu klassifiziert worden. Für jede Klasse werden beispielhaft einige bedeutende Arbeiten genannt.

Als Grundaussage kann festgehalten werden, dass bisher kein Ansatz existiert, der die Integration von Thesauri oder Ontologien ganzheitlich (Analyse und Integration, Anwendung des integrierten Systems) betrachtet.

Insbesondere hat unsere Analyse als wesentliches Problem, das einer guten Skalierbarkeit entgegensteht, ergeben, dass die Vorgehensweise bei der Begriffsintegration entweder zu starr ist

	dokumenten-bestandsbasierte Ansätze	klassische thesaurusbasierte Ansätze	formalisierte thesaurusbasierte Ansätze	Ansätze basierend auf formalen, komplexen Systemen	vergleichende Ansätze	Anfrageübersetzung	Integrationsarchitekturen
	Amba [Amb92,ANOT96]	Rada [Rad87,MR88]; Sneidermann [SB92]; Woodbridge [WSN99]	Sintichtaktis [SC97]	SKC / ONION [MWK99,MWJ99, JMN+99, JW99]	ONIONS [GPG89,GPS08,PAG89]; GALEN [RMC97, RMCGM97,Rec99]	Readi Galen-Kreuzvalidierung [RPR+98]	Observer [Men98,MIKS00]	[3 [AHK+95,Wie96b] Föderierte Datenbanken [Con97]
Informationsmodell								
explizites, formales Informationsmodell für Komponententhesauri	n.a.		+	o (graphbasiertes Modell mit vielen Freiheitsgraden)	o (großer Aufwand für Transformation klassischer Thesauri)	n.a.	o (nur Abstraktionsbeziehungen, Schemabeschreibungsbasiert)	n.a.
explizites, formales Informationsmodell für Multi-Thesaurus-System	n.a.	-	+	o (s.o.)	o (s.o.)	n.a.	+	n.a.
Autonomieerhaltung der Komponententhesauri	n.a.	-	-	+	-	n.a.	+	n.a.
Unterscheidung Indexierung-/Nicht-Indexierungsthesauri	n.a.	-	-	-	-	n.a.	+	n.a.
alle geforderten Inter-Thesaurus-Relationen	n.a.	-	n.a. (Thesaurusvereinigung ohne Intra-Thesaurus-Relationen)	o (s.o.)	o (nur Äquivalenz-beziehung)	n.a.	o (keine Assoziations- und keine Bestands-beziehungen)	n.a.
alle geforderten Begriffstypen	n.a.	o (keine Ergänzenden Begriffe)	+	+	o (keine Ergänzenden Begriffe)	n.a.	o (keine Ergänzenden Begriffe)	n.a.
Invarianten und Konflikte	n.a.	o (keine oder partielle Betrachtung	+ (für Komponenten-thesauri und Thesaurusvereinigung)	-	o (Konfliktauflösung für Zyklen und Polysemie)	n.a.	-	n.a.
Begriffsintegration								
Flexibles Vorgehensmodell	-	-	o (Phasenmodell, nur 2 Phasen werden betrachtet)	-	o (starres, vage definiertes Modell)	-	n.a.	n.a
Unterstützung durch Wissensakquisitionsarchitektur	-	-	-	o (regelbasiertes System, nicht-modular)	-	o (nur Kreuzvalidierung, ohne Berücksichtigung der Ergebnisse für Integration)	n.a.	n.a.
Analyse der Komponententhesauri	-	o	-	-	o (großer Aufwand für Transformation klassischer Thesauri)	o (großer Aufwand für Transformation klassischer Thesauri)	n.a.	n.a.
Vorverarbeitung der Komponententhesauri	-	-	o (Wortstammbildung)	+ (syntaktische Transformationen)	-	-	n.a.	n.a.
Berücksichtigung unterschiedlicher Verwendung	+	-	-	-	-	-	n.a.	n.a.
Kombination unterschiedlicher Integrationsverfahren	-	o (Erweiterungen nur eingeschränkt möglich)	o (Erweiterungen nur eingeschränkt möglich)	o (Erweiterungen nur eingeschränkt möglich)	o (eingeschränkt möglich)	o (Erweiterungen nur eingeschränkt möglich)	n.a.	n.a.
Nutzung externer Wissensquellen	+ (indexierte Dokumentbestände)	-	-	+ (prozedurale Schnittstelle)	-	-	n.a.	n.a.
Ergänzen von Begriffe	-	-	-	o (manuell)	-	-	n.a.	n.a.
Einbringen des Wissens des Integrationsexperten	-	o (keine Werkzeug-unterstützung)	-	+ (Integrationsregeln; einfache Werkzeugunterstützung)	o (keine Werkzeugunterstützung)	-	n.a.	n.a.
Bewertung der Güte des Multi-Thesaurus-Systems								
Evaluierung der iterativen Verbesserung (Güteverbesserung)	n.a.	-	-	-	-	-	n.a.	n.a.
Evaluierung der Erwartungserfüllung (Gesamtgüte)	n.a.	-	-	-	-	-	n.a.	n.a.
Austührungsmaschine								
Überwindung der Entfernung und Heterogenität	n.a.	-	n.a.	n.a.	-	n.a.	-	+
Anfragebearbeitung	n.a.	-	n.a.	n.a.	-	n.a.	+ (eingeschränkt übertragbar)	o (Anfragebearbeitung für verteilte Quellen vorgesehen, Multi-Thesaurus-Systeme nicht berücksichtigt)
Anfrageoptimierung	n.a.	-	n.a.	n.a.	-	n.a.	+ (eingeschränkt übertragbar)	o (s.o.)

Tabelle 3.2: Zusammenfassende Bewertung von Ansätzen zur Thesaurusintegration

und nicht an die unterschiedlichen Randbedingungen angepasst werden kann oder aber beliebige Freiheiten und somit fehlende Führung bei der Integration besitzt. Es fehlt ein flexibleres Vorgehensmodell, das alle Phasen der Analyse und der Integration unter Berücksichtigung der verschiedenen Integrationsverfahren unterstützt. Wichtige Grundlage eines solchen Vorgehensmodells ist die sorgfältige Festlegung der verwendeten Informationsmodelle, um auf definierter semantischer Ebene mit den Komponentensystemen und dem integrierten System umgehen zu können. Zur Unterstützung eines solchen Vorgehensmodells existiert bisher keine Software-Architektur, die die erforderliche Flexibilität hinsichtlich der Einbindung von Integrationsverfahren und des Einbringens von Wissen des menschlichen Integrationsexperten in den Prozess bei gleichzeitiger Modularität und iterativer, flexibler Lösungssuche unterstützt.

Als weiterer wesentlicher Nachteil bisheriger Ansätze wurde festgestellt, dass die Ergebnisse einer Analyse bisher nur in Form von implizitem Expertenwissen in die Begriffsintegration einfließen.

Die Anwendung der integrierten Systeme im Allgemeinen ist bisher wenig untersucht. Bei der Anwendung integrierter Thesauri wird der Benutzer entweder mit vollkommen konsistenten Systemen oder aber mit Systemen mit verborgenen Inkonistenten konfrontiert. Ersteres ist einem Szenario mit autonomen Teilsystemen eine sehr idealisierende Sichtweise. Letzteres ist für den Benutzer eine wenig befriedigende Situation. Zwar wurden bei der Anfrageberbeitung und -optimierung innerhalb des OBSERVER-Ansatz positive Ergebnisse erzielt, doch sind diese aufgrund des an eine Schemabeschreibung angelehnten Datenmodells (jeder Begriff enthält eine Menge von Attributen) nicht direkt auf Thesauri anwendbar. Die Anfrageoptimierung benötigt zudem die Berechnung von Maßen wie Recall und Precision.

Im Rahmen dieser Arbeit wird aufgrund der aufgezeigten Analyseergebnisse ein ganzheitliches Rahmenwerk zur Integration von Thesauri entwickelt. Zur Verbesserung der Skalierbarkeit und der Flexibilität soll dieses Rahmenwerk Lösungen für die beschriebenen Problemstellungen bieten und dabei bewährte Komponenten aus den existierenden Ansätzen einsetzen und flexibel kombinierbar machen.

Kapitel 4

Grundideen des Lösungsansatzes

In Kapitel 2 haben wir das Problem der Integration von Thesauri analysiert und festgestellt, dass die zentralen Herausforderungen die Skalierbarkeit und Flexibilität sind. Die in Kapitel 3 dargestellten Ergebnisse der Untersuchung existierender Ansätze zeigen, dass eine ganzheitliche Betrachtung der Thesaurusintegration (von der Analyse der Komponententhesauri über die Begriffsintegration bis hin zur Anwendung des integrierten Systems) bisher vollständig fehlt. Im Rahmen dieser Arbeit werden wir erstmals einen solchen ganzheitlichen Ansatz bereitstellen, der die Skalierbarkeit und Flexibilität der bisherigen Ansätze signifikant verbessern soll.

In diesem Kapitel wird unser Lösungsansatz in groben Zügen vorgestellt. Wir werden zeigen, welche Bausteine unser Ansatz beinhaltet, um alle Aspekte der Thesaurusintegration zu unterstützen. Zudem werden wir deutlich machen, wie diese Bausteine auf die Ergebnisse existierender Ansätze aufsetzen und diese insbesondere hinsichtlich einer besseren Skalierbarkeit und Flexibilität erweitern.

Die folgenden Kapitel stellen unsere Lösung schließlich detailliert vor. Die Ergebnisse einer exemplarischen Validierung der Teillösungen werden ebenfalls innerhalb der Kapitel dargestellt.

4.1 Aufbau der Lösung

Unser Lösungsansatz zu einer ganzheitlichen Betrachtung einer skalierbaren, flexiblen Thesaurusintegration wird in einer Übersicht in Abbildung 4.1 dargestellt. Er basiert zentral auf der Bereitstellung von Modellen für die semantisch präzise definierte Repräsentation der Komponententhesauri und des Multi-Thesaurus-Systems (Informationsmodelle), einem umfassenden Vorgehensmodell zur Integration von Begriffen in unterschiedlichen Strukturen sowie dessen Umsetzung in eine Wissensakquisitionsarchitektur, Verfahren der Begriffsintegration und schließlich der Unterstützung des Einsatzes des Multi-Thesaurus-Systems zum Zwecke des Information Retrieval (Ausführungsmaschine).

4.2 Bausteine der Lösung

4.2.1 Informationsmodelle

Die Grundlage der semi-automatischen Integration von Thesauri ist die sorgfältige Festlegung der verwendeten Informationsmodelle, um auf definierter *semantischer* Ebene mit den Kom-

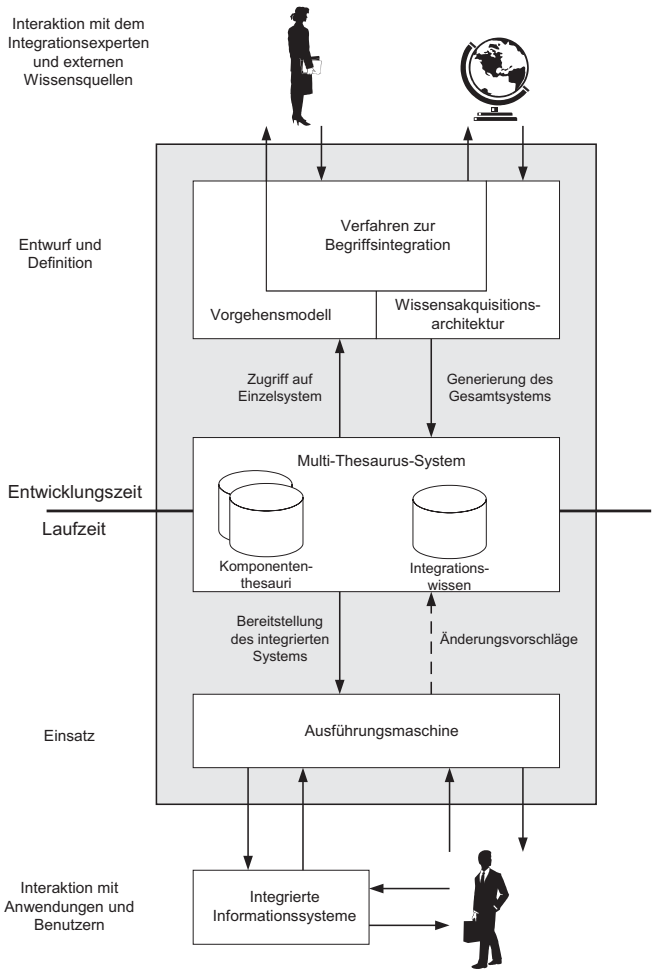

Abbildung 4.1: Aspekte des Lösungsansatzes

ponentensystemen und dem integrierten System umgehen zu können. In Abschnitt 2.3.2.2.3 wurde gezeigt, dass es zwischen den in der Praxis verwendeten Thesaurusmodellen erhebliche Unterschiede gibt und in der Regel von einer präzisen Definition des Modells nicht ausgegangen werden kann. Daher setzen wir bereits an dieser Stelle an und definieren – ausgehend von der in [SC97] dargestellten mengentheoretischen Definition eines Thesaurusmodells – formal ein Informationsmodell für Thesauri, in dem präzise die Eigenschaften des Modells festgelegt werden (vgl. Kapitel 5). Aus diesen Definitionen wird schließlich eine Repräsentation der Thesauri als Graphen hergeleitet, um zu einer intuitiven Form der Darstellung zu gelangen.

Das Informationsmodell für Komponententhesauri wiederum ist Ausgangsbasis für ein Informationsmodell für Thesaurusföderationen (vgl. Kapitel 6), unserer Lösung zur Repräsentation von Multi-Thesaurus-Systemen, die auf lose gekoppelten, autonomen Thesaurussystemen basieren. Um die in Abschnitt 2.3.1 geforderte semantische Reichhaltigkeit, die über die der existierenden Ansätze hinausgeht (vgl. Abschnitt 3.1), zu erreichen, werden Inter-Thesaurus-Relationen und Ergänzende Begriffe sowie eine Unterscheidung zwischen Indexierungs- und Nichtindexierungsthesauri eingeführt. Der gültige Entwurfsraum wird durch die Definition von Invarianten festgelegt und die Behandlung unterschiedlicher Konflikttypen spezifiziert. Da ein Eingriff in die Komponententhesauri deren Autonomie verletzen würde, beinhaltet das Informationsmodell Möglichkeiten zum Markieren von Konflikten und ihrer Ursachen. Eine Behandlung kann dann situationsabhängig zum Zeitpunkt der Benutzeranfrage an das System von der Ausführungsmaschine durchgeführt werden. Das Modell berücksichtigt, dass Komponententhesauri dynamisch ein- und ausgeblendet werden können, so dass dem Benutzer ein „maßgeschneidertes“ Vokabular angeboten werden kann.

Die Informationsmodelle für Komponententhesauri und Thesaurusföderationen sind Grundlage aller Verfahren und Algorithmen, die im Rahmen dieser Arbeit entwickelt werden.

4.2.2 Begriffsintegration

4.2.2.1 Vorgehensmodell

Der Prozess des möglichst vollständigen Auffindens und Etablierens von Inter-Thesaurus-Beziehungen sowie die Konflikterkennung sind wesentliche Aufgabe der Thesaurusintegration auf semantischer Ebene, der Begriffsintegration. Um für die Integrationsexperten in diesem wiederkehrenden Prozess einen Rahmen für ein methodisches Vorgehen mit möglichst weitreichender maschineller Unterstützung, die auf unterschiedlichen Verfahren basiert, bereitzustellen, wird ein Vorgehensmodell benötigt. Anhand der in Abschnitt 2.3 dargestellten Anforderungen können folgende Teilaufgaben des Prozesses der Begriffsintegration hergeleitet werden, die in einem solchen Vorgehensmodell berücksichtigt werden müssen:

1. Die Durchführung vorbereitender Maßnahmen, um die Komponententhesauri in das vom Informationsmodell für Thesaurusföderationen geforderte (vgl. Abschnitt 6.2.1) gemeinsame Informationsmodell zu überführen sowie vergleichbare Benennungen und Definitionen für die Begriffe zu erhalten.

2. Die Analyse der Komponententhesauri, um deren Charakteristika (Stärken und Schwächen) und Kompatibilität festzustellen sowie Integrationserwartungen spezifizieren zu können.

3. Anhand der Ergebnisse der zuvor durchgeführten Schritte sowie der Berücksichtigung der

zur Verfügung stehenden Ressourcen kann die Strategie für das weitere Vorgehen (Integrationsstrategie) geplant bzw. angepasst werden.

4. Die Integrationsstrategie wird schließlich umgesetzt, um potenzielle Inter-Thesaurus-Relationen aufzufinden, zu klassifizieren und zu bewerten, sowie potenzielle Konflikte und Ergänzende Begriffe aufzufinden und zu bewerten. Eine entsprechende Bewertung vorausgesetzt, wird die entsprechende Änderung (Ergänzung oder Revision einer früheren Entscheidung) der Thesaurusföderation etabliert.

   Bei dieser Teilaufgabe sind Interaktionen mit einem menschlichen Experten, der als höchste Entscheidungsinstanz benötigt wird, unbedingt vorzusehen.

5. Eine Bewertung der Güte des Integrationsergebnisses. Diese Bewertung ist die Entscheidungsgrundlage über notwendige weitere Integrationsschritte oder das Beenden des Integrationsprozesses. Des Weiteren sollen Entscheidungen über in weiteren Integrationsschritten einzusetzende Verfahren getroffen werden.

Es gelten weiterhin die generellen Anforderungen der Skalierbarkeit und Flexibilität. Erstere bedeutet als Anforderung für das Vorgehensmodell, sowohl mit komplexen Teilsystemen, in diesem Falle großen Komponententhesauri, umgehen zu können als auch bei der Integration einer größeren Anzahl von Thesauri anwendbar zu sein. Letztere bedeutet als Anforderung, auch bei sehr unterschiedlichen Randbedingungen, d.h. bei inhaltlich und strukturell heterogenen Thesauri mit unterschiedlich stark ausgeprägten Überlappungen, möglichst gute Ergebnisse zu erzielen.

Zur Erfüllung dieser Anforderungen spezifizieren wir ein Vorgehensmodell, das die komplexe Aufgabe der Begriffsintegration in in sich abgeschlossene Teilaufgaben (Phasen) mit geringerem Komplexitätsgrad zerlegt. Durch Ausführen der aufeinander aufbauenden Phasen sowie mögliche Iterationen, die zu einer schrittweisen Verbesserung beitragen, wird schließlich das gesamte Problem der Begriffsintegration gelöst. Solche Phasenmodelle haben sich bei komplexen Prozessen wie dem Software-Engeneering und der Schemaintegration bewährt.

Unser Vorgehensmodell (vgl. Abbildung 4.2) nimmt die anhand der identifizierten Teilaufgaben vorgenommene Prozesszerlegung auf, indem den fünf Teilaufgaben jeweils eine Phase zugeordnet wird.

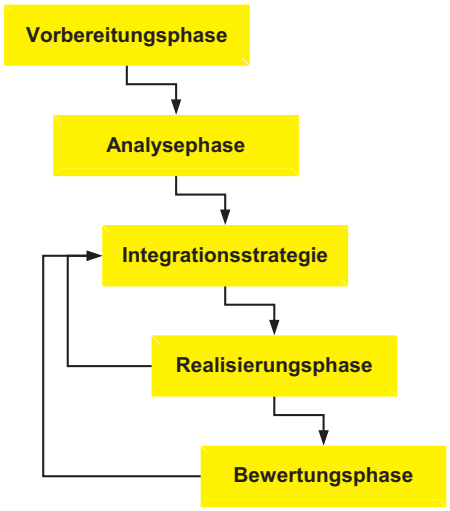

Abbildung 4.2: Die Phasen des Vorgehensmodells

In Abbildung 4.2 ist bereits angedeutet, dass unser Vorgehensmodell Iterationen vorsieht. Dadurch wird eine *schrittweise Verbesserung* des Integrationsergebnisses ermöglicht, wobei die Komplexität einer Iteration innerhalb einer Phase im Vergleich zur Gesamtkomplexität der Phase wiederum verringert werden kann. Bei der Vorstellung der einzelnen Phasen (vgl. Kapitel 8 bis 11) wird ausführlicher dargestellt, wie dieses Potential der schrittweisen Verbesserung ausgenutzt werden kann.

Innerhalb des Vorgehensmodells sind erstmals Phasen vorgesehen, die eine Analyse der Komponententhesauri bzw. des Integrationsergebnisses vorsehen. Die Analyse der Komponententhesauri (vgl. Kapitel 9) ermöglicht es, auch sehr unterschiedliche Thesauri bei der Integration zu berücksichtigen und dabei deren Stärken und Schwächen bei der Konfiguration der Integrationsverfahren zu nutzen. Diese Analyse trägt somit direkt zu einer Verbesserung sowohl der Skalierbarkeit (Umgang mit großen Thesauri, die nicht ohne erheblichen Aufwand manuell analysiert werden können) als auch der Flexibilität (hinsichtlich der beteiligten Komponententhesauri) bei. Die Analyse des Integrationsergebnisses (vgl. Abschnitt 12.3) ermöglicht Aussagen über die Qualität des Ergebnisses als auch über die Verfahren, die zu diesem Ergebnis beigetragen haben.

Die weiteren Phasen (Vorbereitungsphase, Phase der Definition der Integrationsstrategie, Realisierungsphase) sind an Phasen aus Vorgehensmodellen angelehnt, die aus der Literatur bekannt sind. Sie enthalten aber neue Lösungen für Teilaufgaben, z.B. die Anreicherung der Informationsmodelle von Komponententhesauri in der Vorbereitungsphase (vgl. Kapitel 8) zur Herstellung der Konformität mit dem Informationsmodell für Komponententhesauri, die Unterteilung der Informationsstrategie in Subphasen für die Realisierungsphase (vgl. Kapitel 10) zur flexibleren Kombination unterschiedlicher Verfahren und das Speichern von abgelehnten Beziehungen mit der Ablehnungsbegründung in der Realisierungsphase (vgl. Kapitel 11) zur Vermeidung von Wiedervorschlägen während weiterer Iterationen.

4.2.2.2 Architektur

Um das flexible Zusammenspiel der verschiedenen Verfahren phasenübergreifend zu unterstützen sowie die Hinzunahme weiterer Verfahren zu ermöglichen, wird eine blackboardbasierte Wissensakquisitionsarchitektur entwickelt (vgl. Kapitel 7). Die Verfahren werden angewandt, um Integrationshypothesen zu erzeugen und zu bewerten und so einen offenen Ansatz zur Begriffsintegration zu realisieren. Eine Bewertung der Verfahren anhand einer Analyse der Integrationsvorschläge durch qualitätssichernde Verfahren sowie den Integrationsexperten ermöglicht eine dynamische Optimierung.

Durch die flexible Kombination verschiedener Integrationverfahren soll auch bei variierender Ausgangsbasis ein möglichst optimales Integrationsergebnis erreicht werden.

4.2.2.3 Benutzeragent

Die Einbindung eines Experten in den Prozess der Begriffsintegration wird in den verschiedenen Phasen vorgesehen. Als wesentlich werden hierbei die Möglichkeiten einer Einbringung des Expertenwissens in die Phase der Definition der Integrationsstrategie und in die Realisierungsphase angesehen. Unsere Lösung sieht daher einen Benutzeragenten vor, der unterschiedliche Sichten sowohl auf die einzelnen Komponententhesauri ermöglicht als auch vorgeschlagene Integrationsstellen im Kontext anzuzeigen vermag. Aufgrund der darzustellenden Informationsfülle wird als eine graphische Darstellungsform die sehr kompakte Darstellung mittels Fischaugensichten angeboten.

4.2.3 Ausführungsmaschine

Wenn die Begriffsintegration abgeschlossen ist, kann die Thesaurusföderation zur Anfrageformulierung und -erweiterung in Information-Retrieval-Systemen eingesetzt werden. Die von uns entwickelte Ausführungsmaschine bietet durch eine Architektur, die auf der explizit skalierbar gehaltenen I^3-Architektur [Wie94, AHK+95, Wie96] basiert, zum einen die Überwindung der Entfernung und Heterogenität beim Zugriff auf die Komponententhesauri an. Zum anderen beinhaltet sie Komponenten zur Anfragebearbeitung und -optimierung, die eine situationsabhängige Auflösung von Konflikten erlaubt. Hierzu wird auf die Konfliktmarkierungen, die während der Begriffsintegration vorgenommen wurden, zugegriffen sowie auf ein einfaches Benutzerprofil.

Kapitel 5

Informationsmodell für Thesauri

Wie wir bereits gezeigt haben (vgl. Abschnitt 2.3.2.2.3), gibt es zwischen den Thesaurusmodellen, die in der Praxis verwendet werden, erhebliche Unterschiede. In diesem Kapitel soll ein formales Thesaurusmodell vorgestellt werden, dessen Anforderungen die Komponententhesauri mindestens genügen müssen, um an der von uns erarbeiteten Form des Multi-Thesaurus-Systems teilnehmen zu können. Komponententhesauri, die dem in diesem Kapitel vorgestellten Modell nicht entsprechen, müssen nicht automatisch ausgeschlossen werden. Evtl. kann eine entsprechende Informationsanreicherung in einer Vorbereitungsphase dazu beitragen, dass die Anforderungen erfüllt werden.

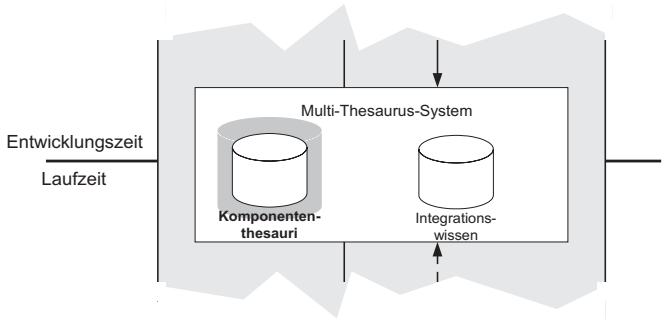

Abbildung 5.1: Modell der Komponententhesauri als Bestandteil der Schnittstelle zwischen Entwicklungs- und Laufzeit

Informal wurde das Thesaurusmodell bereits in Abschnitt 2.1 eingeführt. In diesem Kapitel wird das Modell formal definiert. Abbildung 5.1 zeigt die Bedeutung des Modells als Bestandteil der Schnittstelle zwischen Entwicklungs- und Laufzeit. Die durch die formale Definition gewonnene Präzision ist Voraussetzung der in den Folgekapiteln erarbeiteten Analyse der Komponententhesauri sowie der Verfahren der Begriffsintegration. Das formale Modell wird zusätzlich zu dieser Verwendung während der Entwicklung der Föderation auch zur Laufzeit, also von der Ausführungsmaschine, zum Zugriff auf die Komponententhesauri verwendet.

Alle im weiteren Verlauf der Arbeit vorgestellten Algorithmen basieren auf dem in diesem Kapitel erarbeiteten formalen Informationsmodell.

5.1 Formales Modell monolingualer Thesauri

Ausgangsbasis der in diesem Abschnitt vorgestellten formalen Definitionen sind die in [SC97] gegebenen Definitionen, die hier erweitert und präzisiert werden.

5.1.1 Thesauri

Definition 5.1 *Formal ist ein* monolingualer Thesaurus θ mit Vorzugsbenennungen *definiert als Tupel*

$$\theta = (B, D, D^{def}, D^{erl}, D^{hom}, G, \mathrm{E},\ \mathrm{C},\ \mathrm{A},\ \mathrm{P},\ \mathrm{V})$$

wobei $B = \{b_1, b_2, \cdots, b_k\}$ *die Menge der Benennungen darstellt,* $D = \{d_1, d_2, \cdots, d_l\}$ *die Menge der Deskriptoren oder Vorzugsbenennungen und* $B - D = \{u_1, u_2, \cdots, u_{k-l}\}$ *die Menge der Nicht-Deskriptoren.* $D^{def} = \{p_1, p_2, \cdots, p_m\}$ *ist die Menge der Definitionen,* $D^{erl} = \{q_1, q_2, \cdots, q_n\}$ *ist die Menge der Erläuterungen und* $D^{hom} = \{r_1, r_2, \cdots, r_s\}$ *ist die Menge der Polysem-/Homonymauflösungen.* $G = \{g_1, g_2, \cdots, g_t\}$ *ist die Menge der Gruppen des Thesaurus.* E, C, A, P *und* V *sind die die Äquivalenzrelation, die Benutze-Kombination-Relation, die Abstraktionsrelation, die partitive Hierarchierelation (Bestandsrelation) und die Assoziationsrelation (Verwandtschaftsrelation).*

In den folgenden Abschnitten werden die Elemente von θ näher erläutert.

5.1.2 Begriffe und Benennungen

Definition 5.2 *Jede Benennung* $b \in B$ *ist mit ihrem eindeutigen Identifikator* $\#b$ *durch die Bijektion* $I_\theta : B \to \{1, ..., |B|\} \subseteq \mathbb{N}$ *assoziiert. Das bedeutet* $\forall b \in B : I_\theta(b) = \#b$.

Definition 5.3 *Für jeden Deskriptor* $x \in D$ *ist im Thesaurus jeweils eine Menge von Abstraktionsoberbegriffen* O_x^A*, Bestandsoberbegriffen* O_x^P*, Verwandten Begriffen* R_x *und Gruppen* Q_x *definiert:*

$$\begin{aligned} O_x^A &= \{\#y : y \in D \text{ repräsentiert einen Abstraktionsoberbegriff von } x\} \\ O_x^P &= \{\#y : y \in D \text{ repräsentiert einen Bestandsoberbegriff von } x\} \\ R_x &= \{\#y : y \in D \text{ repräsentiert einen Verwandten Begriff von } x\} \\ Q_x &= \{g : g \in G \text{ ist Gruppe von } x\} \end{aligned}$$

wobei die folgenden Bedingungen

$$\forall x \in D \ : \ \{\#x\} \cap O_x^A = \{\#x\} \cap O_x^P = \{\#x\} \cap R_x = \emptyset \tag{5.1}$$

$$\forall x \in D \ : \ ((O_x^A \neq \emptyset) \vee (O_x^P \neq \emptyset)) \Leftrightarrow Q_x = \emptyset \tag{5.2}$$

gelten. Ein Deskriptor $x \in D$ *wird somit durch ein Tupel*

$$\underline{x} = (\#x, Q_x, O_x^A, O_x^P, R_x)$$

repräsentiert.

Diese Definition sei nun näher erläutert. Die Mengen O_x^A, O_x^P, R_x und Q_x werden von einem oder mehreren Experten festgelegt anhand des folgenden Verständnisses: Ein *Abstraktionsoberbegriff* ist ein direkter oder indirekter Oberbegriff im Sinne eines abstrakteren Begriffes, d.h. der Begriffsinhalt des Oberbegriffes weist mindestens ein Merkmal weniger auf als der Begriffsinhalt seiner direkten oder indirekten Abstraktionsunterbegriffe, wobei der Begriffsinhalt jedes Unterbegriffs eine echte Obermenge des Begriffsinhalts des Oberbegriffs ist (vgl. [DIN93b, DIN92]). Auch Instanzbeziehungen (vgl. S. 16) werden als Abstraktionsbeziehungen betrachtet.

Ein *Bestandsoberbegriff* bezieht sich auf einen Gegenstand[1] als Ganzes und die untergeordneten Begriffe beziehen sich auf die Teile dieses Gegenstandes [DIN93b].

Ein Begriff ist *Verwandter Begriff* eines anderen Begriffs, wenn diese Begriffe in einem Zusammenhang als zusammengehörig gesehen werden und diese Zusammengehörigkeit nicht über die Spezifikation als Abstraktions- oder Bestandsoberbegriff ausgedrückt werden kann. Zudem soll das Ausdrücken des Zusammenhanges zwischen den Begriffen bei Indexierung und Retrieval hilfreich sein (vgl. [DIN87]). Beispiele für Verwandte Begriffe sind gleichgeordnete Begriffe (Begriffe mit dem gleichen Oberbegriff wie Apfel und Birne), nebengeordnete Begriffe (Teile eines gemeinsamen Ganzen, z.B. Baden-Württemberg und Mecklenburg-Vorpommern) oder antonyme Begriffe (Gegensätze, z.B. Hitze, Kälte).

Ein Deskriptor kann eine beliebige Anzahl von Abstraktionsoberbegriffen, Bestandsoberbegriffen bzw. Verwandten Begriffen haben. Allerdings darf ein Deskriptor, wie aus Bedingung 5.1 ersichtlich, nicht sich selbst als Abstraktionsoberbegriff, Bestandsoberbegriff bzw. Verwandten Begriff haben.

Besitzt ein Begriff weder Abstraktions- noch Bestandsoberbegriffe wird er *Topterm* genannnt.

Es wird unten gezeigt, dass über die Mengen O_x^A, O_x^P und R_x die Abstraktionsrelation, die Bestandsrelation und die Assoziationsrelation gebildet werden.

Gruppen spiegeln die Sachgebiete wider, die in dem Thesaurus repräsentiert sind. Sie erlauben es, die Begriffe eines Thesauruses nach einer analytischen Systematik zu ordnen (Beispiele für Gruppen sind „Wirtschaft und Finanzen“ und „Gesundheit und Ernährung“ in GEMET). Die Topterme eines Thesaurus werden je nach Sachgebiet mindestens einer Gruppe zugeordnet. Begriffe, die keine Topterme sind, werden nur indirekt über den oder die zugehörigen Topterme Gruppen zugeordnet. Dies wird aus Eigenschaft 5.2 ersichtlich.

Definition 5.4 *Jeder Deskriptor ist durch die Funktion $J_\theta : D \to D^{def}$ mit einer natürlichsprachigen Definition seiner Bedeutung assoziiert. Diese Definition ist eine textuelle Beschreibung, die entweder leer sein kann oder eindeutig ist.*

Definition 5.5 *Jeder Deskriptor ist durch die Funktion $K_\theta : D \to D^{erl}$ mit einer natürlichsprachigen Erläuterung seiner Verwendung assoziiert. Diese Erläuterung ist eine textuelle Beschreibung, die entweder leer sein kann oder eindeutig ist.*

Definition 5.6 *Jeder Deskriptor ist durch die Funktion $L_\theta : D \to D^{hom}$ mit einer natürlichsprachigen Benennung zur Auflösung von Polysemen und Homonymen assoziiert. Diese Benennung zur Auflösung vom Polysemen oder Homonymen ist eine textuelle Beschreibung, die entweder leer sein kann oder eindeutig ist.*

[1] Ein Gegenstand ist nach [DIN92] ein „beliebiger Ausschnitt aus der wahrnehmbaren oder vorstellbaren Welt“, auch Geschehnisse, Sachverhalte oder Begriffe können Gegenstände sein.

Die drei vorangegangenen Definition erlauben es, zu einem Deskriptor außer der eigentlichen Benennung Definitionen, Erläuterungen und Benennungen zur Auflösung von Polysemen oder Homonymen anzugeben. Nicht jeder Deskriptor muss diese Angaben besitzen, daher wird jeweils ein ausgezeichnetes Element als leeres Element zugelassen, das wir als „“ notieren. Als verkürzende Notation verwenden wir D_x^{def}, D_x^{erl} bzw. D_x^{hom}, um die Definition, Erläuterung bzw. Benennung zur Auflösung von Polysemen oder Homonymen eines Deskriptors $x \in D$ zu bezeichnen.

Definition 5.7 *Für jeden Nicht-Deskriptor $u \in B - D$ wird im Thesaurus entweder genau eine Vorzugsbenennung (einelementige Menge U_u^E) oder eine einzigartige Menge von zu kombinierenden Vorzugsbenennungen U_u^C definiert:*

$$\begin{aligned} U_u^E &= \{\#x : x \in D \textit{ ist eine zu } u \textit{ äquivalente Benennung}\} \\ U_u^C &= \{\#x : x \in D \textit{ ist ein Faktor von } u\} \end{aligned}$$

wobei

$$|U_u^E| \leq 1 \tag{5.3}$$

$$U_u^E \neq \emptyset \Leftrightarrow U_u^C = \emptyset \tag{5.4}$$

$$U_u^E = \emptyset \Leftrightarrow |U_u^C| \geq 2 \tag{5.5}$$

$$\forall u, v \in B - D : \; U_u^C = U_v^C \Rightarrow u = v \tag{5.6}$$

gelten muss.

Auch diese Definition sei näher erläutert: Äquivalenzbeziehungen gibt es ausschließlich zwischen Nicht-Deskriptoren und Deskriptoren. Benennungen sind äquivalent, wenn sie innerhalb des Thesaurus ein- und denselben Begriff bezeichnen. Benennungen werden dabei als äquivalent aufgefasst, wenn es sich um *Synonyme*, also „Bezeichnungen, die im Gebrauch in der fachlichen Kommunikation gleiche Bedeutung haben“ [DIN87], oder um *Quasy-Synonyme*, also „Bezeichnungen, deren Bedeutungen sich zwar in der fachlichen Kommunikation in den einzelnen Aspekten unterscheiden können, die aber für die Zwecke des Dokumentationssystems gleichgesetzt werden“ [DIN87], handelt. Falls U_u^E nicht leer ist, ist der Deskriptor $x \in U_u^E$ der statt u zu verwendende Vorzugsbezeichner.

Falls U_u^C nicht leer ist, ist die *konjunktive Verknüpfung* aller Faktoren $x \in U_u^C$ als äquivalent zu u zu betrachten. Ein *Faktor* ist hier somit ein Begriff (repräsentiert durch einen Deskriptor) der kombiniert mit anderen Faktoren bzw. Begriffen einen neuen zusammengesetzten Begriff darstellt, der nicht durch einen eigenen Deskriptor repräsentiert ist[2].

Wie wir weiter unten zeigen, wird über alle U_u^E die Äquivalenzrelation und über alle U_u^C die Benutze-Kombination-Relation definiert.

Definition 5.8 *Ein Nicht-Deskriptor $u \in B - D$ wird durch ein Tupel*

$$\underline{u} = (\#u, U_u^E, U_u^C)$$

repräsentiert.

[2]Beispiel: Die Faktoren *Luftkühlung* und *Elektromotor* bezeichnen kombiniert den Begriff *luftgekühlter Elektromotor*

Ein *Begriff* ist ein Sachverhalt oder eine Idee selber oder, wie vom Technischen Kommitee 37 (TC37) der ISO in [ISO90] definiert, eine „Einheit des Denkens“. Die Menge der Begriffe wird somit außerhalb des Thesaurus definiert. Innerhalb des Thesaurus werden ausgewählte Begriffe durch Benennungen, also sprachliche Mittel, repräsentiert. Die folgende Definition stellt den Zusammenhang zwischen Begriffen und Benennungen dar.

Definition 5.9 *Ein* Begriff *wird repräsentiert durch eine Teilmenge von B. Es seien* $m \geq 2, n \geq 0$ *und* $x, y_1, \ ..., y_m \in D$ *sowie* $v, u_1, \ ..., u_n \in B - D$. *Dann wird ein Begriff repräsentiert entweder durch* $\{x, u_1, \ ..., u_n\}$, *wobei gilt*

$$u_i \in \{u_1, \ ..., u_n\} \Leftrightarrow U_{u_i}^E = \{\#x\} \tag{5.7}$$

und somit der Begriff durch einen Deskriptor und n Nicht-Deskriptoren dargestellt wird (der Deskriptor ist der eindeutige Vertreter der Äquivalenzklasse), oder der Begriff wird repräsentiert durch $\{v, y_1, \ ..., y_m\}$, *wobei gilt*

$$U_v^C = \{\#y_1, \ ..., \#y_m\} \tag{5.8}$$

und somit wird der Begriff durch mehrere Deskriptoren und einen Nicht-Deskriptor, der in einer Benutze-Kombination-Beziehung zu den Deskriptoren steht, dargestellt.

Im Folgenden unterscheiden wir nicht zwischen den Notationen x und $\underline{x}$ sowie u und $\underline{u}$. Was gemeint ist, ist aus dem Kontext ersichtlich.

5.1.3 Relationen

Im vorangegangenen Abschnitt wurde definiert, wie die Benennungen eines Thesaurus zueinander in Beziehung stehen. In diesem Abschnitt zeigen wir, wie gleichartige Beziehungen zu den zweistelligen Thesaurusrelationen zusammengefasst werden[3].

Ähnlich wie bei WordNet [MBF90] lassen sich die Beziehungen bei Thesauri in syntaktische und semantische Beziehungen einteilen. Syntaktische Beziehungen sind Beziehungen zwischen Bezeichnern, in einem Thesaurus also die Beziehungen zwischen Nicht-Deskriptoren und Deskriptoren. Semantische Beziehungen sind Beziehungen zwischen Bedeutungen oder Begriffen, in einem Thesaurus also die Abstraktions-, die Bestands- und die Assoziationsbeziehungen. Diese semantischen Beziehungen zwischen Begriffen werden ausgedrückt durch Beziehungen zwischen den Begriffsrepräsentanten (s. Definition 5.9). Sind diese Begriffsrepräsentanten ein Nicht-Deskriptor und mehrere Deskriptoren, wird für diesen Begriff keine semantische Beziehung zugelassen.

Eine weitere Ausdifferenzierung der Beziehungstypen ist möglich. Allerdings wird die Verwendung vieler unterschiedlicher Relationen in einem Thesaurus allgemein kritisch beurteilt (s. z.B. [Vie97, S. 56] und [Sch88]). Als Voraussetzung für einen Komponententhesaurus beschränken wir uns daher auf die oben genannten Beziehungstypen.

5.1.3.1 Äquivalenzrelation

Definition 5.10 *Die* Äquivalenzrelation *ist nach DIN-Norm die Relation, die gleichwertige Benennungen (bedeutungsgleich oder bedeutungsähnlich) zu einer Äquivalenzklasse zusammenführt*

[3] Es sei angemerkt, dass in der Literatur und insbesondere auch in den DIN-Normen häufig nicht zwischen den Begriffen *Beziehung* (engl. relationship) und *Relation* (engl. relation) unterschieden wird. Wir treffen diese in der Mathematik gängige Unterscheidung, um möglichst präzise Definitionen bereitstellen zu können.

[DIN87, S. 5]. Formal ist die Äquivalenzrelation E *eine Teilmenge von* $(B-D)\times D$*, die folgende Bedingung erfüllt:*

$$\mathrm{E} = \{(u,x) : u \in B - D \text{ und } \#x \in U_u^E\} \tag{5.9}$$

Bemerkung 5.1 *Im streng mathematischen Sinne ist die Äquivalenzrelation* E *keine Äquivalenzrelation, da sie weder symmetrisch noch reflexiv ist und es aufgrund der disjunkten Grundmengen auch nicht sein kann. Dieser anscheinende Widerspruch ist darauf zurückzuführen, dass unter den äquivalenten Benennungen eine Vorzugsbenennung als eindeutiger Identifikator der Äquivalenzklasse bestimmt wird. Würde auf eine solche Bevorzugung verzichtet werden, könnte die Äquivalenzrelation symmetrisch und auch reflexiv definiert werden.*

5.1.3.2 Benutze-Kombination-Relation

Definition 5.11 *Die* Benutze-Kombination-Relation *(auch 1-zu-n-Äquivalenzrelation genannt) kann verwendet werden, um die Anzahl der Deskriptoren überschaubar zu halten, indem „Begriffe [...] durch die [...] Kombination bereits vorhandener Deskriptoren dargestellt werden" [DIN87, S. 3]. Formal ist die Benutze-Kombination-Relation* C *eine Teilmenge von* $(B-D)\times D$*, die die folgende Bedingung erfüllt:*

$$\mathrm{C} = \{(u,x) : u \in B - D \text{ und } \#x \in U_u^C\} \tag{5.10}$$

Anhand der Definition von U_u^C wird ersichtlich (vgl. Definition 5.7), dass die Benutze-Kombination-Relation 1:n-Beziehungen beinhaltet.

C^{-1} sei die inverse Relation von C.

5.1.3.3 Abstraktionsrelation

Definition 5.12 *Die* Abstraktionsrelation *fasst gerichtete Beziehungen zusammen, die die Über- bzw. Unterordnung der Begriffe im Sinne des allgemeineren bzw. spezifischeren Begriffs darstellen. Dabei besitzt der untergeordnete Begriff (Abstraktionsunterbegriff) alle Merkmale des übergeordneten Begriffs (Abstraktionsoberbegriff) und zusätzlich mindestens ein weiteres spezifizierendes Merkmal [DIN87].*

Formal ist die Abstraktionsrelation A *die kleinste Teilmenge von* $D \times D$ *für die gilt*[4]*:*

$$\mathrm{A} = \{(x,y) : x \in D \text{ und } \#y \in O_x^A\} \tag{5.11}$$

Aufgrund der Definition der Abstraktionsober- bzw. -unterbegriffsbeziehung ist die Abstraktionsrelation A *transitiv, d.h. es gilt:*

$$\forall x,y,z \in D : (x,y) \in \mathrm{A} \text{ und } (y,z) \in \mathrm{A} \Rightarrow (x,z) \in \mathrm{A} \tag{5.12}$$

Die Abstraktionsrelation ist aufgrund Bedingung 5.1 nicht reflexiv. Eine weitere Eigenschaft der Abstraktionsrelation, die im mathematischen Sinne eine Ordnungsrelation ist, ist die Zyklenfreiheit. Ein Pfad in der Abstraktionsrelation von Deskriptor x_1 *zu Deskriptor* x_l *ist eine Folge von Deskriptoren* $x_1, x_2, \ldots, x_l$*, so dass* $l \geq 2$ *und* $\forall i \in [1, l-1] : \exists (x_{i+1}, x_i) \in \mathrm{A}$*. Ein solcher Pfad*

[4] Der erste Parameter ist untergeordneter Begriff des zweiten Begriffs.

wird als $x_1 \overset{\mathrm{A}}{\hookrightarrow} x_l$ *notiert. Die Eigenschaft der* Zyklenfreiheit der Abstraktionsrelation *kann also wie folgt notiert werden:*

$$\forall x, y \in D: \; x \overset{\mathrm{A}}{\hookrightarrow} y \Rightarrow \#x \neq \#y \tag{5.13}$$

Die DIN-Norm [DIN87] gibt keine Beschränkung für die mögliche Anzahl über- bzw. untergeordneter Abstraktionsbegriffe eines Begriffes vor. Ein Begriff, der keinen in der Abstraktionsrelation übergeordneten Begriff besitzt, wird *Topterm der Abstraktionsrelation* genannt.

A^{-1} sei die inverse Relation von A. Die *transitive Hülle* von A bzw. A^{-1} wird notiert als

$$\begin{aligned} \mathrm{A}_x^+ &= \{y \in D : (x, y) \in \mathrm{A}\} \\ (\mathrm{A}_x^{-1})^+ &= \{y \in D : (y, x) \in \mathrm{A}\} \end{aligned}$$

5.1.3.4 Bestandsrelation

Definition 5.13 *Die* Bestandsrelation (partitive Relation) *ist eine Relation, die gerichtete Beziehungen zusammenfasst, die die Über- bzw. Unterordnung der Begriffe im Sinne des Ganzen bzw. eines Teiles darstellen. Dabei entspricht der untergeordnete Begriff (Teilbegriff) einem der Bestandteile des übergeordneten Begriffs (Verbandsbegriff), der ein Ganzes darstellt [DIN87].*

Formal ist die Bestandsrelation P *die kleinste Teilmenge von* $D \times D$ *für die gilt*[5]*:*

$$\mathrm{P} = \{(x, y) : x \in D \text{ und } \#y \in O_x^P\} \tag{5.14}$$

Die Bestandsrelation ist ebenfalls aufgrund Bedingung 5.1 nicht reflexiv. Für die Bestandsrelation gilt wie für die Abstraktionsrelation die Eigenschaft der Zyklenfreiheit. Ein Pfad in der Bestandsrelation von Deskriptor x_1 *zu Deskriptor* x_l *ist eine Folge von Deskriptoren* $x_1, x_2, \ldots, x_l$ *so dass* $l \geq 2$ *und* $\forall i \in [1, l-1] : \exists (x_{i+1}, x_i) \in \mathrm{P}$*. Ein solcher Pfad wird als* $x_1 \overset{\mathrm{P}}{\hookrightarrow} x_l$ *notiert. Die Eigenschaft der* Zyklenfreiheit der Bestandsrelation *kann also wie folgt notiert werden:*

$$\forall x, y \in D: \; x \overset{\mathrm{P}}{\hookrightarrow} y \Rightarrow \#x \neq \#y \tag{5.15}$$

Die DIN-Norm [DIN87] gibt auch für die mögliche Anzahl über- bzw. untergeordneter Bestandsbegriffe eines Begriffes keine Beschränkung vor. Die Bestandsrelation kann in weitere Untertypen unterschieden werden, z.B. in eine Funktionale-Komponente-Relation (administration (Verwaltung) - administration procedure (Verwaltungsvorgang) - circular mail (Rundbrief)) oder in eine Segmentiertes-Ganzes-Relation (forest (Wald) - tree (Baum) - leaf (Blatt)), von denen manche als transitiv angesehen werden, andere nicht (s. z.B. [Sch94]). Da wir nicht zwischen Untertypen der Bestandsrelation unterscheiden, gilt für die Bestandsrelation die Eigenschaft der Transitivität nicht.

P^{-1} sei die inverse Relation von P.

5.1.3.5 Hierarchierelation

Abstraktions- und Bestandsbeziehungen werden in vielen Thesauri zu einer einzigen Relation zusammengefasst, der Hierarchierelation. Aufgrund der grundsätzlich unterschiedlichen Semantik dieser Beziehungstypen erschwert dies aber eine maschinelle Interpretation und Integration

[5] Der erste Parameter ist untergeordneter Begriff des zweiten Begriffs.

von Strukturen aus verschiedenen Thesauri erheblich. Daher fordert unser Informationsmodell zwei verschiedene Relationen. Dennoch ist bei einigen Analysen die gemeinsame Betrachtung der Abstraktions- und Bestandsbeziehungen sinnvoll. Daher führen wir an dieser Stelle die Hierarchierelation als Vereinigung der Abstraktions- und Bestandsrelation ein.

Definition 5.14 *Hierarchiebeziehungen liegen vor, wenn zwei Begriffe zueinander in einem Verhältnis der Über- bzw. Unterordnung stehen, unabhängig davon, ob es sich um eine Abstraktionsbeziehung oder eine Bestandsbeziehung handelt. Die* Hierarchierelation H *ist somit die Vereinigung der Abstraktionsrelation* A *und der Bestandsrelation* P*, für die gilt:*

$$\mathrm{H} = \{(x, y) : (x, y) \in \mathrm{A} \text{ oder } (x, y) \in \mathrm{P}\} \tag{5.16}$$

Für die Hierarchierelation muss wie für die diese Relation definierenden Relationen Abstraktionsrelation und Bestandsrelation die Eigenschaft der Zyklenfreiheit gefordert werden. Ein Pfad in der Hierarchierelation von Deskriptor x_1 zu Deskriptor x_l ist eine Folge von Deskriptoren $x_1, x_2, \ldots, x_l$ so dass $l \geq 2$ und $\forall i \in [1, l-1] : \exists (x_{i+1}, x_i) \in \mathrm{H}$. Ein solcher Pfad wird als $x_1 \hookrightarrow x_l$ notiert. Die Eigenschaft der Zyklenfreiheit der Hierarchierelation *kann also wie folgt notiert werden:*

$$\forall x, y \in D : \; x \hookrightarrow y \Rightarrow \#x \neq \#y \tag{5.17}$$

Bemerkung 5.2 *Die Forderung der Zyklenfreiheit der Hierarchierelation ist stärker als die jeweilige Forderung nach Zyklenfreiheit der Abstraktionsrelation und Bestandsrelation. Es werden somit weitere Einschränkungen bezüglich möglicher Abstraktions- und Bestandsbeziehungen getroffen.*

Die Hierarchierelation ist nicht transitiv. Dies wird sofort anhand der fehlenden Transitivität der Bestandsrelation ersichtlich[6].

Existieren in einem Thesaurus ausschließlich Begriffe mit maximal einem übergeordneten Begriff, wird er *(mono-)hierarchischer Thesaurus* genannt, ansonsten *polyhierarchischer Thesaurus*. Allgemein wird gefordert, dass ein Begriff *eine* Bedeutung hat, die unabhängig vom Pfad ist, über den er erreicht wurde (s. z.B. die beiden Desiderata „poly-hierarchy“ und „concept permanence“ in den häufig zitierten „Desiderata for Controlled Medical Vocabularies“ [Cim98], die in ihrer zentralen Aussage unabhängig von medizinischen Terminologien gelten).

H^{-1} sei die inverse Relation von H.

5.1.3.6 Assoziationsrelation

Definition 5.15 *Die* Assoziationsrelation *umfasst vielfältige Beziehungen zwischen Begriffen, die als wichtig erscheinen, aber weder eindeutig hierarchischer Natur noch äquivalent sind [DIN87]. Formal ist die Assoziationsrelation* V *die kleinste Teilmenge von $D \times D$, für die gilt:*

$$\mathrm{V} = \{(x, y) : x \in D \text{ und } \#y \in R_x\} \tag{5.18}$$

Die Assoziationsrelation ist aufgrund der Symmetrie der Verwandtschaftsbeziehung zwischen zwei Begriffen symmetrisch. D.h. es gilt:

$$\forall x, y \in D : (x, y) \in \mathrm{V} \Rightarrow (y, x) \in \mathrm{V} \tag{5.19}$$

[6] Selbst wenn die Bestandsrelation transitiv wäre, könnte nicht auf die Transitivität der Hierarchierelation als Vereinigung zweier transitiver Relationen geschlossen werden.

5.1.3.7 Paarweise Disjunktheit der Relationen

Während der Definition der Benennungen, Begriffe und Relationen wurden bereits eine Reihe von Eigenschaften eingeführt. An dieser Stelle wird eine weitere, relationsübergreifende Eigenschaft gefordert.

Die *paarweise Disjunktheit der Relationen* fordert, dass die Relationen E, C, A, A^{-1}, P, P^{-1} und V paarweise disjunkt sein sollen. Das bedeutet, dass zwei Benennungen, die durch eine Relation miteinander in Beziehung stehen, durch keine andere Relation miteinander in Beziehung stehen sollen. Die Schnittmenge von E bzw. C mit den anderen Relationen ist nach den Definitionen 5.10 (Äquivalenzrelation), 5.11 (Benutze-Kombination-Relation), 5.12 (Abstraktionsrelation) und 5.13 (Bestandsrelation) offensichtlich immer leer; die Schnittmenge von E und C ist aufgrund der Eigenschaften 5.4 und 5.5 immer leer; ebenso sind die Schnittmengen von A, A^{-1} bzw. P, P^{-1} aufgrund der Zyklenfreiheit der Hierarchierelation (s. Eigenschaft 5.17) leer. Zusätzlich muss – unter Berücksichtigung der Symmetrie von V – also gelten:

$$P \cap A = P \cap A^{-1} = P \cap V = A \cap V = \emptyset \tag{5.20}$$

Da A^{-1} und P^{-1} jeweils die zu A und P inversen Relationen sind, sind die weiteren Schnittmengen mit inversen Relationen identisch mit bereits aufgeführten Schnittmengen und brauchen daher nicht zusätzlich gebildet zu werden.

5.2 Beschreibung von Thesauri als Graphen

Das im vorangegangenen Abschnitt beschriebene Thesaurusmodell kann als gerichteter Graph mit beschrifteten Kanten veranschaulicht werden. Diese Form der Darstellung ist zum einen sehr intuitiv, zum anderen bietet die Graphenrepräsentation eine hervorragende Ausgangsbasis, sowohl zu einer Auswertung der Strukturen als auch zur Formulierung von Algorithmen und Operatoren über den Graphen.

Als Ausgangsbasis zur Herstellung von Inter-Ontologie-Beziehungen wird auch in [MWJ99] ein graphenorientiertes Modell gewählt. Allerdings bleiben dort die Freiheitsgrade zur Ausgestaltung des Graphen hinsichtlich der Kantenbeschriftungen (Kantentypen) sowie der Verbindungen von Knoten durch Kanten sehr groß. Da wir es als erforderlich betrachten, bereits möglichst weitgehende Aussagen über die Semantik der Knoten und Kanten und die Qualität der Graphenstrukturen machen zu können, schränken wir die Freiheitsgrade soweit als möglich ein. Der Vorteil ist, dass wir für solche Thesauri, die im praktischen Einsatz in Information-Retrieval-Systemen überwiegend anzutreffen sind, bereits optimierte Verfahren zur Analyse und Integration anbieten können. Erfüllt ein Komponententhesaurus nicht die Ansprüche unseres Modells, ist eine entsprechende Vorverarbeitung vorzusehen.

5.2.1 Knoten und Kanten

Definition 5.16 *Ein Thesaurus, der dem in Abschnitt 5.1 dargestellten Modell entspricht, kann als gerichteter Graph $T = (N, E)$ dargestellt werden. $N = \{\#\theta\} \cup G \cup B$ ist eine endliche Menge von Knoten des Graphen, wobei $\#\theta \in \mathbb{N}$ der eindeutige Identifikator des Thesaurus ist, G die Menge der Gruppen und B die Menge der Benennungen. E ist eine endliche Menge beschrifteter, gerichteter Kanten.*

Jeder Knoten $n \in N$ ist mit seinem eindeutigen Identifikator $\#n$ durch die Bijektion $I_T : N \rightarrow \{1, ..., |N|\} \subseteq \mathbb{N}$ assoziiert. Das bedeutet $\forall n \in N : I_T(n) = \#n$.

Für jedes $n \in N$ liefert eine Funktion $\lambda(n)$ ein Tupel $(typ,\ \#n,\ s)$ wobei

$$typ = \begin{cases} \textit{Thesaurus}, & \textit{falls } n = \#\theta \\ \textit{Gruppe}, & \textit{falls } n \in G \\ \textit{Deskriptor}, & \textit{falls } n \in D \\ \textit{Äquivalenz-Nicht-Deskriptor}, & \textit{falls } n \in B - D, U_n^E \neq \emptyset \\ \textit{Kombinations-Nicht-Deskriptor}, & \textit{falls } n \in B - D, U_n^C \neq \emptyset \end{cases}$$

ist. s ist der Name des Knotens, also eine Zeichenkette.

Eine Kante $e \in E$ wird notiert als (n_1, α, n_2), wobei $n_1, n_2 \in N$ und α die Beschriftung der Kante ist. Die Funktion $\delta(e)$ liefert die Beschriftung α einer Kante, wobei gilt $\alpha \in \{hatGruppe, hatTopterm, hatBKTopterm, hatElement, BS, BK, UA, UP, VB\}$.

VB ist eine ungerichtete Kante, alle anderen Kanten sind gerichtet. Sie können auch gegen die Richtung gelesen werden. Die impliziten Kantenbeschriftungen, die wir der besseren Lesbarkeit ebenfalls verwenden, entprechen dann $istGruppeVon, istToptermIn, istBKToptermIn, istElementVon, BF, KB, OA, OP$.

Für den Typ, den Identifikator und den Namen eines Knotens notieren wir kurz $\lambda_n^{typ}, \lambda_n^{\#}, \lambda_n^{s}$. Eine Kante z.B. des Typs BK nennen wir abgekürzt auch BK-Kante. Die zweibuchstabigen Beschriftungen entsprechen mehrheitlich den in [DIN87] aufgeführten deutschen Kurzzeichen (BS: Benutze Synonym, BF: Benutzt für Synonym, BK: Benutze Kombination, KB: Benutzt in Kombination, OA: Oberbegriff der Abstraktionsrelation, UA: Unterbegriff der Abstraktionsrelation, VB: Verwandter Begriff). Nur bei OP (Oberbegriff der Bestandsrelation) und UP (Unterbegriff der Bestandsrelation) wurde davon abgewichen, um eine einfachere Lesbarkeit zu erzielen.

Die Menge E der Kanten ist nicht zu verwechseln mit der Äquivalenzrelation, die ebenfalls mit E denotiert wird. Die Bedeutung ist anhand des Kontextes immer ersichtlich. Die Kanten $e \in E$ werden in den folgenden Definitionen spezifiziert. Illustrierend wird ein so definierter Thesaurus in Abbildung 5.2 gezeigt.

Definition 5.17 *Der Knoten $\#\theta$ mit dem Thesaurusidentifikator ist der Wurzelknoten des Thesaurusgraphen T. Vom Wurzelknoten gibt es zu jedem Gruppenknoten $g \in G$ genau eine Kante des Typs hatGruppe, zu jedem Topterm $d \in D$ genau eine Kante des Typs hatTopterm und zu jedem Nicht-Deskriptorknoten $u \in B - D$, der in einer Benutze-Kombination-Beziehung steht, genau eine Kante des Typs hatBKTopterm. Weitere Kanten gehen vom Wurzelknoten nicht aus. Das bedeutet*

$$(\#\theta, hatGruppe, g) \in E \quad \Leftrightarrow \quad g \in G \tag{5.21}$$

$$(\#\theta, hatTopterm, d) \in E \quad \Leftrightarrow \quad d \in D, O_d^A = O_d^P = \emptyset \tag{5.22}$$

$$(\#\theta, hatBKTopterm, u) \in E \quad \Leftrightarrow \quad u \in B - D, |U_u^C| \geq 2 \tag{5.23}$$

In der vorangegangenen Definition 5.17 werden alle Kanten definiert, die vom Wurzelknoten ausgehen. In der folgenden Definition 5.18 werden die Kanten definiert, die von einem Knoten des Typs *Gruppe* ausgehen.

Definition 5.18 *Von allen Gruppenknoten $g \in G$ gibt es eine Kante des Typs hatElement zu jedem Deskriptorknoten $d \in D$, für den $g \in Q_d$ gilt. Das bedeutet*

$$(g, hatElement, d) \in E \quad \Leftrightarrow \quad g \in Q_d \tag{5.24}$$

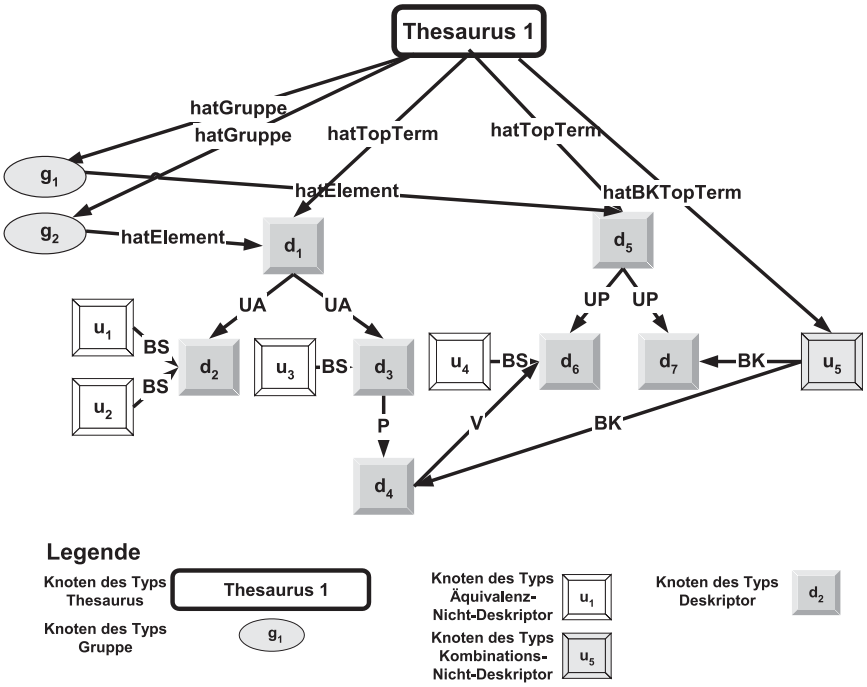

Abbildung 5.2: Graphische Darstellung eines Thesaurusgraphen

Aufgrund Bedingung 5.2 ist jeder Deskriptorknoten, auf den von einer Gruppe verwiesen wird, Topterm.

Die folgenden Definitionen 5.19 bis 5.21 definieren die Kanten, die von einem Deskriptor ausgehen.

Definition 5.19 *Von einem Deskriptor $d_1 \in D$ geht eine Kante des Typs UA zu einem anderen Deskriptor $d_2 \in D$, wenn d_1 ein Abstraktionsoberbegriff von d_2 ist und die Kante nicht aufgrund der Transitivität der Abstraktionsrelation redundant ist. Das bedeutet*

$$(d_1, UA, d_2) \in E \quad \Leftrightarrow \quad (d_2, d_1) \in A, \nexists d_3 \in A_{d_1}^{+} : \#d_2 \in O_{d_3}^{A} \tag{5.25}$$

Definition 5.20 *Wenn ein Deskriptor $d_1 \in D$ Bestandsoberbegriff von einem anderen Deskriptor $d_2 \in D$ ist, geht eine Kante des Typs UP von d_1 nach d_2. Das bedeutet*

$$(d_1, UP, d_2) \in E \quad \Leftrightarrow \quad (d_2, d_1) \in \mathrm{P} \tag{5.26}$$

Definition 5.21 *Wenn ein Deskriptor $d_1 \in D$ assoziierter Begriff von einem anderen Deskriptor $d_2 \in D$ ist, geht eine Kante des Typs VB von d_1 nach d_2. Das bedeutet*

$$(d_1, VB, d_2) \in E \quad \Leftrightarrow \quad (d_1, d_2) \in \mathrm{V} \wedge \#d_1 < \#d_2 \tag{5.27}$$

Die Zusatzbedingung $\#d_1 < \#d_2$ ist erforderlich, da aufgrund der Definition 5.15 der Assoziationsrelation V ansonsten ebenfalls eine Kante des Typs VB in die umgekehrte Richtung existieren würde. VB-Kanten sind die einzigen Kanten, die ungerichtet sind.

Die Definitionen 5.22 und 5.23 definieren Kanten, die von Nicht-Deskriptorknoten ausgehen.

Definition 5.22 *Von einem Nicht-Deskriptor $u \in B - D$ geht eine Kante des Typs BS zu einem Deskriptor $d \in D$ genau dann, wenn es eine Äquivalenzbeziehung zwischen u und d gibt:*

$$(u, BS, d) \in E \quad \Leftrightarrow \quad (u, d) \in \mathrm{E} \tag{5.28}$$

Definition 5.23 *Von einem Nicht-Deskriptor $u \in B - D$ geht eine Kante des Typs BK zu einem Deskriptor $d \in D$ genau dann, wenn es eine Benutze-Kombination-Beziehung zwischen u und d gibt:*

$$(u, BK, d) \in E \quad \Leftrightarrow \quad (u, d) \in \mathrm{C} \tag{5.29}$$

Bemerkung 5.3 *Durch die Verwendung des Knotens mit dem Thesaurusidentifikator als Wurzelknoten wird der Thesaurusgraph ein zusammenhängender, gerichteter Graph. Von dem Wurzelknoten aus sind bis auf die Nicht-Deskriptorknoten alle anderen Knoten des Graphen erreichbar. Um die Nicht-Deskriptorknoten zu erreichen, müssen die BS- und BK-Kanten in umgekehrte Richtung gelesen werden dürfen.*

5.2.2 Pfade

Definition 5.24 *Ein* Pfad *p in einem Thesaurusgraphen T = (N, E) von einem Knoten n_1 zu einem Knoten n_l ($l \geq 2$) ist definiert als $p = n_1.e_1.n_2.e_2...e_{l-1}.n_l$, wobei*

$$\forall i \in [1, l-1] : \exists (n_i, e_i, n_{i+1}) \in E \tag{5.30}$$

Ein solcher Pfad wird auch notiert als $n_1 \hookrightarrow n_l$.

Definition 5.25 *Ein wie in 5.24 definierter Pfad heißt* Abstraktionspfad, *wenn zusätzlich gilt*

$$\forall i \in [1, l-1] : \delta(e_i) = UA \tag{5.31}$$

Ein Abstraktionspfad wird auch notiert als $n_1 \overset{UA}{\hookrightarrow} n_l$.

Definition 5.26 *Ein wie in 5.24 definierter Pfad heißt* Bestandspfad, *wenn zusätzlich gilt*

$$\forall i \in [1, l-1] : \delta(e_i) = UP \tag{5.32}$$

Ein Bestandspfad wird auch notiert als $n_1 \overset{UP}{\hookrightarrow} n_l$.

Definition 5.27 *Ein wie in 5.24 definierter Pfad heißt* Hierarchiepfad *oder* Hierarchischer Pfad, *wenn zusätzlich gilt*

$$\forall i \in [1, l-1] : \delta(e_i) \in \{UA, UP\} \tag{5.33}$$

Ein Hierarchiepfad wird auch notiert als $n_1 \overset{U}{\hookrightarrow} n_l$.

Bemerkung 5.4 *Aufgrund möglicher Polyhierarchie müssen Abstraktions-, Bestands- und Hierarchiepfade nicht eindeutig sein. Das bedeutet beispielsweise, dass zwei verschiedene Abstraktionspfade gleicher oder auch unterschiedlicher Länge zwischen zwei Knoten existieren können.*

5.2.3 Invarianten

Für einen Thesaurusgraphen gelten eine ganze Reihe von Invarianten, die anhand der Relationen, auf denen sie basieren, hergeleitet werden können (s. auch [SC97, S. 132] und [Vie97, S. 68f]). An dieser Stelle werden sie explizit aufgeführt, um als Überblick für ein späteres Prüfen der Einhaltung der Invarianten eines zu analysierenden Thesaurus zu dienen.

5.2.3.1 Keine Selbstverweise

Eine Kante, die von einem Knoten $n \in N$ ausgeht, darf nicht wiederum zu dem gleichen Knoten n führen. Diese wird für Deskriptoren und die über diesen definierten Relationen (und somit den Kanten zwischen diesen) durch Bedingung 5.1 gefordert, für Nicht-Deskriptoren anhand der Definitionen 5.10 und 5.11 der Relationen über diese. Für Gruppen und den Thesaurusidentifikator $\#\theta$ gilt dies offensichtlich, da es keine Kante von einem Knoten solchen Typs zu einem Knoten gleichen Typs gibt.

5.2.3.2 Einzigartigkeit einer Kante

Zwei Knoten dürfen in eine Richtung ausschließlich durch eine einzige Kante verbunden sein. Dies gilt aufgrund der paarweisen Disjunktheit der Relationen (vgl. Abschnitt 5.1.3.7) und der Definitionen der Kanten, die von Gruppen und dem Thesaurusnamen ausgehen (Definitionen 5.17 und 5.18).

5.2.3.3 Zyklenfreiheit der Abstraktionspfade

Es darf keinen Abstraktionspfad geben, der einen Zyklus enthält. Das bedeutet

$$n \overset{UA}{\hookrightarrow} m \Rightarrow n \neq m \tag{5.34}$$

Die Zyklenfreiheit der Abstraktionspfade gilt aufgrund der in Bedingung 5.13 geforderten Zyklenfreiheit der Abstraktionsrelation.

5.2.3.4 Zyklenfreiheit der Bestandspfade

Es darf keinen Bestandspfad geben, der einen Zyklus enthält. Das bedeutet

$$n \overset{UP}{\hookrightarrow} m \Rightarrow n \neq m \tag{5.35}$$

Die Zyklenfreiheit der Bestandspfade gilt aufgrund der in Bedingung 5.15 geforderten Zyklenfreiheit der Bestandsrelation.

5.2.3.5 Zyklenfreiheit der Hierarchiepfade

Es darf keinen Hierarchiepfad geben, der einen Zyklus enthält. Das bedeutet

$$n \overset{U}{\hookrightarrow} m \Rightarrow n \neq m \tag{5.36}$$

Die Zyklenfreiheit der Hierarchiepfade gilt aufgrund der in Bedingung 5.17 geforderten Zyklenfreiheit der Hierarchierelation.

5.2.3.6 Redundanzfreiheit der Abstraktionspfade

Durch die Transitivität der Abstraktionsrelation implizierte Kanten dürfen nicht redundant ausgedrückt werden. Dies gilt aufgrund Definition 5.19.

5.2.3.7 Verbundenheit der Nicht-Deskriptoren

Es darf keinen Nicht-Deskriptor geben, von dem keine Kante des Typs BS oder des Typs BK ausgeht. Das bedeutet, dass Nicht-Deskriptoren nur in Abhängigkeit der Dekriptoren existieren können, also jeder Nicht-Deskriptor durch die Äquivalenzrelation oder die Benutze-Kombination-Relation mindestens einem Deskriptor zugeordnet ist. Die Verbundenheit der Nicht-Deskriptoren gilt aufgrund der Definition 5.7 und hier insbesondere aufgrund der Eigenschaften 5.3 bis 5.4.

5.2.3.8 Einzigartigkeit einer Menge von BK-Kanten

Die Menge der Deskriptoren, auf die ein Nicht-Deskriptor per BK-Kanten verweist, muss einzigartig sein. Das bedeutet, wenn u, v Nicht-Deskriptorknoten und $D_u = \{d : (u, BK, d) \in E\}$, $D_v = \{d : (v, BK, d) \in E\}$ definiert sind, dann muss gelten

$$(D_u \neq \emptyset \neq D_v) \wedge (D_u = D_v) \Rightarrow u = v \tag{5.37}$$

Die Einzigartigkeit der Menge von BK-Kanten gilt aufgrund Definition 5.7.

5.3 Resümee

In diesem Kapitel wurde ein formales Thesaurusmodell und dessen Darstellung als Graph eingeführt. Somit konnte die mathematische Präzision bei der Definition des mengentheoretischen Modells mit der Anschaulichkeit einer Repräsentation durch Graphen verbunden werden.

Das entwickelte Thesaurusmodell ist die Grundlage für die Analyse und Integration von Komponententhesauri. Daher ist es erforderlich, dass die zu integrierenden Komponententhesauri – ggf. nach einer Vorverarbeitung – dem Thesaurusmodell entsprechen. Als Grundlage für eine Entscheidung, ob ein vorliegender Thesaurus dem Modell entspricht, wurden außer den eigentlichen Definitionen eine Reihe von Invarianten aufgestellt.

Dadurch, dass wir in diesem Kapitel konkret definiert haben, wie die zu integrierenden Thesauri auszusehen haben, können wir in den Folgekapiteln eine auf dieses Thesaurusmodell maßgeschneiderte Integration unter weitestgehender Berücksichtigung der Semantik anbieten. Andererseits ist das Modell aus der Betrachtung einer ganzen Reihe von in der Praxis eingesetzten Thesauri entstanden, so dass wir mit dem Modell möglichst offen für eine Vielzahl von Thesauri geblieben sind.

Nicht zuletzt erfüllt das Thesaurusmodell die in den Abschnitten 2.3.1.1.1 und 2.3.2.2.3 geforderten Anforderungen an ein Modell für Komponententhesauri und ist integraler und zentraler Bestandteil des im folgenden Kapitel 6 entwickelten Informationsmodells für Thesaurusföderationen.

Kapitel 6

Informationsmodell für Thesaurusföderationen

Grundlage eines Multi-Thesaurus-Systems, das den Anforderungen der Skalierbarkeit und Flexibilität Rechnung trägt, ist ein geeignetes *Informationsmodell* (vgl. Abschnitt 2.3.1). Ein solches Informationsmodell beinhaltet das im vorangegangenen Kapitel 5 definierte Modell für Komponententhesauri sowie ein Modell des zusätzlichen Integrationswissens, das in diesem Kapitel den Schwerpunkt bildet. Das Informationsmodell ist die Schnittstelle zwischen Entwicklungs- und Laufzeit eines Multi-Thesaurus-Systems (vgl. Abbildung 6.1).

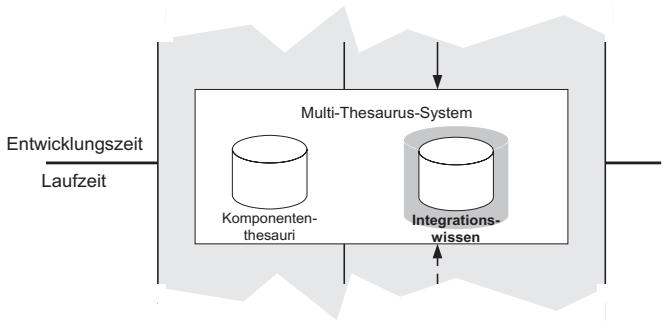

Abbildung 6.1: Modell des Integrationswissens als Bestandteil der Schnittstelle zwischen Entwicklungs- und Laufzeit

Da die semantische Reichhaltigkeit der bekannten Informationsmodelle für Multi-Thesaurus-Systeme und Multi-Ontologie-Systeme diesen Anforderungen nicht entspricht (vgl. Abschnitt 3.1), wird ein eigenes Informationsmodell entwickelt. Zuvor erfolgt eine vertiefte Analyse der Stärken und Schwächen der bekannten Modelle, die die Ausgangsbasis für das eigene Modell bilden. Das formal definierte Informationsmodell bildet die Grundlage für die folgenden Kapitel.

6.1 Analyse

6.1.1 Thesauri

Die im Abschnitt 2.3.1.1.1 geforderte Autonomie der Komponententhesauri wird bereits von verschiedenen Multi-Thesaurus-Systemen erfüllt (vgl. Abschnitt 3.1.3). Von diesen erfüllen Thesaurus-Wechsel-Systeme und die vorgestellten Multi-Ontologie-Systeme die Anforderung nach einer Unterscheidung zwischen Indexierungs- und Nichtindexierungsthesaurus bzw. -ontologien (vgl. Abschnitt 2.3.1.1.2) zumindest teilweise: Es wird zwar das Wechseln von einem System (Nichtindexierungsthesaurus) in das andere (Indexierungsthesaurus) unterstützt, wann es aber erforderlich wird, ist implizites Wissen, das es in einem vollständigen Informationsmodell explizit auszudrücken gilt. Dieser Mangel in den existierenden Modellen kann darauf zurückgeführt werden, dass eine integrale Betrachtung der Integration von Thesauri bzw. Ontologien und der zugehörigen föderierten Informationssysteme bisher nicht stattgefunden hat.

6.1.2 Relationen

Die Anforderungen zur Repräsentation der unterschiedlichen Typen von Beziehungen (vgl. Abschnitt 2.3.1.2.1) werden von Thesaurus-Wechsel-Systemen mit Interlingua und von den vorgestellten Multi-Ontologie-Ansätzen OBSERVER und SKC bereits erfüllt. Bei Thesaurus-Wechsel-Systemen ist dies allerdings eine theoretische Möglichkeit, da uns kein System bekannt ist, das bereits alle geforderten Beziehungstypen zwischen den Komponententhesauri ausdrücken kann. Die OBSERVER- und SKC-Informationsmodelle unterstützen eine größere Vielfalt von Beziehungstypen, die im Thesaurusbereich zum Teil unbekannt sind. Einschränkungen sind notwendig, um z.B. Werkzeuge für einzelne Thesauri einfach für Multi-Thesaurus-Systeme wiederverwenden zu können. Dies gilt insbesondere für das in dieser Hinsicht völlig offene SKC-Modell, das zwar die größtmögliche Flexibilität bietet, aber von den Entwicklern der Multi-Ontologie-Systeme eine Spezifikation der Beziehungstypen und entsprechende Verfahren zum Auffinden von Beziehungen der unterschiedlichen Typen erwartet. Aufgrund der sehr großen Freiheitsgrade greift die Unterstützung durch Standard-Werkzeuge nur begrenzt.

Abgelehnte Beziehungen hingegen werden bisher von keinem Multi-Thesaurus-System unterstützt. Die in jüngster Zeit entstandenen Ergebnisse der Multi-Ontologie-Systeme OBSERVER und SKC belegen durch ihre Unterstützung solch abgelehnter Beziehungen die Erforderlichkeit, die insbesondere zur Auflösung potenzieller Homonyme und Polyseme benötigt werden: Auch wenn davon ausgegangen wird, dass in den Komponententhesauri die unterschiedlichen Bedeutungen von Homonymen (Homographen) und Polysemen aufgelöst und durch näher bestimmende Zusätze kenntlich gemacht sind (vgl. Abschnitt 5.1), kann durch die Integration die Homonymproblematik wiedereingeführt werden (s. z.B. [NK99, MWJ99]). Je breiter das abgedeckte Gebiet ist, desto größer ist die Gefahr solcher Mehrdeutigkeiten [Rec99, S. 3]. Homographen und Polyseme, die bei der Integration erkannt werden, sollen daher explizit als abgelehnte Inter-Thesaurus-Äquivalenzbeziehungen geführt werden. Dem Benutzer kann dann durch die Darstellung des Kontextes (mindestens: aus welchem Thesaurus der Begriff kommt) die Bedeutung ersichtlich gemacht werden.

Bei den bekannten Multi-Ontologie-Systemen fehlt eine Begründung für die Ablehnung, so dass sowohl bei weiteren Integrationsprozessen als auch beim Betrieb nicht erkannt werden kann, warum diese Beziehung abgelehnt wurde. Eine explizite Begründung (z.B. Homonym oder fehlerhafter Vorschlag durch das System) soll ergänzt werden.

6.1.3 Begriffe

Durch die Verbindung mit Inter-Thesaurus-Relationen entstehen *Föderierte Begriffe*. Föderierte Begriffe können daher von solchen Modellen dargestellt werden, die das Ausdrücken von Beziehungen zwischen Begriffen verschiedener Komponententhesauri erlauben[1] (Thesaurus-Wechsel-Systeme, Thesaurus-Kopplungen und alle vorgestellten Multi-Ontologie-Systeme). Wesentlicher Unterschied dieser Modelle ist, dass die föderierten Begriffe entweder als eigenständige Begriffe repräsentiert werden (als Begriffe der Interlingua bei Thesaurus-Wechsel-Systemen mit einer solchen und im SKC- und ONIONS-Modell) oder durch in Beziehung stehende Begriffe aus den Komponententhesauri dargestellt werden (Thesaurus-Wechsel-System ohne Interlingua, OBSERVER-Modell). *Ergänzende Begriffe* hingegen können ausschließlich in Modellen mit Interlingua dargestellt werden (Thesaurus-Wechsel-Systeme mit Interlingua, SKC und ONIONS).

Da die Verwendung bzw. Nicht-Verwendung einer Interlingua zur Darstellung von Begriffen in einem Multi-Thesaurus- bzw. Multi-Ontologie-System der Hauptunterschied ist, sollen die Vor- und Nachteile einer Interlingua analysiert werden. Als Vorteile sind aufzuführen:

- Zur Darstellung von Ergänzenden Begriffen ebenso wie zu einer Modifikation der Beziehungen in Komponententhesauri zum Zwecke einer einheitlichen Sicht auf das Gesamtvokabular ist eine Interlingua (im weitesten Sinne) erforderlich.

- Mit der Interlingua können beliebige Föderierte Begriffe dargestellt werden, deren Zusammenhänge zu Begriffen aus den Komponentensystemen durch Beziehungen unterschiedlicher Typen differenziert werden können.

- Die Interlingua bietet eine zentrale Sicht auf die integrierten Komponentensysteme.

Folgende Nachteile können festgestellt werden:

- Die in den untersuchten Systemen verwendeten Formen einer Interlingua sind wenig flexibel bezogen auf die Teilnahme/Nichtteilnahme von Komponentensystemen. Die Interlingua selbst kann bei unterschiedlich ausgewählten Teilmengen der Komponentensysteme immer nur dieselbe Sicht aller Interlingua-Begriffe und Beziehungen anbieten. So wird beispielsweise ein Ergänzender Begriff, der zum Ausgleich einer Abstraktionsniveaudifferenz zwischen Schwesterknoten zweier Thesauri entstand, auch angezeigt, wenn einer dieser beiden Komponententhesauri gar nicht angezeigt werden soll. Eine Gesamtsicht auf eine Teilmenge der Komponentensysteme ist nicht ohne komplexe Anfragen möglich.

- Die Aufgabe der Datenhaltung für die integrierte Sicht wird im Wesentlichen vom Informationsvermittler durchgeführt. Der eigentliche Informationsanbieter muss dazu für den Betrieb seine Thesaurusdaten aus der Hand geben. In der Praxis ist festzustellen, dass dies häufig nicht erwünscht ist.

- Der Speicherbedarf einer Interlingua mit allen Begriffen und der überwiegenden Zahl aller Intra-Thesaurus-Beziehungen aus allen Komponententhesauri zusätzlich der ergänzten Inter-Thesaurus-Beziehungen und Ergänzender Begriffe ist groß.

[1] Da Thesaurus-Vereinigungen die Autonomie der Komponententhesauri nicht unterstützen, werden sie hier nicht weiter betrachtet. Falls die Autonomie unterstützt wird und Beziehungen zu Komponententhesauri dargestellt werden, kann die Thesaurus-Vereinigung als Interlingua eines Thesaurus-Wechsel-Systems betrachtet werden.

Die Verwendung einer Interlingua (im weitesten Sinne) wird von uns aufgrund des erstgenannten Vorteils als unbedingt erforderlich angesehen. Als Tendenz ist zu beobachten, dass in den Projekten versucht wird, die Interlingua möglichst klein zu halten. In SKC etwa sind in der Interlingua nur Begriffe enthalten, die direkt zur Abbildung benötigt werden oder aber Ergänzende Begriffe. Eine weitere Verkleinerung der Interlingua soll den Informationsanbietern weitestmögliche Datenhoheit bieten. Wesentliche Herausforderung aber ist es, die Flexibilität des Gesamtsystems besser zu unterstützen. Hierzu soll eine semantisch reichere Modellierung entwickelt werden, die die Ansicht von Teilmengen der integrierten Komponentensysteme unterstützt.

6.1.4 Gruppen

In einigen neueren Thesauri, z.B. GEMET [CNR97], erlaubt die Zuordnung der Begriffe zu Gruppen, die Menge der Begriffe thematisch geordnet zugänglich zu machen. Dies erscheint bereits bei einem einzelnen Thesaurus sinnvoll, um dem Benutzer, aber auch dem Thesaurusersteller, eine weitere übergeordnete Orientierungsmöglichkeit zu verschaffen. Eine solche Ordnungsmöglichkeit wird in existierenden Modellen vermisst. Dabei gilt es hier eine noch größere Menge an Begriffen übersichtlich zugänglich zu machen und dabei zum Teil mit einer sehr großen Anzahl an Toptermen der Komponententhesauri umzugehen.

6.1.5 Invarianten und Konflikte

Alle uns bekannten Systeme weisen große Defizite hinsichtlich der expliziten Aufführung von Invarianten auf. Obwohl DIN- und ISO-Normen eine Reihe dieser Invarianten für einzelne Thesauri angeben, wird deren Einhaltung nur in wenigen Systemen überprüft und die Übertragung auf das Multi-Thesaurus-System – falls überhaupt – nur sehr rudimentär angegangen. Die vorgestellten Multi-Ontologie-Systeme verzichten auf die Festlegung von Invarianten und das Überprüfen auf Verstöße gegen diese sogar vollkommen. Der in solchen Fällen erforderliche Umgang mit inkonsistenten Informationen wird ebenfalls nicht weiter betrachtet, sondern allein dem Benutzer überlassen.

In Radas Arbeiten [MR88], die zu den ersten, bedeutenden Arbeiten zur computer-unterstützen Thesaurus-Integration zählen, wurden als einzige Konflikte der Verstoß gegen die paarweise Disjunktheit der Relationen und sowie das Vorhandensein von Oberbegriffen auf unterschiedlichen Abstraktionsniveaus betrachtet. Der erste Fall schließt dabei direkt gegensätzliche Hierarchierelationen ein. Eine Konfliktauflösung findet statt, indem die stärker bindende Hierarchiebeziehung durch eine Assoziationsbeziehung ersetzt wird. Der zweite Fall betrachtet ausschließlich Begriffe, die mehrere Oberbegriffe haben, von denen wiederum mindestens zwei Oberbegriffe einen gemeinsamen direkten oder indirekten Oberbegriff haben. Unterschiedlich lange Pfade werden als Konflikt erkannt, die Konfliktauflösung beseitigt den kürzeren Pfad.

In [SC97] wird ebenfalls die paarweise Disjunktheit für Inter-Thesaurus-Relationen betrachet. Die Zyklenfreiheit wird verallgemeinert auf Zyklen, die über mehrere Stufen entstehen. Diese beiden Invarianten werden formal spezifiziert und überprüft. Wie auf Verstöße reagiert wird, wird dem menschlichen Integrator überlassen.

Weitere explizite Invarianten und der Umgang bei Verstößen gegen diese sind uns nicht bekannt. Wenn das Ergebnis der Integration wiederum ein konsistenter Thesaurus sein soll (Thesaurus-Vereinigung), müssen die Thesaurus-Invarianten weiter gelten und es stellt sich ausschließlich die Frage, wie Konflikte aufgelöst werden. In allen anderen Fällen allerdings ist zusätzlich zu analysieren, welche Invarianten gelten sollen, damit der Entwurfsraum vollständig beschrieben

ist. Die folgenden Beispiele zeigen Freiheitsgrade, die durch Invarianten eingeschränkt werden sollen, um dem Benutzer einen kohärenten Eindruck des Gesamtvokabulars anbieten zu können. Sie belegen auch, dass erst durch eine solche Konfliktanalyse eine Integration verschiedener Komponententhesauri möglich wird, die zur Definition der Begriffe neben den Benennungen die Semantik der Relationen angemessen berücksichtigen.

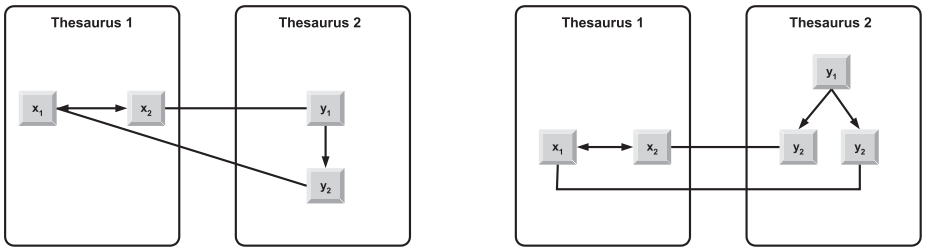

Abbildung 6.2: Unterschiedliche Verwandtschaftstypen zwischen Föderierten Begriffen

Paarweise Disjunktheit der Relationen: Die Beziehung zwischen zwei Begriffen muss eine eindeutige semantische Bedeutung haben. Das heisst, dass zwei Begriffe eines Multi-Thesaurus-Systems durch maximal eine Beziehung miteinander in Beziehung stehen dürfen. Dies schließt z.B. aus, dass zwei Begriffe sowohl Ober-/Unterbegriffspaar einer Hierarchiebeziehung sind als auch durch eine Assoziationsbeziehung verbunden sind. Ebenso wird dadurch ausgeschlossen, dass eine Relation sowohl als akzeptierte als auch als abgelehnte Beziehung aufgeführt wird.

Zyklenfreiheit der Hierarchierelation: Ein Begriff x, der (indirekter) Oberbegriff eines anderen Begriffs y ist, darf nicht zugleich auch (indirekter) Unterbegriff dieses Begriffs y sein. Damit darf die Hierarchierelation des Multi-Thesaurus-Systems keine Zyklen enthalten.

Redundante Beziehungen: Die Problematik der durch die Integration möglicherweise entstehenden redundanten Beziehungen wurde bereits in Abschnitt 2.3.1.4 vorgestellt. Da die Verwendung solcher redundanten Beziehungen dazu führt, dass ein Begriff gleichzeitig Unterbegriff auf verschiedenen Ebenen ist, ist die resultierende Forderung – wie für Komponententhesauri auch – die Redundanzfreiheit der Hierarchierelation in Multi-Thesaurus-Systemen.

Unterschiedliche Verwandtschaftstypen: Durch die Integration von Begriffen aus verschiedenen Thesauri können die Föderierten Begriffe aufgrund unterschiedlicher Beziehungen in den Komponententhesauri in verschiedenen Beziehungen miteinander stehen. In Abbildung 6.2 wird links der Fall dargestellt, dass die Föderierten Begriffe (x_2, y_1) und (x_1, y_2) sowohl in einer Hierarchiebeziehung als auch in einer Assoziationsbeziehung stehen. Dieser Fall wird durch eine auch für Föderierte Begriffe geltende Forderung der paarweisen Disjunktheit der Relationen abgefangen.

In Abbildung 6.2 rechts wird der Fall dargestellt, dass die Föderierten Begriffe (x_2, y_2) und (x_1, y_3) sowohl Schwesterknoten als auch assoziierte Begriffe sind. Im Allgemeinen soll eine solche Assoziationsbeziehung vermieden werden, da durch den gemeinsamen Oberbegriff bereits ein Bezug zwischen den Begriffen ausgedrückt ist.

Ein Spezialfall unterschiedlicher Beziehungen zwischen Föderierten Begriffen ist der Fall der unterschiedlichen Typen von Hierarchiebeziehungen (vgl. Abbildung 6.3). Der links

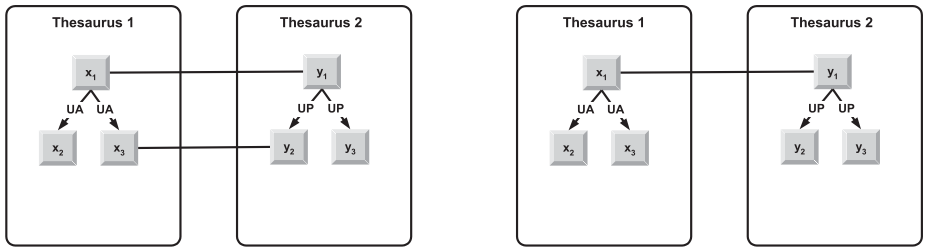

Abbildung 6.3: Unterschiedliche Typen von Hierarchiebeziehungen zwischen Föderierten Begriffen.

dargestellte Fall kann wiederum auf die Forderung der paarweisen Disjunktheit der Relationen zurückgeführt werden. Anders hingegen der rechts dargestellte Fall, bei dem Begriffe, die in den Komponententhesauri entweder ausschließlich Abstraktions- oder Bestandsrelationsunterbegriffe hatten, als Föderierter Begriff beide Typen von Unterbegriffen hat. Eine solche Vermischung wird allgemein als qualitätsmindernd angesehen.

Verwendung Ergänzender Begriffe: Die Notwendigkeit der Verwendung Ergänzender Begriffe wurde bereits erläutert. Mit der Einführung solcher Ergänzenden Begriffe ergeben sich aber weitere Freiheitsgrade: Können Ergänzende Begriffe miteinander in Beziehung stehen? Falls ja, welche Beziehungstypen sind erlaubt? Dürfen Ergänzende Begriffe eingeführt werden, die keine Beziehung zu einem Begriff aus einem Komponentenhesaurus eingehen?

Verwendung von 1:n-Äquivalenzbeziehungen: Auch wenn 1:n-Äquivalenzbeziehungen recht frei verwendet werden können, soll zumindest der in Abbildung 6.4 (links) dargestellte Fall einer 1:n-Äquivalenzbeziehung zu einem Begriff und zu einem der direkten oder indirekten Abstraktionsunterbegriffe untersagt werden. Da der untergeordnete Begriff in solch einem Falle eine Spezialisierung des übergeordneten Begriffs ist, ist die Beziehung zu dem übergeordneten Begriff überflüssig.

Abstraktionsniveaudifferenzen: Zwischen zwei Begriffen besteht eine Abstraktionsniveaudifferenz, wenn sich die Begriffe auf unterschiedlichen Abstraktionsebenen, befinden, wie dies z.B. für Landwirtschaft und Betrieb der Fall sein kann. Solche Abstraktionsniveaudifferenzen können z.T. nur durch die Interpretation der Bedeutung erkannt werden. Abbildung 6.4 (rechts) verdeutlicht so einen Fall, bei dem durch die Integration zweier Begriffe Schwesterknoten auf unterschiedlichem Abstraktionsniveau entstanden sind, dies aber durch Betrachtung des Graphen nicht erkannt werden kann.

Da Abstraktionsniveaudifferenzen die darunterliegenden Hierarchien und damit die Struktur des Multi-Thesaurus-Systems stören, werden solche Unterschiede allgemein als nicht erwünscht betrachtet (s. z.B. [MR88]). Die Feststellung einer Abstraktionsniveaudifferenz kann entweder zusätzliches semantisches Wissen über das Abstraktionsniveau erfordern oder aber es existieren weitere Beziehungen im Graphen, die das Erkennen möglich machen. Beispiele für solche Beziehungen sind Hierarchiebeziehungen wie in Abbildung 2.4, S. 18, dargestellt, weitere Äquivalenzrelationen (s. Abbildung 6.5, links) oder gemeinsame Unterbegriffe, die über verschieden lange Pfade erreicht werden (s. Abbildung 6.5, rechts). Die beiden erstgenannten Fälle können auf eine Verletzung der Redundanzfreiheit der Hierarchierelation zurückgeführt werden, der letztgenannte Fall hingegen muss gesondert betrachtet werden. Wie bei redundanten Beziehungen entsteht das Problem, dass ein und

derselbe Begriff beim Navigieren durch die Hierarchie z.B. mit dem Ziel, Begriffe für eine Anfrageerweiterung zu finden, auf verschiedenen Abstraktionsebenen gefunden wird.

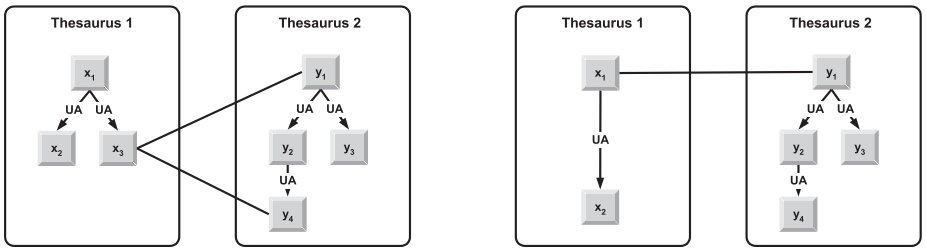

Abbildung 6.4: 1:n-Äquivalenzbeziehung zu einem Begriff und seinem (indirekten) Abstraktionsunterbegriff (links). Unterschiedliches Abstraktionsniveau der Schwesterknoten, das ohne zusätzliches semantisches Wissen anhand des Graphen nicht erkannt werden kann (rechts).

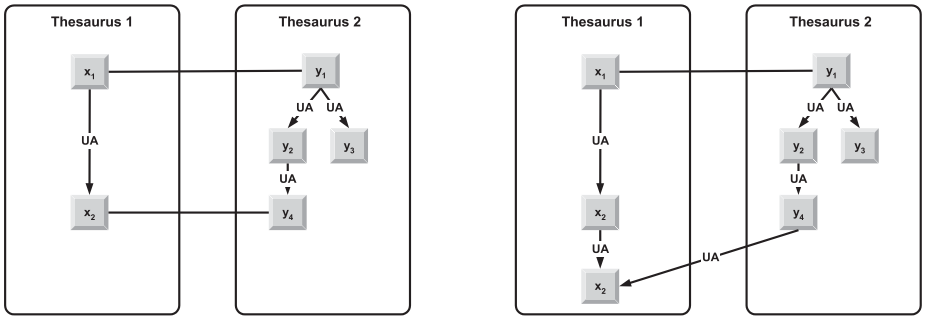

Abbildung 6.5: Abstraktionsniveaudifferenzen, die anhand von weiteren Äquivalenzbeziehungen (links) bzw. gemeinsamen untergeordneten Begriffen erkannt werden können (rechts).

6.2 Informationsmodell

Nachdem die im vorangegangenen Abschnitt dargelegten Ergebnisse der Analyse die Stärken und Schwächen der aus der Literatur bekannten Modelle für Multi-Thesaurus-Systeme aufgezeigt haben, wird ausgehend von diesen Modellen ein neues Modell entwickelt. Dieses erfüllt, wie wir zeigen werden, erstmals alle in Abschnitt 2.3.1 aufgeführten Anforderungen. Multi-Thesaurus-Systeme, die diesem Modell genügen, werden *Thesaurusföderationen* genannt.

Eine Thesaurusföderation ist eine der losen Kopplung von Komponenten-Informationssystemen zu übergreifenden Informationssystemen adäquate Kopplung der entsprechenden Thesauri dieser Systeme. Grundidee der Thesaurusföderation ist es, eine *integrierte Sicht auf das Gesamtvokabular* der Komponententhesauri unter *Beibehaltung ihrer Autonomie* anzubieten (vgl. Abschnitt 3.1.1.3).

Das Informationsmodell enthält als Bestandteile ein Modell für Komponententhesauri und ein Modell für das zusätzliche Integrationswissen, das für die integrierte Sicht auf das Gesamtvokabular erforderlich ist.

6.2.1 Komponententhesauri

Da die Komponententhesauri aufgrund ihrer Autonomie unverändert in die Thesaurusföderation eingehen, ist das Modell für die Komponententhesauri integraler Bestandteil des Modells für Thesaurusföderationen. Es wurde bereits in Kapitel 5 vorgestellt.

Es sei angemerkt, dass an einer Thesaurusföderation sowohl Komponententhesauri mit Gruppen als auch ohne Gruppen teilnehmen können (falls Gruppen vorhanden sind, müssen alle Begriffe genau einer Gruppe zugeordnet sein) und auch eine Unterscheidung von Hierarchierelationen in Abstraktions- und Bestandsrelationen nicht von vorn herein erforderlich ist. Weitere Eigenschaften von Komponententhesauri (z.B. Supergruppen und Themen in GEMET) werden innerhalb der Föderation nicht berücksichtigt.

6.2.2 Integrationswissen

Anhand der in Abschnitt 6.1 vorgestellten Ergebnisse der Analyse entwickeln wir nun das Informationsmodell für das Integrationswissen, das aus einer Menge von Thesauri eine Thesaurusföderation werden lässt. Wir führen das Modell informell ein, um das Verständnis des Modelles zu erleichtern. In Abschnitt 6.3 folgt dann die formale Definition des gesamten Informationsmodelles.

6.2.2.1 Metainformationen über Komponententhesauri

Innerhalb der Thesaurusföderation werden Informationen darüber benötigt, in welchen Informationssystemen ein Komponententhesaurus als Indexierungs- und Retrievalthesaurus verwendet wird. Diese Informationen werden als Metainformationen über die Komponententhesauri vorgehalten. Auch solche Komponententhesauri können an der Föderation teilnehmen, die in keinem beteiligten Informationssystem als Indexierungsthesaurus verwendet werden. Somit wird die Möglichkeit der Unterscheidung von Indexierungs- und Nichtindexierungsthesauri gegeben. Zudem kann eine Anfragekomponente die Informationen erhalten, welche Begriffe aus welchen Thesauri ein bestimmtes Informationssystem versteht, und eine entsprechende Abbildung vornehmen.

6.2.2.2 Relationen

Die in Abschnitt 2.3.1.2.1 dargestellten Anforderungen können erfüllt werden, indem die aus den klassischen Thesauri bekannten Relationen (Äquivalenz, Hierarchie, Assoziation) übertragen werden und als *Inter-Thesaurus-Relationen* die *Verknüpfungen* zwischen den Begriffen aus den verschiedenen Thesauri bilden. Die Hierarchierelation wird dabei weiter unterteilt in eine Abstraktionsrelation (generische Hierarchiebeziehung) und eine Bestandsrelation (partitive Hierarchiebeziehung). Instanzbeziehungen werden als Spezialfall der Abstraktionsrelation betrachtet und nicht weiter unterschieden.

Um auch die in Abschnitt 2.3.1.2.2 geforderten abgelehnten Beziehungen ausdrücken zu können, kann jede Inter-Thesaurus-Relation den Status *akzeptiert* oder *abgelehnt* besitzen. Im Falle einer abgelehnten Beziehung kann eine Begründung für die Ablehnung ergänzt werden.

Die Verwendung dieser eingeschränkten Menge von Relationstypen hat mehrere Vorteile:

- Alle in Abschnitt 2.3.1.2.2 geforderten Typen von Beziehungen können ausgedrückt wer-

den.

- Die eingeführten Inter-Thesaurus-Relationen sind den Anwendern bereits als Intra-Thesaurus-Relationen vertraut. Es ist nicht erforderlich, sich mit neuer Semantik vertraut zu machen.
- Werkzeuge, die mit einzelnen Thesauri umgehen können (z.B. Thesaurus-Browser), können einfach erweitert werden, so dass sie mit Thesaurusföderationen umgehen können (ähnliche Struktur).
- Verfahren zur Thesaurusintegration können von einer klar definierten Semantik der Inter-Thesaurus-Relationen ausgehen.

Die Intra-Thesaurus-Äquivalenzbeziehungen und die Intra-Thesaurus-Benutze-Kombination-Beziehungen in Komponententhesauri sind zwischen Nicht-Deskriptoren und Deskriptoren etabliert, die gemeinsam einen Begriff repräsentieren. Die Deskriptoren sind dabei ausgezeichnete Vorzugsrepräsentanten des Begriffs, die z.B. zur Indexierung in einem Information-Retrieval-System verwendet werden. Die Inter-Thesaurus-Äquivalenzbeziehungen und die Inter-Thesaurus-Benutze-Kombination-Beziehungen hingegen verbinden Deskriptoren in den verschiedenen Thesauri. Der neue Föderierte Begriff besitzt dann nicht länger einen einzigen Vorzugsrepräsentanten. Die Deskriptoren aus den verschiedenen Thesauri werden als gleichwertigeVorzugsrepräsentanten angesehen.

6.2.2.3 Begriffe

Wie bereits in Abschnitt 6.1.3 erwähnt, entstehen *Föderierte Begriffe* durch die Verbindung mit Inter-Thesaurus-Relationen. Im Gegensatz zu äquivalenten Bezeichnern in einem Komponententhesaurus, von denen ein Vorzugsbezeichner zur Repräsentation des Begriffes ausgewählt wird, wird aus einer Menge von äquivalenten Begriffen der Föderation keinerlei Auswahl festgelegt. Die vollständige Repräsentation findet durch die Menge aller entsprechenden Deskriptoren statt[2] (vgl. Abschnitt 2.3.1.3.1).

Durch das Ausdrücken von Föderierten Begriffen durch Inter-Thesaurus-Beziehungen sind keine eigenständigen Begriffe innerhalb einer Interlingua erforderlich. Das Modell der Thesaurusföderation sieht als einzige Begriffe einer „Interlingua“ *Ergänzende Begriffe* vor. Solche Ergänzenden Begriffe werden eingeführt, wenn Schwesterbeziehungen dargestellt oder Abstraktionsniveaudifferenzen ausgeglichen werden sollen (vgl. Abschnitte 2.3.1.3.2 und 6.1.3). Die Komponententhesauri, aufgrund derer ein Ergänzender Begriff eingefügt wurde, sind anhand der Inter-Thesaurus-Beziehungen erkenntlich. Zwischen den Ergänzenden Begriffen können Intra-Thesaurus-Beziehungen etabliert werden, um semantische Bezüge ausdrücken zu können. Ergänzende Begriffe ohne Beziehungen zu Komponententhesauri dürfen nur eingefügt werden, wenn sie zur Strukturierung anderer Ergänzender Begriffe verwendet werden, d.h. Schwesterknoten unter einem gemeinsamen Vaterknoten zusammenfassen oder Abstraktionsniveaudiffernzen ausgleichen.

6.2.2.4 Gruppen

Wir greifen das aus verschiedenen neueren Thesauri bekannte Konzept einer den Toptermen übergeordneten thematischen Ordnung auf und führen es als Bestandteil der Thesaurusfödera-

[2]Lexikalisch identische Deskriptoren können zur Darstellung selbstverständlich zusammengefasst werden.

tion ein. Die Thesaurusföderation besitzt somit thematische Gruppen, so genannte *Föderierte Gruppen*, denen die Topterme der Komponententhesauri zugeordnet werden. Ein navigatorischer Einstieg über eine überschaubar zu haltende Anzahl von Gruppen wird damit ebenso möglich wie eine thematisch eingeschränkte Suche nach Begriffen. Da die Gruppen eine den Toptermen und somit der gesamten Begriffstruktur übergeordnete Ebene sind, erlauben wir – wie durch die Polyhierarchie innerhalb der Struktur auch – eine Mehrfachzuordnung der Topterme zu den Gruppen. Die Föderierten Gruppen können inhaltlich ganz oder teilweise mit evtl. vorhandenen Gruppen der Komponententhesauri übereinstimmen – gefordert wird vom Informationsmodell eine solche Übereinstimmung aber nicht.

6.2.2.5 Invarianten und Konflikte

Es wird vorausgesetzt, dass die an einer Thesaurusföderation teilnehmenden Komponententhesauri konsistent enstprechend den in Abschnitt 5.2.3 aufgeführten Invarianten sind. Falls dies nicht der Fall ist, werden inkonsistente Beziehungen markiert und innerhalb der Föderation nicht weiter berücksichtigt.

Anhand der in Abschnitt 6.1.5 durchgeführten Analyse können eine Reihe von Invarianten für Thesaurusföderationen hergeleitet werden. Auf eine vollzählige Aufführung wird hier verzichtet, da alle Invarianten im folgenden Abschnitt 6.3 formal definiert werden.

Zentrale Idee der Thesaurusföderationen im Zusammenhang mit Invarianten ist es, zwischen Invarianten zu unterscheiden, bei denen Verstöße gegen diese *a priori* bzw. *a posteriori* aufgelöst werden. Invarianten, bei denen Verstöße a priori aufgelöst werden, werden somit bereits während der Erstellung der Thesaurusföderation erzwungen. Werden Verstöße gegen Invarianten a posteriori aufgelöst, bedeutet dies, dass diese Verstöße bei der Erstellung der Thesaurusföderation nicht beseitigt, sondern markiert werden und dann situationsabhängig bei der Anfrage aufgelöst werden (z.B. Verstöße gegen die Redundanzfreiheit der Hierarchierelation). So kann etwa eine Beziehung bei Betrachtung zweier Thesauri angezeigt werden, die bei Betrachtung eines weiteren Thesaurus als redundant ausgeblendet wird. Oder ein Zyklus in der Hierarchierelation kann in Abhängigkeit des vom Benutzer präferierten Thesaurus aufgelöst werden. Vorteil ist die gewonnene Flexibilität des Systems.

Das Modell für Thesaurusföderationen enthält somit zusätzlich zu den eigentlichen Invarianten Möglichkeiten, Verstöße gegen Invarianten in Form von Konfliktmarkierungen auszudrücken. Dabei werden Konflikttyp und alle konfliktverursachenden Beziehungen vermerkt.

6.2.2.6 Zusammenfassung

Aufgrund der Anforderungen der Flexibilität, der erwünschten Datenhoheit der Thesaurus-Anbieter und der Minimalität (vgl. Abschnitt 6.1.3) verzichten wir zur Darstellung des Integrationswissens auf eine Interlingua im eigentlichen Sinne (Abbildung der Thesaurusbegriffe geschieht in jedem Fall indirekt über die Begriffe der Interlingua). Stattdessen enthält die von uns entwickelte „Interlingua“ ausschließlich die Ergänzenden Begriffe, die Inter-Thesaurus-Relationen, Föderierte Gruppen (vgl. Abschnitt 6.2.2.4) sowie Konfliktmarkierungen (vgl. Abschnitt 6.2.2.5). Um den Unterschied zu einer Interlingua zu verdeutlichen, sprechen wir daher von *Extralingua*. Diese Extralingua enthält ausschließlich *zusätzliches Integrationswissen*, das in den Komponententhesauri nicht enthalten ist. Gemeinsam mit den Komponententhesauri stellt diese Extralingua somit die Thesaurusföderation dar.

6.3 Formales Thesaurusföderations-Modell

6.3.1 Thesaurusföderation

Definition 6.1 *Formal wird eine* Thesaurusföderation *definiert als Tupel*

$$\Im = (\Theta, I, M, \omega, G_\Im, K, \mathcal{E}, \mathcal{C}, \mathcal{A}, \mathcal{P}, \mathcal{V}, \mathsf{A}, \mathsf{P}, \mathsf{V})$$

wobei $\Theta = \{\theta_1, \theta_2, \cdots, \theta_k\}$ *die Menge der an der Föderation beteiligten Komponententhesauri und* $I = \{i_1, i_2, \cdots, i_l\}$ *die Menge der Informationssysteme, in denen die Komponententhesauri zum Indexieren und Retrieval verwendet werden, bezeichnet.* M *enthält für jedes* $\theta \in \Theta$ *eine Teilmenge von* I*, die die Informationssysteme enthält, in denen* θ *verwendet wird.* ω *bezeichnet einen Thesaurus ohne Gruppen, der die Ergänzenden Begriffe sowie die Beziehungen zwischen diesen enthält.* $G_\Im = \{g_1, g_2, \cdots, g_m\}$ *ist die Menge der Föderierten Gruppen und* $K = \{k_1, k_2, \cdots, k_n\}$ *die Menge der Konfliktmarkierungen.* $\mathcal{E}, \mathcal{C}, \mathcal{A}, \mathcal{P}$ *und* $\mathcal{V}$ *sind die Inter-Thesaurus-Äquivalenzrelation, die Inter-Thesaurus-Benutze-Kombination-Relation, die Inter-Thesaurus-Abstraktionsrelation, die Inter-Thesaurus-Bestandsrelation und die Inter-Thesaurus-Assoziationsrelation (Verwandtschaftsrelation).* A, P *und* V *sind die implizierten Intra-Thesaurus-Abstraktions-, -Bestands- und -Assoziationsrelationen.*

In den folgenden Abschnitten werden die Elemente von $\Im$ näher erläutert. Vorgreifend jedoch eine Anmerkung zu den neu eingeführten implizierten Beziehungen: Implizierte Beziehungen sind solche Beziehungen, die nicht explizit etabliert wurden, aus vorhandenen Beziehungen und den Relationseigenschaften aber abgeleitet werden können. Werden etwa zwei Deskriptoren durch eine Inter-Thesaurus-Äquivalenzbeziehung verbunden, werden alle Beziehungen des jeweils einen Deskriptors zu weiteren Deskriptoren auch für den jeweils anderen Deskriptor gültig, die Beziehungen werden impliziert (vgl. Abbildung 6.6). Denn diese Deskriptoren werden durch die Inter-Thesaurus-Äquivalenzbeziehung Repräsentanten ein und desselben Begriffs.

Bemerkung 6.1 *Definition 6.1 besagt, dass implizierte Beziehungen explizit aufgeführt werden. Wie noch gezeigt werden wird, wird der erforderliche Mehraufwand zur expliziten Nennung implizierter Beziehungen durch den Vorteil der besseren Übersichtlichkeit und höheren Performanz ausgeglichen.*

Wie wir später zeigen werden, können auch Intra-Thesaurus-Beziehungen impliziert werden. Während die implizierten Inter-Thesaurus-Beziehungen innerhalb der Intra-Thesaurus-Relationen repräsentiert werden können, sind, da keine Eingriffe in die Komponententhesauri möglich sind, für solche implizierten Beziehungen separate Relationen vorzusehen.

6.3.2 Komponententhesauri und Metainformationen

Definition 6.2 *Jeder* Thesaurus $\theta \in \Theta$ *ist mit seinem* eindeutigen Identifikator $\#\theta$ *durch die Bijektion* $I_\Theta : \Theta \to \{1, ..., |\Theta|\} \subseteq \mathbb{N}$ *assoziiert. Das bedeutet* $\forall \theta \in \Theta : I_\Theta(\theta) = \#\theta$.

Für die Komponententhesauri $\theta \in \Theta$ wurden bereits im vorangegangenen Kapitel 5 ausführliche Definitionen gegeben. Wir gehen im Folgenden davon aus, dass sie diesen Definitionen genügen. Ein Element eines Komponententhesaurus kann durch Voranstellen des Bezeichners des Komponententhesauruses mit einem Punkt bezeichnet werden. So bedeutet z.B. $\theta_1.B$ die Menge der Benennungen des Thesaurus θ_1.

Die von der Thesaurusföderation verwalteten Metainformationen betreffen Informationen darüber, in welchen Informationssystemen ein Komponententhesaurus jeweils zum Indexieren und Retrieval eingesetzt wird.

Definition 6.3 *Jedes Informationssystem $i \in I$ ist mit seinem eindeutigen Identifikator $\#i$ durch die Bijektion $I_I : I \to \{1, ..., |I|\} \subseteq \mathbb{N}$ assoziiert. Das bedeutet $\forall i \in I : I_I(i) = \#i$.*

Definition 6.4 *Für jeden Thesaurus $\theta \in \Theta$ ist in der Thesaurusföderation jeweils eine Menge $m_\theta \in M$ von Informationssystemen als Teilmenge von I wie folgt definiert:*

$$m_\theta = \{\#i \ : \ i \in I \text{ ist ein Informationssystem, in dem } \theta \text{ zum Indexieren und Retrieval verwendet wird}\}$$

6.3.3 Begriffe und Benennungen

Definition 6.5 *Ein* Ergänzender Begriff *wird wie ein Begriff eines Komponententhesaurus definiert*[3]*, d.h. zu jedem Ergänzenden Begriff muss ein Deskriptor angegeben werden. Des Weiteren können als äquivalent betrachtete Nicht-Deskriptoren, eine Definition, eine Erläuterung sowie eine Homonymauflösung spezifiziert werden. Zwischen Ergänzenden Begriffen können bis auf die Benutze-Kombination-Beziehung die üblichen Thesaurus-Beziehungen etabliert werden. Eine Zuordnung zu eigenständigen Gruppen ist aber nicht möglich. Ein Ergänzender Begriff muss mindestens in einer Inter-Thesaurus-Beziehung vorkommen, die keine Inter-Thesaurus-Äquivalenzbeziehung ist*[4]*, oder aber mindestens zwei Begriffe als Unterbegriffe besitzen*[5]*.*

Wir betrachten die Menge der Ergänzenden Begriffe daher als einen ausgezeichneten Komponententhesaurus ω, der die in Abschnitt 5.1 definierten Thesaurus-Eigenschaften besitzt und für den zusätzlich gilt:

$$\begin{aligned}
\omega.G &= \emptyset \\
\omega.\mathrm{C} &= \emptyset \\
d \in \omega.\mathrm{D} \ &\Rightarrow \ [\exists \theta \in \Theta : \exists x \in \theta.D : (d,x) \in \mathcal{A} \vee (x,d) \in \mathcal{A} \vee \\
&\quad (d,x) \in \mathcal{P} \vee (x,d) \in \mathcal{P} \vee (d,x) \in \mathcal{V}] \vee \\
&\quad [\exists d_1, d_2 \in \omega.\mathrm{D} : (\#d \in \omega.O^{\mathrm{A}}_{d_1} \wedge \#d \in \omega.O^{\mathrm{A}}_{d_2}) \vee \\
&\quad (\#d \in \omega.O^{\mathrm{P}}_{d_1} \wedge \#d \in \omega.O^{\mathrm{P}}_{d_2})]
\end{aligned}$$

Definition 6.6 *Wir notieren kurz $\mathbf{B} = \theta_1.B \cup \theta_2.B \ \ldots \ \cup \theta_k.B \cup \omega.B$ als die Menge aller Benennungen, $\mathbf{B}_{\overline{\theta}} = \mathbf{B} - \theta.B$ als die Menge aller Benennungen ohne die Benennungen aus Thesaurus θ, $\mathbf{D} = \theta_1.D \cup \theta_2.D \ \ldots \ \cup \theta_k.D \cup \omega.D$ als die Menge aller Deskriptoren und $\mathbf{D}_{\overline{\theta}} = \mathbf{D} - \theta.D$ als die Menge aller Deskriptoren ohne die Deskriptoren aus Thesaurus θ.*

Definition 6.7 *Die Funktion $\gamma : \mathbf{B} \to \{\theta_1, \ \ldots \ , \theta_k, \omega\}$ liefert zu jeder Benennung aus einem Komponententhesaurus bzw. aus den Ergänzenden Begriffen die Bezeichnung des entsprechenden Thesaurus bzw. der Menge der Ergänzenden Begriffe.*

[3] Wir treffen zur Vereinfachung unserer Betrachtungen die Einschränkung, dass kein Ergänzender Begriff eingeführt werden kann, der durch einen Nicht-Deskriptor und n $\geq$ 2 Deskriptoren, auf die vom Nicht-Deskriptor durch eine Benutze-Kombination-Beziehung verwiesen wird, repräsentiert wird. Stattdessen soll in einem solchen Fall der komplexere Begriff durch einen eigenen Deskriptor repräsentiert werden.

[4] Stände ein Ergänzender Begriff ausschließlich in einer Inter-Thesaurus-Äquivalenzbeziehung, bräuchte er nicht eingeführt zu werden, da die Föderation um keinen neuen Begriff ergänzt würde, sondern dieser bereits in einem Komponententhesaurus vorhanden wäre.

[5] Die Einführung strukturierender Ergänzender Begriffe wird somit erlaubt.

Definition 6.8 *Jede Benennung $b \in \mathbf{B}$ ist mit ihrem eindeutigen Identifikator $\star b$ durch die Bijektion $I_{\mathbf{B}} : \mathbf{B} \to \{\gamma(a).\#a : a \in \mathbf{B}\}$ assoziiert, wobei $I_{\mathbf{B}}(b) = \gamma(b).\#b = \star b$ gilt.*

Definition 6.9 *Die Funktion $\gamma^\star : \{\star b : b \in \mathbf{B}\} \to \Theta \cup \omega$ liefert zu einem eindeutigen Bezeichner einer Benennung die Bezeichnung des entsprechenden Thesaurus. Sie wird definiert als Verknüpfung der zu $I_{\mathbf{B}}$ inversen Funktion $I_{\mathbf{B}}^{-1}$ und der Funktion γ, das bedeutet $\gamma^\star = \gamma \circ I_{\mathbf{B}}^{-1}$.*

Definition 6.10 *Die auf den Mengen X und Y definierte Projektionsfunktion $\iota : X \times Y \to X$ bildet zweielementige Tupel auf das erste Element ab, d.h. es gilt: $\forall x \in X, y \in Y : \iota((x, y)) = x$.*

Definition 6.11 *Sei $x \in \mathbf{D}$ ein Deskriptor aus einem Thesaurus θ (der hier ein Komponententhesaurus oder der Thesaurus mit den Ergänzenden Begriffen ist). Für jeden Deskriptor x sind zusätzlich zu den thesaurusinternen Mengen jeweils eine Menge von Äquivalenten Begriffen $U_x^{\mathcal{E}}$, Benutze-Kombination-Begriffen $U_x^{\mathcal{C}}$, Abstraktionsoberbegriffen $O_x^{\mathcal{A}}$, Bestandsoberbegriffen $O_x^{\mathcal{P}}$ und Verwandten Begriffen $\mathcal{R}_x$ definiert. Entsprechend sind abgelehnte Äquivalente Begriffe $\overline{U_x^{\mathcal{E}}}$, abgelehnte Benutze-Kombination-Begriffe $\overline{U_x^{\mathcal{C}}}$, abgelehnte Abstraktionsoberbegriffe $\overline{O_x^{\mathcal{A}}}$, abgelehnte Bestandsoberbegriffe $\overline{O_x^{\mathcal{P}}}$ und abgelehnte Verwandte Begriffe $\overline{\mathcal{R}_x}$ sowie entsprechende Mengen $\widetilde{U_x^{\mathcal{E}}}$, $\widetilde{U_x^{\mathcal{C}}}$, $\widetilde{O_x^{\mathcal{A}}}$, $\widetilde{O_x^{\mathcal{P}}}$ und $\widetilde{\mathcal{R}_x}$ mit abgelehnten Begriffen, die zusätzlich eine Begründung der Ablehnung enthalten, definiert. O_x^{A} definiert implizierte Abstraktionsoberbegriffe, O_x^{P} implizierte Bestandsoberbegriffe und R_x implizierte Assoziationsbegriffe. Schließlich sind Föderierte Gruppen $Q_{\Im_x}$ mit thesaurusübergreifendem Charakter definiert.*

$$
\begin{aligned}
U_x^{\mathcal{E}} &= \{\star y : y \in \mathbf{D} \\
&\quad \textit{repräsentiert einen zu } x \textit{ äquivalenten Begriff}\} \\
U_x^{\mathcal{C}} &= \{\star y : y \in \mathbf{D}_{\overline{\theta}} \textit{ ist ein Faktor von } x\} \\
O_x^{\mathcal{A}} &= \{\star y : y \in \mathbf{D}_{\overline{\theta}} \textit{ repräsentiert einen Abstraktionsoberbegriff von } x\} \\
O_x^{\mathcal{P}} &= \{\star y : y \in \mathbf{D}_{\overline{\theta}} \textit{ repräsentiert einen Bestandsoberbegriff von } x\} \\
\mathcal{R}_x &= \{\star y : y \in \mathbf{D}_{\overline{\theta}} \textit{ repräsentiert einen Verwandten Begriff von } x\} \\
\widetilde{U_x^{\mathcal{E}}} &= \{(\star y, s) : y \in \mathbf{D}_{\overline{\theta}} \textit{ repräsentiert einen zu } x \\
&\quad \textit{mit der Begründung } s \textit{ abgelehnten äquivalenten Begriff}\} \\
\widetilde{U_x^{\mathcal{C}}} &= \{(\star y, s) : y \in \mathbf{D}_{\overline{\theta}} \textit{ ist ein mit der Begründung } s \\
&\quad \textit{abgelehnter Faktor von } x\} \\
\widetilde{O_x^{\mathcal{A}}} &= \{(\star y, s) : y \in \mathbf{D}_{\overline{\theta}} \textit{ repräsentiert einen} \\
&\quad \textit{mit der Begründung } s \textit{ abgelehnten Abstraktionsoberbegriff von } x\} \\
\widetilde{O_x^{\mathcal{P}}} &= \{(\star y, s) : y \in \mathbf{D}_{\overline{\theta}} \textit{ repräsentiert einen} \\
&\quad \textit{mit der Begründung } s \textit{ abgelehnten Bestandsoberbegriff von } x\} \\
\widetilde{\mathcal{R}_x} &= \{(\star y, s) : y \in \mathbf{D}_{\overline{\theta}} \textit{ repräsentiert einen} \\
&\quad \textit{mit der Begründung } s \textit{ abgelehnten Verwandten Begriff von } x\} \\
\overline{U_x^{\mathcal{E}}} &= \iota(\widetilde{U_x^{\mathcal{E}}}) \\
\overline{U_x^{\mathcal{C}}} &= \iota(\widetilde{U_x^{\mathcal{C}}}) \\
\overline{O_x^{\mathcal{A}}} &= \iota(\widetilde{O_x^{\mathcal{A}}}) \\
\overline{O_x^{\mathcal{P}}} &= \iota(\widetilde{O_x^{\mathcal{P}}}) \\
\overline{\mathcal{R}_x} &= \iota(\widetilde{\mathcal{R}_x})
\end{aligned}
$$

$$
\begin{aligned}
O_x^{\mathrm{A}} &= \{\star y : y \in \theta.D \textit{ repräsentiert einen implizierten} \\
&\quad \textit{Intra-Thesaurus-Abstraktionsoberbegriff von } x\} \\
O_x^{\mathrm{P}} &= \{\star y : y \in \theta.D \textit{ repräsentiert einen implizierten} \\
&\quad \textit{Intra-Thesaurus-Bestandsoberbegriff von } x\} \\
\mathsf{R}_x &= \{\star y : y \in \theta.D \textit{ repräsentiert einen implizierten} \\
&\quad \textit{Verwandten Begriff von } x\} \\
Q_{\Im_x} &= \{g : g \in G_{\Im} \textit{ ist Föderierte Gruppe von } x\}
\end{aligned}
$$

wobei aufgrund der Definition von $\mathbf{D}_{\overline{\theta}}$ sichergestellt ist, dass die jeweiligen Elemente der Mengen aus einem von $\gamma^\star(\star x)$ verschiedenen Thesaurus kommen und damit auch ein Selbstbezug ausgeschlossen ist. Ausnahme ist die Menge $U_x^{\mathcal{E}}$, die einen Selbstbezug erlaubt, da ein Begriff auch zu sich selbst äquivalent ist.

Es gelten weiterhin die folgenden Bedingungen

$$\forall \star y_1, \star y_2 \in U_x^{\mathcal{E}} : \gamma^\star(\star y_1) = \gamma^\star(\star y_2) \Leftrightarrow \star y_1 = \star y_2 \tag{6.1}$$

$$\forall x, y, z \in \mathbf{D} : \star y \in U_x^{\mathcal{E}} \wedge \star z \in U_y^{\mathcal{E}} \Rightarrow \star z \in U_x^{\mathcal{E}} \tag{6.2}$$

$U_x^{\mathcal{C}} \neq \emptyset \Rightarrow$

$$\forall \star u \in U_x^{\mathcal{C}} : \gamma^\star(\star u) \neq \gamma^\star(\star x) \wedge \exists \star v \in U_x^{\mathcal{C}} : \ \star u \neq \star v \wedge \gamma^\star(\star u) = \gamma^\star(\star v) \tag{6.3}$$

$$\forall u, v \in U_x^{\mathcal{C}} : u \notin \gamma^\star(\star v).O_v^{\mathrm{A}} \tag{6.4}$$

$$\forall x, y \in \mathbf{D} : \star y \in U_x^{\mathcal{E}} \Rightarrow \forall \star z \in U_x^{\mathcal{C}} : \gamma^\star(\star y) \neq \gamma^\star(\star z) \tag{6.5}$$

$\forall x_1, x_2 \in \theta_i.D, \forall y \in \theta_j.D, \#\theta_i \neq \#\theta_j :$

$$x_1 \in \theta_i.O_{x_2}^{\mathrm{A}} \wedge \star y \in U_{x_2}^{\mathcal{E}} \Rightarrow \star x_1 \in O_y^{\mathcal{A}} \tag{6.6}$$

$\forall \star y_1 \in U_x^{\mathcal{E}} - \{\star x\}, \forall y_2 \in \gamma^\star(\star y_1).D :$

$$y_1 \in \gamma^\star(\star y_1).O_{y_2}^{\mathrm{A}} \Rightarrow \star x \in O_{y_2}^{\mathcal{A}} \tag{6.7}$$

$\forall x \in \theta_i.D, \forall y \in \theta_j.D, \forall z \in \theta_r.D, \#\theta_i, \#\theta_j \#\theta_r$ *sind paarweise verschieden* :

$$\star x \in O_y^{\mathcal{A}} \wedge \star y \in U_z^{\mathcal{E}} \Rightarrow \star x \in O_z^{\mathcal{A}} \tag{6.8}$$

$\forall x \in \theta_i.D, \forall y \in \theta_j.D, \forall z \in \theta_r.D, \#\theta_i, \#\theta_j \#\theta_r$ *sind paarweise verschieden* :

$$\star x \in U_y^{\mathcal{E}} \wedge \star y \in O_z^{\mathcal{A}} \Rightarrow \star x \in O_z^{\mathcal{A}} \tag{6.9}$$

$\forall x_1, x_2 \in \theta_i.D, \forall y_1, y_2 \in \theta_j.D, \#\theta_i \neq \#\theta_j :$

$$x_1 \in \theta_i.O_{x_2}^{\mathrm{A}} \wedge \star y_1 \in U_{x_1}^{\mathcal{E}} \wedge \star y_2 \in U_{x_2}^{\mathcal{E}} \Rightarrow \star y_1 \in O_{y_2}^{\mathrm{A}} \tag{6.10}$$

$\forall x \in \theta_i.D, \forall y_1, y_2 \in \theta_j.D, \#\theta_i \neq \#\theta_j :$

$$\star y_1 \in O_x^{\mathcal{A}} \wedge \star y_2 \in U_x^{\mathcal{E}} \Rightarrow \star y_1 \in O_{y_2}^{\mathrm{A}} \tag{6.11}$$

$\forall x \in \theta_i.D, \forall y_1, y_2 \in \theta_j.D, \#\theta_i \neq \#\theta_j :$

$$\star x \in O_{y_2}^{\mathcal{A}} \wedge \star y_1 \in U_x^{\mathcal{E}} \Rightarrow \star y_1 \in O_{y_2}^{\mathrm{A}} \tag{6.12}$$

$$\forall x, y, z \in \mathbf{D} : (\star y \in \Re_x \vee \#y \in \gamma^\star(\star y).\mathrm{R}_x) \wedge \star z \in U_x^{\mathcal{E}} \Rightarrow \star y \in \Re_z \tag{6.13}$$

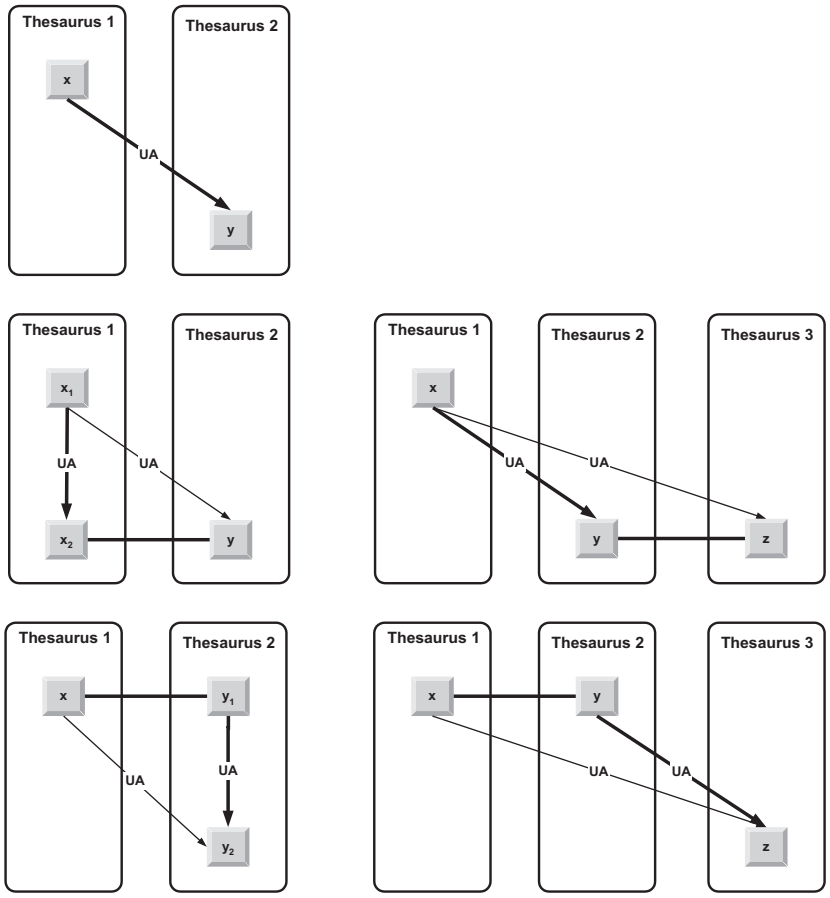

Abbildung 6.6: Thesaurusübergreifende Abstraktionsober-/-unterbegriffspaare. Die mit einer dünnen Linie verbundenen Ober-/Unterbegriffe werden durch die dicker eingezeichneten Abstraktionsober-/-unterbegriffspaare und durch die Äquivalenzpaare impliziert.

$$\forall x_1, x_2 \in \theta_i.D, \forall y_1, y_2 \in \theta_j.D, \#\theta_i \neq \#\theta_j : \\ \#x_1 \in \gamma^\star(\star x_1).\mathrm{R}_{x_2} \wedge \star y_1 \in U^{\mathcal{E}}_{x_1} \wedge \star y_2 \in U^{\mathcal{E}}_{x_2} \Rightarrow \star y_1 \in \mathrm{R}_{y_2} \quad (6.14)$$

$$\forall x \in \theta_i.D, \forall y_1, y_2 \in \theta_j.D, \#\theta_i \neq \#\theta_j : \\ \star y_1 \in \mathcal{R}_x \wedge \star y_2 \in U^{\mathcal{E}}_x \Rightarrow \star y_1 \in \mathrm{R}_{y_2} \quad (6.15)$$

$$U^{\mathcal{E}}_x \cap \overline{U^{\mathcal{E}}_x} = U^{\mathcal{C}}_x \cap \overline{U^{\mathcal{C}}_x} = O^{\mathcal{A}}_x \cap \overline{O^{\mathcal{A}}_x} = O^{\mathcal{P}}_x \cap \overline{O^{\mathcal{P}}_x} = R_x \cap \overline{\mathcal{R}_x} = \emptyset \quad (6.16)$$

$$\forall x \in \mathbf{D} : ((\gamma^\star(\star x).O^{\mathrm{A}}_x \neq \emptyset) \vee (\gamma^\star(\star x).O^{\mathrm{P}}_x \neq \emptyset)) \Leftrightarrow Q_{\Im_x} = \emptyset \quad (6.17)$$

Diese zentrale Definition sei nun näher erläutert. Die Mengen $U^{\mathcal{E}}_x, U^{\mathcal{C}}_x, O^{\mathcal{A}}_x, O^{\mathcal{P}}_x, \mathcal{R}_x, \overline{U^{\mathcal{E}}_x}, \overline{U^{\mathcal{C}}_x}, \overline{O^{\mathcal{A}}_x}, \overline{O^{\mathcal{P}}_x}, \overline{\mathcal{R}_x}$ und $Q_{\Im_x}$ werden anhand des bereits in Abschnitt 5.1.2, S. 59f, dargelegten Verständnisses von Abstraktions- und Bestandsoberbegriffen, Verwandten Begriffen, Äquivalenzbeziehungen und thematischen Gruppen von einem oder mehreren Experten festgelegt. Sie stellen das Ergebnis einer *idealen* Thesaurusföderation dar, in der *alle relevanten* Beziehungen etabliert sind und zusätzlich *alle relevanten nicht zu etablierenden*

Beziehungen als abgelehnte Beziehungen aufgeführt sind. Ein solches ideales Ergebnis, das von allen Experten als einziges Ergebnis akzeptiert wird, wird in der Praxis aufgrund der fehlenden Exaktheit der Bedeutung von Begriffen, der Komplexität der Begriffsstrukturen und der nicht immer eindeutigen Entscheidung der Relevanz nicht zu erreichen sein. Im Rahmen der Kapitel 7ff unserer Arbeit versuchen wir, weitreichende Unterstützung für die Konstruktion von Thesaurusföderationen anzubieten, die einer idealen Thesaurusföderation möglichst nahe kommen.

Im Gegensatz zu den Komponententhesauri werden für Thesaurusföderationen über $U_x^{\mathcal{E}}$ und $U_x^{\mathcal{C}}$ äquivalente Begriffe und Benutze-Kombination-Begriffe (jeweils statt Benennungen) festgelegt. Ausgedrückt werden diese Beziehungen durch Beziehungen zwischen den Deskriptoren. Dabei muss jeder Begriff ungleich x, der als äquivalenter Begriff angegeben wird, aus einem vom Thesaurus von x verschiedenen Thesaurus stammen (die Thesauri der äquivalenten Begriffe sind paarweise disjunkt; vgl. Bedingung 6.1). Ein Begriff wird als zu sich selbst äquivalent betrachtet, d.h. $\star x \in U_x^{\mathcal{E}}$. Bedingung 6.2 gilt aufgrund der Transitivität des Äquivalenzbegriffes. Für die in $U_x^{\mathcal{C}}$ angegebenen Benutze-Kombination-Begriffe gilt, dass wenn hier Begriffe aufgeführt werden, mindestens zwei Begriffe, die nicht ein Abstraktionsober-/-unterbegriffspaar sind[6], aus einem anderen Thesaurus stammen, die dann in Kombination für den Begriff x zu verwenden sind (vgl. Bedingungen 6.3 und 6.4). Sind Begriffe aus verschiedenen Thesauri aufgeführt, ist die jeweilige Kombination der Begriffe aus *einem* Thesaurus äquivalent zu x. Bedingung 6.5 besagt, dass zu einem Begriff aus einem Thesaurus θ_i *entweder* ein äquivalenter Begriff in einem von diesem verschiedenen Thesaurus θ_j *oder* eine Menge von Benutze-Kombination-Begriffen in diesem Thesaurus θ_j angegeben werden kann, nicht aber beides gleichzeitig.

Für die Mengen $\widetilde{U_x^{\mathcal{E}}}$, $\widetilde{U_x^{\mathcal{C}}}$, $\overline{U_x^{\mathcal{E}}}$ und $\overline{U_x^{\mathcal{C}}}$ gelten die Bedingungen 6.1 und 6.3 entsprechend übertragen ebenfalls. Wir führen sie zur Vereinfachung aber nicht getrennt auf.

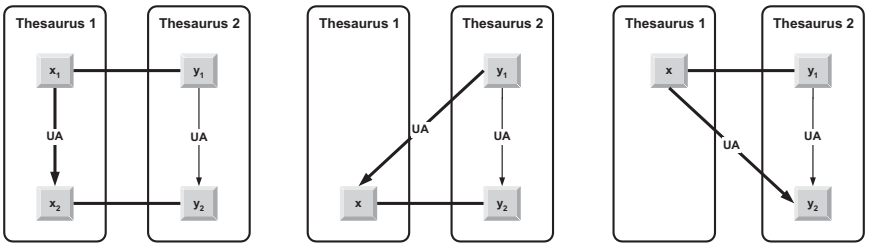

Abbildung 6.7: Implizierte Intra-Thesaurus-Abstraktionsober-/-unterbegriffspaare.

Die Mengen $O_x^{\mathcal{A}}$, $O_x^{\mathcal{P}}$, $\mathcal{R}_x$ und Q_x entsprechen von ihrer Semantik her den aus den Komponententhesauri bekannten Mengen bis darauf, dass hier jeweils Elemente aus anderen Thesauri (also solchen, die nicht x enthalten) aufgeführt werden. Allerdings gilt für $O_x^{\mathcal{A}}$ und $O_x^{\mathcal{P}}$ zu berücksichtigen, dass thesaurusübergreifende Abstraktionsober-/-unterbegriffspaare von anderen Ober-/Unterbegriffspaaren und von Äquivalenzpaaren impliziert werden können (vgl. Abbildung 6.6). Diese implizierten Beziehungen müssen ebenfalls explizit aufgeführt werden (vgl. Bedingung 6.6 und Abbildung 6.6 mitte links, Bedingung 6.7 und Abbildung 6.6 unten links, Bedingung 6.8 und Abbildung 6.6 mitte rechts, Bedingung 6.9 und Abbildung 6.6 unten rechts).

Durch Inter-Thesaurus-Beziehungen können auch Intra-Thesaurus-Beziehungen impliziert werden, die ebenfalls explizit vermerkt werden (vgl. Abbildung 6.7 links und Bedingung 6.10, Abbildung 6.7 mitte und Bedingung 6.11 sowie Abbildung 6.7 rechts und Bedingung 6.12).

[6]Durch die Kombination eines direkten oder indirekten Ober-/Unterbegriffspaares gewänne der Unterbegriff keine weiteren Eigenschaften und der Umfang des Oberbegriffs würde nicht erweitert.

Für $O_x^{\mathcal{P}}$ gelten die Bedingungen 6.6 bis 6.12 entsprechend, auf eine erneute Aufführung wurde verzichtet.

Auch für $\mathcal{R}_x$ gilt, dass durch Äquivalenzpaare thesaurusübergreifende und thesaurusinterne Verwandtschaftspaare impliziert werden, die ebenfalls explizit ausgedrückt werden (vgl. Bedingung 6.13 und Bedingungen 6.14, 6.15).

Bemerkung 6.2 *Wie wir gezeigt haben, können auch Intra-Thesaurus-Beziehungen impliziert werden. Da aufgrund der Autonomie der Komponententhesauri diese direkt aber nicht verändert werden können, werden diese Intra-Thesaurus-Beziehungen ebenfalls als Bestandteil des Integrationswissens gehandhabt. Das explizite Auffinden von zusätzlichen Intra-Thesaurus-Beziehungen wird jedoch nicht als Aufgabe der Thesaurusintegration betrachtet, so dass innerhalb des Integrationswissens ausschließlich aufgrund von Inter-Thesaurus-Beziehungen implizierte Intra-Thesaurus-Beziehungen existieren.*

Abgelehnte Beziehungen können, wenn dies für den weiteren Integrationsprozess relevant erscheint, in eigenen Mengen mitgeführt werden. Da bei weiteren Integrationsschritten die Begründung für die Ablehnung von Beziehungen ersichtlich sein soll, wird diese ebenfalls in die Mengen $\widetilde{U_x^{\mathcal{E}}}$, $\widetilde{U_x^{\mathcal{C}}}$, $\widetilde{O_x^{\mathcal{A}}}$, $\widetilde{O_x^{\mathcal{P}}}$ und $\widetilde{\mathcal{R}_x}$ aufgenommen. Um einfacher auf abgelehnte Elemente zugreifen zu können, definieren wir zusätzlich Mengen, die diese Begründung nicht weiter mitführen. Diese Mengen $\overline{U_x^{\mathcal{E}}}$, $\overline{U_x^{\mathcal{C}}}$, $\overline{O_x^{\mathcal{A}}}$, $\overline{O_x^{\mathcal{P}}}$ und $\overline{\mathcal{R}_x}$ können mit der Projektionsfunktion ι abgeleitet werden.

Die Schnittmengen der etablierten und der abgelehnten Inter-Thesaurus-Beziehungen sind jeweils leer (vgl. Bedingung 6.16), d.h. eine Beziehung kann nicht zugleich etabliert und abgelehnt sein.

Alle Begriffe, die in Komponententhesauri Topterme sind, können Föderierten Gruppen zugeordnet werden (vgl. Bedingung 6.17). Eine Einschränkung auf Topterme innerhalb der Föderation wird nicht getroffen, damit die Flexibilität bzgl. einer Auswahl zu betrachtender Komponententhesauri nicht beschränkt wird.

Nicht-Deskriptoren aus den Komponententhesauri tauchen in der Definition 6.11 nicht auf. Dies bedeutet, dass für Nicht-Deskriptoren keine direkten Beziehungen zu Benennungen in anderen Komponententhesauri hergestellt werden. Dies gilt insbesondere auch für Nicht-Deskriptoren, die in einem Komponententhesaurus in einer Benutze-Kombination-Beziehung stehen und somit einen Begriff repräsentieren. Da innerhalb des Komponententhesaurus kein eigener Deskriptor eingeführt wurde, betrachten wir diesen Begriff sogar innerhalb des Komponententhesaurus als weniger bedeutend und halten eine Berücksichtigung als eigenständigen Begriff innerhalb der Föderation für nicht erforderlich.

Betrachten wir Begriffe im Zusammenhang mit Thesaurusföderationen, sprechen wir von Föderierten Begriffen.

Definition 6.12 *Ein* Föderierter Begriff F_d *wird repräsentiert durch eine Teilmenge von* $\mathbf{B}$*, wobei* d *einer der den Föderierten Begriff repräsentierenden Deskriptoren sei, und es gilt*

$$\begin{aligned} F_d \quad = \quad & U_d^{\mathcal{E}} \cup \{\star u : u \in \mathbf{B} - \mathbf{D} \wedge \star y \in U_d^{\mathcal{E}} \wedge \gamma(y).U_u^{\mathrm{E}} = \{\#y\}\} \cup \\ & U_d^{\mathcal{C}} \cup \{\star u : u \in \mathbf{B} - \mathbf{D} \wedge \star y \in U_d^{\mathcal{C}} \wedge \gamma(y).U_u^{\mathrm{E}} = \{\#y\}\} \end{aligned}$$

Ein Föderierter Begriff wird somit durch mindestens einen Deskriptor aus einem Komponententhesaurus oder dem Thesaurus der Ergänzenden Begriffe repräsentiert. Hinzu kommen alle Benennungen aus allen Thesauri, die zu diesem Deskriptor äquivalent sind. Dabei kann es sich

sowohl um Deskriptoren als auch um Nicht-Deskriptoren handeln. Ebenso kommen alle in einer in Kombination zu benutzenden Begriffe hinzu.

6.3.4 Konfliktmarkierungen

Die strikten Bedingungen eines Thesaurus (z.B. Zyklenfreiheit der Abstraktionsrelation) können in einer Thesaurusföderation nicht immer eingehalten werden. Um mit Verstößen gegen solche Bedingungen umgehen zu können, werden so genannte Konfliktmarkierungen eingeführt. Da das Konzept der Konfliktmarkierungen im formalen Modell für Thesauri nicht erforderlich war, erläutern wir es an dieser Stelle.

Ziel der Konfliktmarkierung ist es, Inkosistenzen zwischen den Komponententhesauri, die bei der Integration erkannt werden, aufgrund der Autonomie der Komponententhesauri zuzulassen, gleichzeitig aber auch als Konflikte explizit zu markieren. Solche Konfliktmarkierungen sind Basis einer möglichen Konfliktauflösung während der Verwendung der Thesaurusföderation.

Eine Konfliktmarkierung ist ein Tupel $k = (t, v, r_1, r_2, s)$ wobei t eine Zeichenfolge ist und den Konfliktttyp bezeichnet, $v \subset \mathbf{D} \times \mathbf{D}$ die Menge aller Paare von Deskriptoren, die den Konflikt verursachen, $r_1, r_2 \subset v$ die Mengen der Deskriptorpaare, die in Konflikt miteinander stehen, und $s \subset v$ die Deskriptorpaare, zwischen denen die Beziehungen entfernt werden müssten, um den Konflikt zu beseitigen (Konfliktauflösung durch Beziehungsentfernung). Zwischen den Deskriptoren der in s aufgeführten Paare ist eine Intra- oder eine Inter-Thesaurus-Beziehung etabliert. Handelt es sich um eine Intra-Thesaurus-Beziehung, gilt diese durch die Konfliktmarkierung innerhalb der Föderation als nicht vorhanden zu betrachten. Bei einer Inter-Thesaurus-Beziehung ist die Beziehung selbst und, wenn es sich um eine implizierte Beziehung (vgl. Abbildung 6.6) handelt, die diese Beziehung implizierende Inter-Thesaurus-Äquivalenz- bzw. Intra-Thesaurus-Beziehung zu entfernen sowie ggf. durch andere Beziehungen zu ersetzen.

6.3.5 Implizierte Intra-Thesaurus-Relationen

Wie wir in Abschnitt 6.3.3, S. 86, gezeigt haben, werden durch die Integration auch Beziehungen zwischen Begriffen *eines* Thesaurus impliziert. Entsprechend den Intra-Thesaurus-Hierarchie-Relationen und der Intra-Thesaurus-Assoziationsrelation werden daher die Relation der implizierten Intra-Thesaurus-Abstraktionsbeziehungen A (kurz: Implizierte Intra-Thesaurus-Abstraktionsrelation), die Relation der implizierten Intra-Thesaurus-Bestandsbeziehungen P (kurz: Implizierte Intra-Thesaurus-Bestandsrelation) sowie die Relation der implizierten Intra-Thesaurus-Assoziationsbeziehungen V (kurz: Implizierte Intra-Thesaurus-Assoziationsrelation) anhand der für jeden Deskriptor $x \in \mathbf{D}$ gegebenen Mengen $O_x^{\mathsf{A}}, O_x^{\mathsf{P}}$ und R_x definiert.

6.3.6 Inter-Thesaurus-Relationen

Wie bei den Intra-Thesaurus-Relationen, wie wir die in Abschnitt 5.1.3 definierten Relationen innerhalb eines Thesaurus zur besseren Unterscheidbarkeit nennen, werden die im vorangegangenen Abschnitt vorgestellten Beziehungen zwischen verschiedenen Thesauri, sofern sie gleichartig sind, zu zweistelligen Inter-Thesaurus-Relationen zusammengefasst.

6.3.6.1 Inter-Thesaurus-Äquivalenzrelation

Definition 6.13 *Die* Inter-Thesaurus-Äquivalenzrelation *ist die Relation, die gleichwertige Deskriptoren, also Deskriptoren die innerhalb der Thesaurusföderation* einen *Begriff repräsentieren, aus verschiedenen Thesauri zu einer Äquivalenzklasse zusammenführt. Formal ist die Inter-Thesaurus-Äquivalenzrelation* $\mathcal{E}$ *eine Teilmenge von* $\mathbf{D} \times \mathbf{D}$*, die folgende Bedingung erfüllt:*

$$\mathcal{E} = \{(x, y) : x \in \mathbf{D} \text{ und } \star\, y \in U_x^{\mathcal{E}}\}$$

Bemerkung 6.3 *Die Inter-Thesaurus-Äquivalenzrelation unterscheidet sich von der Intra-Thesauru-Äquivalenzsrelation zum einen dadurch, dass sie nur über Deskriptoren statt über Deskriptoren und Nicht-Deskriptoren definiert wird. Zum anderen besitzt sie aufgrund der Definition von* $U_x^{\mathcal{E}}$ *die Eigenschaften der Reflexivität, Symmetrie und Transitivität, so dass sie auch im mathematischen Sinne eine Äquivalenzrelation ist.*

6.3.6.2 Inter-Thesaurus-Benutze-Kombination-Relation

Definition 6.14 *Die* Inter-Thesaurus-Benutze-Kombination-Relation *wird verwendet, um Äquivalenzen zwischen einem Begriff eines Thesaurus und der konjunktiven Verknüpfung von zwei oder mehr Begriffen eines anderen Thesaurus auszudrücken. Formal ist die Benutze-Kombination-Relation* $\mathcal{C}$ *eine Teilmenge von* $\mathbf{D} \times \mathbf{D}$*, die die folgende Bedingung erfüllt:*

$$\mathcal{C} = \{(x, y) : x \in \mathbf{D} \wedge \star y \in U_x^{\mathcal{C}}\}$$

6.3.6.3 Inter-Thesaurus-Abstraktionsrelation

Definition 6.15 *Die* Inter-Thesaurus-Abstraktionsrelation *fasst gerichtete Beziehungen zusammen, die die Über- bzw. Unterordnung von Begriffen aus* verschiedenen *Thesauri im Sinne des allgemeineren bzw. spezifischeren Begriffs darstellen. Wie bei der Intra-Thesaurus-Abstraktionsrelation besitzt der untergeordnete Begriff (Abstraktionsunterbegriff) alle Merkmale des übergeordneten Begriffs (Abstraktionsoberbegriff) und zusätzlich mindestens ein weiteres spezifizierendes Merkmal.*

Formal ist die Inter-Thesaurus-Abstraktionsrelation $\mathcal{A}$ *die kleinste Teilmenge von* $\mathbf{D} \times \mathbf{D}$*, für die gilt*[7]*:*

$$\mathcal{A} = \{(x, y) : x \in \mathbf{D} \wedge \star y \in O_x^{\mathcal{A}}\}$$

Wie die Intra-Thesaurus-Abstraktionsrelation ist auch die Inter-Thesaurus-Abstraktionsrelation transitiv und nicht reflexiv (x und y stammen aufgrund der Definition von $O_x^{\mathcal{A}}$ *aus verschiedenen Thesauri). Anders als bei der Intra-Thesaurus-Abstraktionsrelation wird die Zyklenfreiheit über die Kombination von Inter- und Intra-Thesaurus-Abstraktionsbeziehungen nicht strikt gefordert. Wird gegen die Zyklenfreiheit verstoßen, muss dieser Verstoß aber als Konflikt vermerkt sein. Ein Pfad über die Inter-, Intra-* und *implizierten Intra-Thesaurus-Abstraktionsrelationen von Deskriptor* x_1 *zu Deskriptor* x_l *ist eine Folge von Deskriptoren* $x_1, x_2, \ldots, x_l \in \mathbf{B}$*, so dass* $l \geq 2$ *und* $\forall i \in [1, l-1]$ *gilt:* $(x_{i+1}, x_i) \in \mathcal{A} \vee (x_{i+1}, x_i) \in \mathsf{A} \vee (x_{i+1}, x_i) \in \gamma(x_i).\mathsf{A}$*. Ein solcher Pfad wird als* $x_1 \overset{\mathrm{A}^\star}{\hookrightarrow} x_l$ *notiert.* $(x_i, x_j) \in x_1 \overset{\mathrm{A}^\star}{\hookrightarrow} x_l$ *bedeutet, dass* x_i*,* x_j *ein Paar von Deskriptoren*

[7] Der erste Parameter ist untergeordneter Begriff des zweiten Begriffs.

ist, das in dem Pfad $x_1 \overset{\mathrm{A}^\star}{\hookrightarrow} x_l$ *vorkommt. Ein Pfad, in dem* x *vorkommt, wird auch mit* $x_1 \overset{\mathrm{A}^\star_x}{\hookrightarrow} x_l$ *bezeichnet.*

Die Eigenschaft der Markierung bei Verstoß gegen die Zyklenfreiheit der Inter- und Intra-Thesaurus-Abstraktionsrelationen *kann also wie folgt notiert werden:*

$\forall x, y \in \mathbf{D}$:

$$x \overset{\mathrm{A}^\star}{\hookrightarrow} y \Rightarrow (\star x \neq \star y) \vee [(\star x = \star y) \wedge \exists k = (t, v, r_1, r_2, s) \in K \ \textit{mit} \quad (6.18)$$

$$\begin{aligned} t &= \text{„}\textit{Abstraktionszyklus}\text{“} \\ v &= \{(x_j, x_i) : (x_i, x_j) \in x \overset{\mathrm{A}^\star}{\hookrightarrow} y\} \\ r_1 &= \{(x_j, x_i) : (x_i, x_j) \in x \overset{\mathrm{A}^\star}{\hookrightarrow} y\} \cap \mathcal{A} \\ r_2 &= v - r_1 \\ s &= \{(x_m, x_n) \in v\}] \end{aligned}$$

Des Weiteren gilt die Eigenschaft der Markierung bei Verstoß gegen die Redundanzfreiheit der Inter- und Intra-Thesaurus-Abstraktionsrelationen*:*

$\forall x_1, x_l, y \in \mathbf{D}, x_1, y, x_l$ *sind paarweise verschieden* :

$$(((x_1, x_l) \in \gamma(x_1).\mathrm{A}) \vee ((x_1, x_l) \in \mathcal{A})) \wedge (x_1 \overset{\mathrm{A}^\star_y}{\hookrightarrow} x_l) \Rightarrow \exists k = (t, v, r_1, r_2, s) \in K \ \textit{mit} \quad (6.19)$$

$$\begin{aligned} t &= \text{„}\textit{Abstraktionsredundanz}\text{“} \\ v &= \{(x_l, x_1)\} \cup \{(x_j, x_i) : (x_i, x_j) \in x_1 \overset{\mathrm{A}^\star_y}{\hookrightarrow} x_l\} \\ r_1 &= \{(x_l, x_1)\} \\ r_2 &= v - r_1 \\ s &= \{(x_m, x_n) \in v\} \end{aligned}$$

Abbildung 6.8 zeigt ein Beispiel für einen Zyklus, der durch gegensätzliche Hierarchien in den Thesauri ensteht. Konfliktverursacher sind die Paare $(x_2, y_1), (y_1, y_2)$ und (y_2, x_2), in Konflikt stehen das erst- und letztgenannte Paar mit dem zweiten Paar, zur Konfliktauflösung (Auftrennung des Zyklus) muss zwischen einem der Paare die Beziehung „entfernt“ werden. Welches Paar dies sein sollte, wird in s festgelegt.

Entsprechend den Markierungen bei Zyklen in der Abstraktionsrelation werden auch redundante Abstraktionsbeziehungen markiert. Abbildung 6.9 zeigt Beispiele für redundante Beziehungen, d.h., zwischen zwei Begriffen gibt es eine direkte Abstraktionsbeziehung (jeweils zwischen y_1 und y_2) sowie einen Pfad über mindestens einen weiteren Begriff (x_2 bzw. x_2 und x_3). Die direkte Abstraktionsbeziehung wird als im Konflikt mit den indirekten Abstraktionsbeziehungen stehend markiert, zur Konfliktauflösung ist die Entfernung der direkten Beziehung oder einer Beziehung zwischen den Elementen des Pfades erforderlich. Wie aus den Beispielen in der Abbildung ersichtlich wird, können solche redundanten Beziehungen durch das Etablieren von

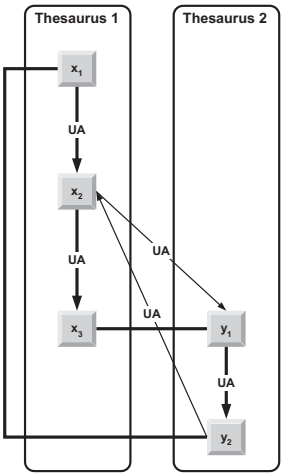

Abbildung 6.8: Zyklus in der Abstraktionsrelation, verursacht durch gegensätzliche Beziehungen in den Komponententhesauri.

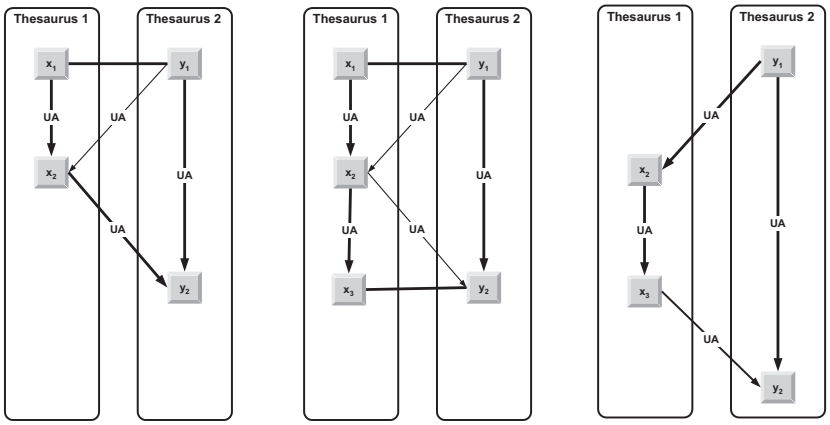

Abbildung 6.9: Beispiele für Redundanz in der Abstraktionsrelation.

Inter-Thesaurus-Äquivalenzbeziehungen und Inter-Thesaurus-Abstraktionsbeziehungen entstehen.

Zusätzlich gilt die Forderung nach identischen Abstraktionsniveaus. Das bedeutet, dass verschiedene Pfade innerhalb der Abstraktionsrelation von einem Begriff zu einem untergeordneten Begriff gleich lang sein sollen. Aufgrund von Inter-Thesaurus-Beziehungen feststellbare Abstraktionsniveaudifferenzen können in verschiedenen Fällen auf eine Verletzung der Redundanzfreiheit der Hierarchierelation zurückgeführt werden (vgl. Abbildung 6.10, rechts. Die Rückführung auf – sogar mehrfache – Verletzung der Redundanzfreiheit gelingt dank der Berücksichtigung der implizierten Beziehungen; aufgedeckt werden hier durch die Integration unterschiedliche Abstraktionsniveaus innerhalb eines Thesaurus). Es gibt aber auch Fälle, in denen die Redundanzfreiheit nicht verletzt wird, aber verschiedene Abstraktionsniveaus festgestellt werden (vgl. Abbildung 6.10, links).

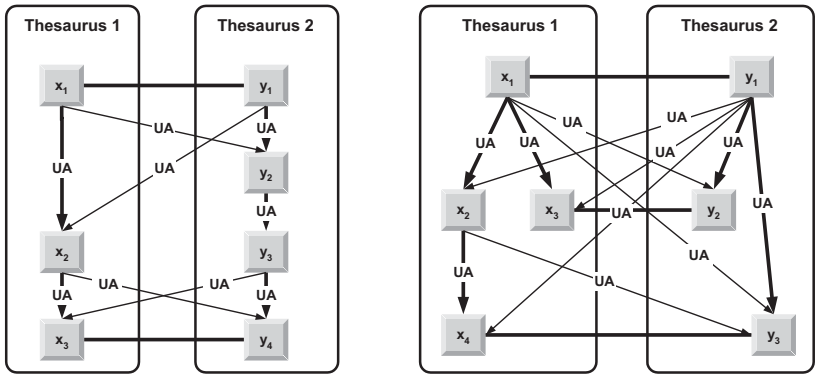

Abbildung 6.10: Abstraktionsniveaudifferenz über mehrere Hierarchiestufen (links) und innerhalb eines Komponententhesaurus (rechts).

Da redundante Abstraktionsbeziehungen immer auch verschiedene Abstraktionsniveaus bedeuten, berücksichtigen wir diesen Fall nicht erneut als Verletzung der Forderung nach identischen Abstraktionsniveaus. Wir können somit für die Eigenschaft der *Markierung bei verschiedenen Abstraktionsniveaus* notieren:

$\forall x_2, x_3, y_1, y_3 \in \mathbf{D}, x_2, x_3, y_1, y_3$ sind paarweise verschieden:

$$|y_1 \overset{\mathrm{A}^\star_{x_2}}{\hookrightarrow} x_3| < |y_1 \overset{\mathrm{A}^\star_{y_3}}{\hookrightarrow} x_3| \wedge (x_3, x_2) \in \gamma(x_3).\mathrm{A} \wedge (x_3, y_3) \in \mathcal{A} \Rightarrow \exists k = (t, v, r_1, r_2, s) \in K \text{ mit} \quad (6.20)$$

$$\begin{aligned} t &= \text{„Abstraktionsniveaudifferenz“} \\ v &= \{(z_j, z_i) : (z_i, z_j) \in y_1 \overset{\mathrm{A}^\star_{x_2}}{\hookrightarrow} x_3 \vee (z_i, z_j) \in y_1 \overset{\mathrm{A}^\star_{y_3}}{\hookrightarrow} x_3\} \\ r_1 &= \{(z_j, z_i) : (z_i, z_j) \in y_1 \overset{\mathrm{A}^\star_{x_2}}{\hookrightarrow} x_3\} \\ r_2 &= v - r_1 \\ s &= \emptyset \vee s = \{(a, b) : (a, b) \in y_1 \overset{\mathrm{A}^\star_{y_3}}{\hookrightarrow} x_3 \wedge a \text{ oder } b \text{ sollen entfernt werden}\} \end{aligned}$$

Wird ein Abstraktionsniveauunterschied festgestellt, gilt dieser Unterschied auch für alle direkten und indirekten Abstraktionsunterbegriffe. Damit diese Unterschiede nicht zusätzlich markiert

werden, werden die Begriffe, über die die verschieden langen Pfade gehen, so eingeschränkt, dass einer davon direkter Oberbegriff innerhalb der Inter-Thesaurus-Abstraktionsrelation und der andere direkter Oberbegriff innerhalb der Intra-Thesaurus-Abstraktionsrelation ist.

Abstraktionsniveauunterschiede können durch Einführen von Ergänzenden Begriffen in den kürzeren Pfad aufgelöst werden. In diesem Fall ist s die leere Menge – allerdings werden die durch den oder die Ergänzenden Begriffe redundant gewordenen Abstraktionsbeziehungen entsprechend markiert. Wurde kein Ergänzender Begriff zur Konfliktauflösung eingeführt, werden entsprechend der Längenunterschiede der Pfade Deskriptoren aus dem längeren Pfad entfernt. Diese Deskriptoren werden markiert, in dem in die Menge s alle Tupel, die zu entfernende Deskriptoren enthalten, aufgenommen werden.

Wie für die Intra-Thesaurus-Abstraktionsrelation wird auch für die Inter-Thesaurus-Abstraktionsrelation keine Beschränkung für die mögliche Anzahl über- bzw. untergeordneter Abstraktionsbegriffe eines Begriffes vorgegeben. Ein Begriff, der weder in der Inter-Thesaurus-Abstraktionsrelation noch in der Intra-Thesaurus-Abstraktionsrelation einen übergeordneten Begriff besitzt, wird *Topterm der Abstraktionsrelation* genannt.

$\mathcal{A}^{-1}$ sei die inverse Relation von $\mathcal{A}$. Die *transitive Hülle* von $\mathcal{A}$ bzw. $\mathcal{A}^{-1}$ wird notiert als

$$\begin{aligned} \mathcal{A}_x^+ &= \{y \in \mathbf{D} : (x,y) \in \mathcal{A}\} \\ (\mathcal{A}_x^{-1})^+ &= \{y \in \mathbf{D} : (y,x) \in \mathcal{A}\} \end{aligned}$$

6.3.6.4 Inter-Thesaurus-Bestandsrelation

Definition 6.16 *Die* Inter-Thesaurus-Bestandsrelation (partitive Relation) *ist eine Relation, die gerichtete Beziehungen zusammenfasst, die die Über- bzw. Unterordnung von Begriffen aus verschiedenen Thesauri im Sinne des Ganzen bzw. eines Teiles darstellen. Dabei entspricht der untergeordnete Begriff (Teilbegriff) einem der Bestandteile des übergeordneten Begriffs (Verbandsbegriff), der ein Ganzes darstellt.*

Formal ist die Inter-Thesaurus-Bestandsrelation $\mathcal{P}$ die kleinste Teilmenge von $\mathbf{D} \times \mathbf{D}$, für die gilt[8]*:*

$$\mathcal{P} = \{(x,y) : x \in \mathbf{D} \wedge \star y \in O_x^{\mathcal{P}}\}$$

Die Inter-Thesaurus-Bestandsrelation ist wie die Intra-Thesaurus-Bestandsrelation nicht reflexiv und nicht transitiv. Wie bei der Inter-Thesaurus-Abstraktionsrelation muss bei einem Verstoß gegen die Zyklenfreiheit dieser Verstoß als Konflikt markiert sein. Ein Pfad in den Intra-, implizierten Intra- und Inter-Thesaurus-Bestandsrelationen von Deskriptor x_1 zu Deskriptor x_l ist eine Folge von Deskriptoren $x_1, x_2, \ldots, x_l$ so dass $l \geq 2$ und $\forall i \in [1, l-1] : (x_{i+1}, x_i) \in \mathcal{P} \vee (x_{i+1}, x_i) \in \mathsf{P} \vee (x_{i+1}, x_i) \in \gamma(x_i).\mathrm{P}$. Ein solcher Pfad wird als $x_1 \overset{\mathrm{P}^\star}{\hookrightarrow} x_l$ notiert. $(x_i, x_j) \in x_1 \overset{\mathrm{P}^\star}{\hookrightarrow} x_l$ bedeutet, dass x_i, x_j ein Paar von Deskriptoren ist, das in dem Pfad $x_1 \overset{\mathrm{P}^\star}{\hookrightarrow} x_l$ vorkommt.

Die Eigenschaft der Markierung bei Verstoß gegen die Zyklenfreiheit der Inter-Thesaurus-

[8] Der erste Parameter ist untergeordneter Begriff des zweiten Begriffs.

Bestandsrelation kann also wie folgt notiert werden:

$\forall x, y \in \mathbf{D}$:

$$x \overset{\mathrm{P}^\star}{\hookrightarrow} y \Rightarrow (\star x \neq \star y) \vee [(\star x = \star y) \wedge \exists k = (t, v, r_1, r_2, s) \in K \ mit \quad (6.21)$$

$$\begin{aligned} t &= \text{„}Bestandszyklus\text{“} \\ v &= \{(x_j, x_i) : (x_i, x_j) \in x_1 \overset{\mathrm{P}^\star}{\hookrightarrow} x_l\} \\ r_1 &= \{(x_j, x_i) : (x_i, x_j) \in x \overset{\mathrm{P}^\star}{\hookrightarrow} y\} \cap \mathcal{P} \\ r_2 &= v - r_1 \\ s &= \{(x_m, x_n) \in v\}] \end{aligned}$$

Die Markierung bei Verstößen gegen die Zyklenfreiheit der Bestandsrelation entspricht somit den Markierungen der Abstraktionsrelation. Redundanzfreiheit wird aufgrund der nicht vorhandenen Transitivität der Bestandsrelation nicht verlangt.

Für die mögliche Anzahl über- bzw. untergeordneter Bestandsbegriffe eines Begriffes wird keine Beschränkung vorgegeben. Ein Begriff, der weder in der Inter-Thesaurus-Bestandsrelation noch in der Intra-Thesaurus-Bestandsrelation einen übergeordneten Begriff besitzt, wird *Topterm der Bestandsrelation* genannt.

$\mathcal{P}^{-1}$ sei die inverse Relation von $\mathcal{P}$.

6.3.6.5 Inter-Thesaurus-Hierarchierelation

Definition 6.17 *Inter-Thesaurus-Hierarchiebeziehungen liegen vor, wenn zwei Begriffe aus verschiedenen Thesauri zueinander in einem Verhältnis der Über- bzw. Unterordnung stehen, unabhängig davon, ob es sich um eine Inter-Thesaurus-Abstraktionsbeziehung oder eine Inter-Thesaurus-Bestandsbeziehung handelt. Die* Inter-Thesaurus-Hierarchierelation $\mathcal{H}$ *ist somit die Vereinigung der Inter-Thesaurus-Abstraktionsrelation $\mathcal{A}$ und der Inter-Thesaurus-Bestandsrelation $\mathcal{P}$, für die gilt:*

$$\mathcal{H} = \{(x, y) : (x, y) \in \mathcal{A} \ oder \ (x, y) \in \mathcal{P}\}$$

Für die Inter-Thesaurus-Hierarchierelation wird wie für die diese Relation definierenden Abstraktions- und Bestandsrelation die Eigenschaft der Markierung von Verstößen gegen die Zyklenfreiheit gefordert.

Aufgrund der großen Ähnlichkeit zu den Zyklenmarkierungen der Abstraktions- bzw. Bestandsrelation (ersetze A durch H, „Abstraktionszyklus“ durch „Hierarchiezyklus“) verzichten wir auf die formale Notation der Markierung bei Verstoß gegen die Zyklenfreiheit der Inter- und Intra-Thesaurus-Hierarchierelationen.

Die Inter-Thesaurus-Hierarchierelation ist wie die Intra-Thesaurus-Hierarchierelation nicht transitiv (fehlende Transitivität der Inter-Thesaurus-Besandsrelation) und nicht reflexiv.

$\mathcal{H}^{-1}$ sei die inverse Relation von $\mathcal{H}$.

6.3.6.6 Inter-Thesaurus-Assoziationsrelation

Definition 6.18 *Die* Inter-Thesaurus-Assoziationsrelation *umfasst entsprechend der Intra-Thesaurus-Assoziationsrelation Beziehungen zwischen Begriffen, die als wichtig erscheinen, aber weder eindeutig hierarchischer Natur noch äquivalent sind. Formal ist die Inter-Thesaurus-Assoziationsrelation $\mathcal{V}$ die kleinste Teilmenge von $\mathbf{D} \times \mathbf{D}$ für die gilt*

$$\mathcal{V} = \{(x, y) : x \in \mathbf{D} \text{ und } \star y \in \mathcal{R}_x\}$$

Die Inter-Thesaurus-Assoziationsrelation ist aufgrund der Symmetrie der Verwandtschaftsbeziehung zwischen zwei Begriffen symmetrisch.

6.3.7 Relationsübergreifende Eigenschaften

Bereits bei der Einführung von Begriffen und Benennungen sowie den Inter-Thesaurus-Relationen wurde eine Reihe von geforderten Eigenschaften genannt. In diesem Abschnitt werden zusätzliche Eigenschaften aufgeführt, die relationsübergreifenden Charakter haben.

6.3.7.1 Paarweise Disjunktheit der Inter-Thesaurus-Relationen

Die Eigenschaft der *paarweisen Disjunktheit der Inter-Thesaurus-Relationen* fordert entsprechend der paarweisen Disjunktheit der Intra-Thesaurus-Relationen (vgl. Abschnitt 5.1.3.7, S. 65), dass die Inter-Thesaurus-Relationen $\mathcal{E}, \mathcal{C}, \mathcal{A}, \mathcal{A}^{-1}, \mathcal{P}, \mathcal{P}^{-1}$ und $\mathcal{V}$ paarweise disjunkt sind. Das bedeutet, dass zwei Deskriptoren, die durch eine Inter-Thesaurus-Relation miteinander in Beziehung stehen, durch keine andere Inter-Thesaurus-Relation miteinander in Beziehung stehen sollen.

Aus Bedingung 6.5 folgt bereits die Disjunktheit der Relationen $\mathcal{E}$ und $\mathcal{C}$. Zusätzlich muss also gelten:

$$\begin{aligned}\mathcal{E} \cap \mathcal{A} = \mathcal{E} \cap \mathcal{A}^{-1} = \mathcal{E} \cap \mathcal{P} = \mathcal{E} \cap \mathcal{P}^{-1} = \mathcal{E} \cap \mathcal{V} \\ = \mathcal{C} \cap \mathcal{A} = \mathcal{C} \cap \mathcal{A}^{-1} = \mathcal{C} \cap \mathcal{P} = \mathcal{C} \cap \mathcal{P}^{-1} = \mathcal{C} \cap \mathcal{V} \\ = \mathcal{A} \cap \mathcal{A}^{-1} = \mathcal{A} \cap \mathcal{P} = \mathcal{A} \cap \mathcal{P}^{-1} = \mathcal{A} \cap \mathcal{V} \\ = \mathcal{P} \cap \mathcal{P}^{-1} = \mathcal{P} \cap \mathcal{V} = \emptyset \end{aligned} \quad (6.22)$$

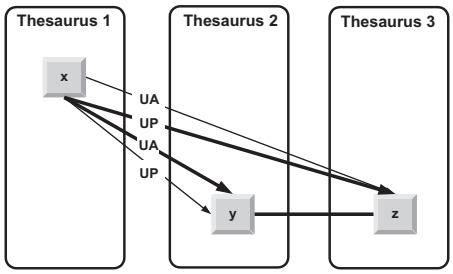

Abbildung 6.11: Verstoß gegen die paarweise Disjunktheit der Inter-Thesaurus-Relationen mit ausschließlich Inter-Thesaurus-Beziehungen als Verursachern.

In Situationen, in denen an einer Verletzung gegen diese Bedingung ausschließlich Inter-Thesaurus-Beziehungen beteiligt sind (vgl. Abbildung 6.11), wird die Konfliktauflösung durch

die Beseitigung einer der in Konflikt stehenden Beziehungen gefordert. Stehen jedoch Inter- und Intra-Thesaurus-Beziehungen in Konflikt (vgl. Abbildung 6.12) und wird der Inter-Thesaurus-Beziehung im Rahmen der Föderation mehr Gewicht zuteil als der Intra-Thesaurus-Beziehung, würde die Konfliktauflösung die Entfernung der Intra-Thesaurus-Beziehung erfordern. Da diese Entfernung aufgrund der Autonomie der Komponententhesauri nicht möglich ist, erfolgt eine Konfliktmarkierung:

$$\forall x_1, x_2, y_1, y_2 \in \mathbf{D}:$$
$$\star y_1 \in U_{x_1}^{\mathcal{E}} \wedge \star y_2 \in U_{x_2}^{\mathcal{E}} \wedge (x_2, x_1) \in A$$
$$\Rightarrow (y_2, y_1) \notin P \vee$$
$$[(y_2, y_1) \in P \wedge \exists k = (t, v, r_1, r_2, s) \in K \text{ mit} \quad (6.23)$$

$$\begin{aligned} t &= \text{„Beziehungstypdifferenz“} \\ v &= \{(x_1, y_1), (x_2, y_2), (x_2, x_1), (y_2, y_1)\} \\ r_1 &= \{(x_2, x_1)\} \\ r_2 &= \{(y_2, y_1)\} \\ s &= r_1 \vee r_2] \end{aligned}$$

Direkt in Konflikt stehen in diesem Fall die unterschiedlichen Beziehungen zwischen den äquivalenten Begriffen. Da die Inter-Thesaurus-Äquivalenzbeziehungen etabliert wurden, wird eine der Intra-Thesaurus-Beziehungen markiert, um Hinweise zur Konfliktauflösung während der Anwendung der Thesaurusföderation zu geben.

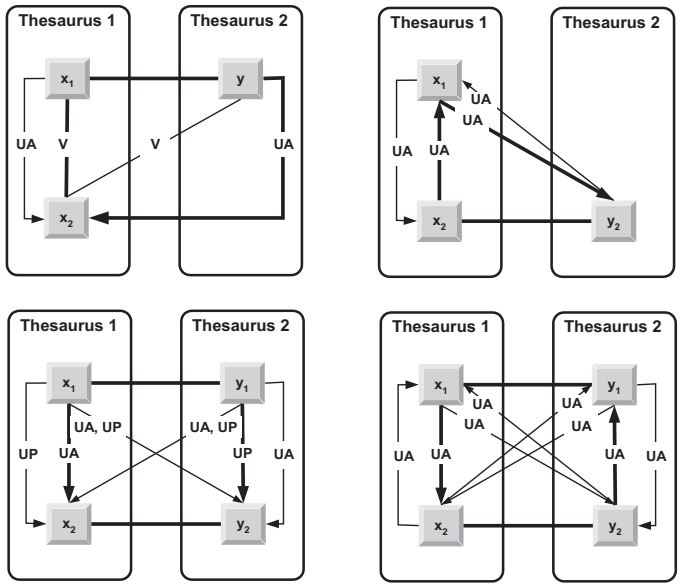

Abbildung 6.12: Beispiele für den Verstoß gegen die paarweise Disjunktheit der Inter-Thesaurus-Relationen mit Beteiligung von Intra-Thesaurus-Beziehungen.

Diese Eigenschaft gilt wie angegeben für das Relationspaar $(\mathcal{A}, \mathcal{P})$ sowie wie für die Paare $(\mathcal{A}, \mathcal{V})$, $(\mathcal{P}, \mathcal{V})$, $(\mathcal{A}, \mathcal{A}^{-1})$, $(\mathcal{A}, \mathcal{P}^{-1})$ sowie $(\mathcal{P}, \mathcal{P}^{-1})$. Aufgrund der großen Ähnlichkeit der Eigenschaften verzichten wir für die weiteren Fälle auf die ausführliche Formulierung.

Für Fälle wie in Abbildung 6.12 oben links dargestellt, ist es zusätzlich notwendig zu notieren:

$$\begin{aligned}\forall x_1, x_2, y_1 \in \mathbf{D}: \\ \star y_1 \in U^{\mathcal{E}}_{x_1} \wedge (\nexists y_2 \in \gamma^\star(\star y_1).\mathbf{D}: y_2 \in U^{\mathcal{E}}_{x_2}) \wedge (x_2, x_1) \in V \\ \Rightarrow (x_2, y_1) \notin \mathcal{A} \vee \\ [(x_2, y_1) \in \mathcal{A} \wedge \exists k = (t, v, r_1, r_2, s) \in K \text{ mit} \end{aligned} \quad (6.24)$$

$$\begin{aligned} t &= \text{„Beziehungstypdifferenz“} \\ v &= \{(x_1, y_1), (x_2, y_1), (x_2, x_1)\} \\ r_1 &= \{(x_2, x_1)\} \\ r_2 &= \{(x_2, y_1)\} \\ s &= r_1] \end{aligned}$$

Wiederum wird auf die Wiederholung der Notation für die weiteren Relationenpaare verzichtet. Entsprechend gilt diese Formulierung auch für die durch Abbildung 6.12 oben rechts repräsentierten Fälle.

Wird gegen die Bedingung der paarweisen Disjunktheit der Inter-Thesaurus-Relationen durch $\mathcal{A} \cap \mathcal{A}^{-1} \neq \emptyset$, $\mathcal{P} \cap \mathcal{P}^{-1} \neq \emptyset$ oder $\mathcal{A} \cap \mathcal{P}^{-1} \neq \emptyset$ verstoßen, ist dies zusätzlich als Konflikt gegen die Zyklenfreiheit der Inter-Thesaurus-Abstraktionsrelation, -Bestandsrelation bzw. Hierachierelation markiert (vgl. Abschnitte 6.3.6.3 bis 6.3.6.5).

6.3.7.2 Keine Assoziationsbeziehungen zwischen Schwesterknoten

Begriffe, die über eine Inter-Thesaurus-Äquivalenzrelation verbunden sind, sollen nicht gleichzeitig Schwesterknoten und assoziierte Begriffe sein. Ein Beispiel für einen Verstoß ist in Abbildung 6.2 (rechts), S. 75, dargestellt. Dieser Fall kann aufgrund der implizierten Beziehungen zurückgeführt werden auf den Fall, dass zwei Begriffe, die Schwestern innerhalb der Inter-Thesaurus-Hierarchierelation sind, zusätzlich in einer Assoziationsbeziehung stehen (vgl. Abbildung 6.13).

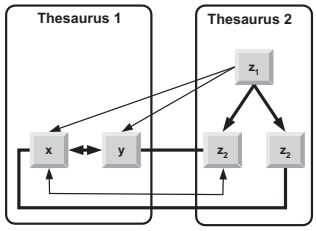

Abbildung 6.13: Assoziationsbeziehung zwischen Schwesterknoten der Inter-Thesaurus-Hierarchierelation.

Die Eigenschaft der *Markierung bei Assoziationsbeziehungen zwischen Schwesterknoten* wird wie folgt notiert:

$$\begin{aligned}\forall x, y \in \mathbf{D}: \#y \in R_x \vee \star y \in \mathcal{R}_x \\ \Rightarrow (O^{\mathcal{A}}_x \cap O^{\mathcal{A}}_y = O^{\mathcal{P}}_x \cap O^{\mathcal{P}}_y = O^{\mathcal{A}}_x \cap O^{\mathcal{P}}_y = O^{\mathcal{P}}_x \cap O^{\mathcal{A}}_y = \emptyset) \vee \\ [\exists z_1 \in \mathbf{D}: \star z_1 \in O^{\mathcal{A}}_x \cap O^{\mathcal{A}}_y \vee \star z_1 \in O^{\mathcal{P}}_x \cap O^{\mathcal{P}}_y \vee \\ \star z_1 \in O^{\mathcal{A}}_x \cap O^{\mathcal{P}}_y \vee \star z_1 \in O^{\mathcal{P}}_x \cap O^{\mathcal{A}}_y \end{aligned}$$

$$\wedge \exists k = (t, v, r_1, r_2, s) \in K \text{ mit} \tag{6.25}$$

$$\begin{aligned} t &= \text{„Schwesternassoziation“} \\ v &= \{(x, y), (x, z_1), (y, z_1)\} \\ r_1 &= \{(x, y)\} \\ r_2 &= v - r_1 \\ s &= r_1] \end{aligned}$$

6.3.7.3 Beibehaltung des Hierarchierelationstyps

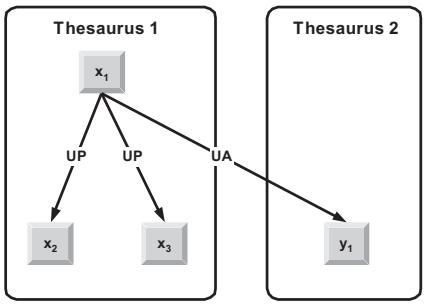

Abbildung 6.14: Unterschiedliche Typen von Hierarchierelationen bei Schwesterknoten.

Wenn innerhalb eines Komponetenthesaurus alle Unterbegriffe über denselben Hierarchierelationstypen (Abstraktionsrelation oder Bestandsrelation) verbunden sind, soll durch das Etablieren von Inter-Thesaurus-Beziehungen diese Einheitlichkeit nicht gestört werden. Ein Beispiel für einen Verstoß gegen diese Forderung zeigt Abbildung 6.14.

Wir notieren die Eigenschaft der *Markierung bei verschiedenen Hierarchierelationstypen bei Schwestern* wie folgt:

$\forall x_1 \in \mathbf{D}$:

$$\begin{aligned} &\exists x_2 \in \gamma(x_1).D : (x_2, x_1) \in \gamma(x_1).\mathrm{A} \\ \wedge\ &\nexists x_3 \in \gamma(x_1).D : (x_3, x_1) \in \gamma(x_1).\mathrm{P} \wedge \exists y \in \mathbf{D} : (y, x_1) \in \mathcal{P} \\ &\Rightarrow \exists k = (t, v, r_1, r_2, s) \in K \text{ mit} \end{aligned} \tag{6.26}$$

$$\begin{aligned} t &= \text{„AP-Schwestern“} \\ v &= \{(z, x_1) : (z, x_1) \in \gamma(x_1).\mathrm{A} \vee (z, x_1) \in \mathcal{P}\} \\ r_1 &= \{(z, x_1) : (z, x_1) \in \gamma(x_1).\mathrm{A}\} \\ r_2 &= v - r_1 \\ s &= r_1 \vee r_2 \end{aligned}$$

Und:

$\forall x_1 \in \mathbf{D}$:

$$\begin{aligned} &\exists x_2 \in \gamma(x_1).D : (x_2, x_1) \in \gamma(x_1).\mathrm{P} \\ \wedge\ &\nexists x_3 \in \gamma(x_1).D : (x_3, x_1) \in \gamma(x_1).\mathrm{A} \wedge \exists y \in \mathbf{D} : (y, x_1) \in \mathcal{A} \\ &\Rightarrow \exists k = (t, v, r_1, r_2, s) \in K \text{ mit} \end{aligned} \tag{6.27}$$

$$
\begin{aligned}
t &= \text{„PA-Schwestern“} \\
v &= \{(z, x_1) : (z, x_1) \in \gamma(x_1).\mathrm{P} \vee (z, x_1) \in \mathcal{A}\} \\
r_1 &= \{(z, x_1) : (z, x_1) \in \gamma(x_1).\mathrm{P}\} \\
r_2 &= v - r_1 \\
s &= r_1 \vee r_2
\end{aligned}
$$

Konfliktverursacher als auch die in Konflikt stehenden Beziehungen sind die Intra-Thesaurus-Hierarchiebeziehungen des einen Typs (Abstraktion oder Bestand) und die Inter-Thesaurus-Hierarchiebeziehungen des anderen Typs (entsprechend Bestand oder Abstraktion).

6.3.8 Konsistenz der Konfliktmarkierungen

Um eine einheitliche Behandlung gleichartiger Konflikte zu gewährleisten, fordern wir zusätzlich die Konsistenz gleichartiger Konfliktmarkierungen. Zwei Konfliktmarkierungen sind in diesem Sinne gleichartig, wenn

- es sich um den gleichen Konflikttypen handelt und
- die Deskriptoren, die Konfliktverursacher und in der Hierarchie am höchsten sind, identisch oder durch eine Inter-Thesaurus-Äquivalenzbeziehung verbunden sind und
- die Deskriptoren, die Konfliktverursacher und in der Hierarchie am niedrigsten sind, identisch oder durch eine Inter-Thesaurus-Äquivalenzbeziehung verbunden sind.

Konsistent sind zwei gleichartige Konfliktmarkierungen k_1 und k_2 genau dann, wenn es in $k_1.s$ kein Tupel mit einem Deskriptor gibt, der identisch oder äquivalent zu einem Deskriptor aus $k_2.v$ ist und nicht in mindestens einem Tupel in $k_2.s$ vorkommt.

Beispiel 6.1 *Abbildung 6.10 links zeigt ein Beispiel für eine Abstraktionsniveaudifferenz, die vier gleichartige Konfliktmarkierungen erfordert. Durch unterschiedliche Längen stehen die Pfade*

- $x_1 \overset{A^\star_{x_2}}{\hookrightarrow} x_3$ *und* $x_1 \overset{A^\star_{y_3}}{\hookrightarrow} x_3$
- $x_1 \overset{A^\star_{x_2}}{\hookrightarrow} y_4$ *und* $x_1 \overset{A^\star_{y_3}}{\hookrightarrow} y_4$
- $y_1 \overset{A^\star_{x_2}}{\hookrightarrow} x_3$ *und* $y_1 \overset{A^\star_{y_3}}{\hookrightarrow} x_3$
- $y_1 \overset{A^\star_{x_2}}{\hookrightarrow} y_4$ *und* $y_1 \overset{A^\star_{y_3}}{\hookrightarrow} y_4$

in Konflikt. Wird in einer Markierung des ersten Konfliktes y_2 als zu entfernender Deskriptor markiert, indem alle Tupel der Pfade, in denen y_2 vorkommt, in die Menge s aufgenommen werden ($s = \{(y_2, x_1), (y_3, y_2)\}$), muss dies aufgrund der Forderung der Konsistenz der Konfliktmarkierungen ebenfalls in allen gleichartigen Konfliktmarkierungen geschehen. Das impliziert z.B. für den vierten Konflikt bereits eindeutig folgende Markierung: $s = \{(y_2, y_1), (y_3, y_2)\}$.

6.4 Beschreibung von Thesaurusföderationen als Graphen

Wie bereits das Thesaurusmodell in Abschnitt 5.2 soll nun das Thesaurusföderationsmodell als gerichteter Graph mit beschrifteten Kanten veranschaulicht werden.

6.4.1 Knoten und Kanten

Definition 6.19 *Eine Thesaurusföderation $\Im$, die dem in Abschnitt 6.3 dargestellten Modell entspricht, kann als gerichteter Graph $F = (N, E)$ dargestellt werden. Seien $\theta_1, \theta_2, \cdots, \theta_k$ die Komponententhesauri der Thesaurusföderation. Dann ist $N = \{\#\Im\} \cup G_\Im \cup \#\omega \cup \omega.B \cup (\cup_{1\ldots m} \#\theta_i \cup \theta_i.G \cup \theta_i.B)$ eine endliche Menge von Knoten des Graphen, wobei $\#\Im \in \mathbb{N}$ der eindeutige Identifikator der Thesaurusföderation ist, $G_\Im$ die Menge der Föderierten Gruppen, $\#\omega \in \mathbb{N}$ der eindeutige Identifikator des Thesaurus der Ergänzenden Begriffe, $\omega.B$ die Menge der Benennungen des Thesaurus der Ergänzenden Begriffe und $\#\theta_i \in \mathbb{N}$ die eindeutigen Identifikatoren der Komponententhesauri sind, $\theta_i.G$ die Mengen der Gruppen der Komponententhesauri sowie $\theta_i.B$ die Mengen der Benennungen der Komponententhesauri. E ist eine endliche Menge beschrifteter, gerichteter Kanten.*

Jeder Knoten $n \in N$ ist mit seinem eindeutigen Identifikator $\star n$ durch die Bijektion $I_F : N \to \{1, ..., |N|\} \subseteq \mathbb{N}$ assoziiert. Das bedeutet $\forall n \in N : I_F(n) = \star n$.

Für jedes $n \in N$ liefert eine Funktion $\lambda(n)$ ein Tupel $(typ,\ \star n,\ s)$ wobei

$$typ = \begin{cases} \textit{Thesaurusföderation}, & \textit{falls } n = \#\Im \\ \textit{Föderierte Gruppe}, & \textit{falls } n \in G_\Im \\ \textit{Thesaurus der Erg. Begriffe}, & \textit{falls } n = \#\omega \\ \textit{Thesaurus}, & \textit{falls } n \in \{\#\theta_1, \ldots, \#\theta_k\} \\ \textit{Gruppe}, & \textit{falls } n \in \#\theta_1.G \cup \ldots \cup \#\theta_k.G \\ \textit{Deskriptor}, & \textit{falls } n \in \mathbf{D} \\ \textit{Äquivalenz-Nicht-Deskriptor}, & \textit{falls } n \in \\ & \{u \in \mathbf{B} - \mathbf{D} : \gamma^\star(\star u).U_u^E \neq \emptyset\} \\ \textit{Kombinations-Nicht-Deskriptor}, & \textit{falls } n \in \\ & \{u \in \mathbf{B} - \mathbf{D} : \gamma^\star(\star u).U_u^C \neq \emptyset\} \end{cases}$$

ist. Nur die ersten drei potenziellen Werte für typ sind neu, die anderen sind – eingeschränkt auf einen Thesaurus – bereits aus dem Thesaurusgraphen bekannt. s ist wie dort der Name des Knotens, also eine Zeichenkette.

Eine Kante $e \in E$ wird wiederum notiert als (n_1, α, n_2), wobei $n_1, n_2 \in N$ und $\alpha = (\alpha_1, \alpha_2)$ die Beschriftung der Kante ist. Die Funktion $\delta_1(e)$ liefert die erste Komponente α_1 der Beschriftung α einer Kante, wobei gilt $\alpha_1 \in \{\mathcal{HATTHESAURUS}, \mathcal{HATERGBEGRIFFE}, \mathcal{HATGRUPPE}, \mathcal{HATTOPTERM}, \mathcal{HATELEMENT}, \mathcal{BS}, \mathcal{BK}, \mathcal{UA}, \mathcal{UP}, \mathcal{VB}\} \cup \{hatGruppe, hatTopterm, hatBKTopterm, hatElement, BS, BK, UA, UP, VB\} \cup \{\mathsf{UA}, \mathsf{UP}, \mathsf{VB}\}$. Die erste Menge sind die neuen Kantenbeschriftungen der Thesaurusföderation, die zweite Menge beinhaltet bereits die vom Thesaurusgraphen her bekannten Kantentypen innerhalb der Komponententhesauri. Die dritte Menge schließlich beinhaltet die implizierten Intra-Thesaurus-Beziehungen. α_1 wird im Folgenden auch als Kantentyp bezeichnet. Wenn die zweite Komponente der Beschriftung keine Rolle spielt, notieren wir auch vereinfachend (n_1, α_1, n_2).

Die Funktion $\delta_2(e)$ liefert als zweite Komponente α_2 der Kantenbeschriftung α die Menge der Konfliktmarkierungen dieser Kante. Es gilt $\alpha_2 = \emptyset \vee$

$\{(\star k_1, ktyp_1, kr_1, ks_1), \ldots, (\star k_m, ktyp_m, kr_m, ks_m)\}$, *wobei* $m \in \mathbb{N}$ *die Anzahl der Konflikte, an der die Kante beteiligt ist,* $\star k_i \in \mathbb{N}$ *den eindeutigen Identifikator eines Konfliktes,* $ktyp_i \in \{$*Abstraktionszyklus, Abstraktionsredundanz, Bestandszyklus, Hierarchiezyklus, Beziehungstypdifferenz, Schwesternassoziation, Abstraktionsniveaudifferenz, AP-Schwestern, PA-Schwestern*$\}$ *den Konfliktttyp und* $kr_i \in \{1, 2\}$ *die Konfliktmenge bezeichnet (vgl. Abschnitt 6.3.4) sowie* $ks_i \in \{0, 1\}$ *angibt, ob diese Beziehung zu entfernen ist, um den Konflikt zu beseitigen (bei* $ks_i = 1$ *ist die Beziehung zu entfernen, sonst nicht).*

VB, VB und BS sind ungerichtete Kanten, alle anderen Kanten sind gerichtet. Alle gerichteten Kanten können entsprechend den Kanten der Thesaurusgraphen auch gegen die Richtung gelesen werden, die erste Komponente der Kantenbeschriftungen entspricht dann ISTTHESAURUSIN, ISTERGBEGRIFFIN, HATELEMENT, ISTTOPTERM, ISTELEMENTVON, KB, OA, OP, *istGruppeVon, istToptermIn, istBKToptermIn, istElementVon, BF, KB, OA, OP,* OA *oder* OP.

Für den Typ, den Identifikator und den Namen eines Knotens notieren wir kurz $\lambda_n^{typ}, \lambda_n^{\star}, \lambda_n^{s}$.

Die Kanten innerhalb eines Thesaurus wurden bereits in Abschnitt 5.2.1 definiert. An dieser Stelle definieren wir ausschließlich die durch die Thesaurusföderation entstehenden zusätzlichen Kanten (vgl. Abbildung 6.15, in der aus Gründen der besseren Lesbarkeit auf Nicht-Deskriptoren, durch die dargestellten Kanten implizierte Kanten und Konfliktmarkierungen verzichtet wurde).

Definition 6.20 *Wurzelknoten des Thesaurusföderationsgraphen F ist der Knoten* $\#\Im$ *mit dem Thesaurusföderationsidentifikator. Vom Wurzelknoten gibt es zu jedem Komponententhesaurus* $\theta \in \Theta$ *genau eine Kante des Typs* HATTHESAURUS, *zu dem Knoten des Thesaurus der Ergänzenden Begriffe eine Kante des Typs* HATERGBEGRIFFE, *zu jedem Knoten des Typs Föderierte Gruppe* $g \in G_{\Im}$ *genau eine Kante des Typs* HATGRUPPE *und zu jedem Topterm der Föderation* $d \in \mathbf{D}$ *genau eine Kante des Typs* HATTOPTERM. *Weitere Kanten gehen vom Wurzelknoten nicht aus. Das bedeutet*

$$(\#\Im, \mathcal{HATTHESAURUS}, \#\theta) \in E \quad \Leftrightarrow \quad \theta \in \Theta \tag{6.28}$$

$$(\#\Im, \mathcal{HATERGBEGRIFFE}, \#\omega) \in E \quad \Leftrightarrow \quad \omega = \Im.\omega \tag{6.29}$$

$$(\#\Im, \mathcal{HATGRUPPE}, g) \in E \quad \Leftrightarrow \quad g \in G_{\Im} \tag{6.30}$$

$$(\#\Im, \mathcal{HATTOPTERM}, d) \in E \quad \Leftrightarrow \tag{6.31}$$

$$d \in \mathbf{D}, \gamma^{\star}(\star d).O_d^{\mathrm{A}} = \gamma^{\star}(\star d).O_d^{\mathrm{P}} = O_d^{\mathcal{A}} = O_d^{\mathcal{P}} = \emptyset$$

Definition 6.21 *Von allen Gruppenknoten* $g \in G_{\Im}$ *gibt es eine Kante des Typs* HATELEMENT *zu jedem Deskriptorknoten* $d \in \mathbf{D}$, *für den* $g \in Q_{\Im_d}$ *gilt. Das bedeutet*

$$(g, \mathcal{HATELEMENT}, d) \in E \quad \Leftrightarrow \quad g \in Q_{\Im_d} \tag{6.32}$$

Jeder Deskriptorknoten, auf den von einer Föderierten Gruppe verwiesen wird, ist Topterm der Thesaurusföderation. Das bedeutet, dass von den Föderierten Gruppen nicht auf alle Topterme der Komponententhesauri verwiesen wird (vgl. Abbildung 6.15).

Die folgende Definition definiert die Kanten, die von Deskriptoren ausgehen. Aufgrund der Definitionen der zugrunde liegenden Inter-Thesaurus-Relationen stammen die jeweils verbundenen Knoten aus verschiedenen Thesauri. Es gilt zu beachten, dass im Gegensatz zum Thesaurusgraphen auch die BS- und BK-Kanten zwischen Deskriptoren definiert sind.

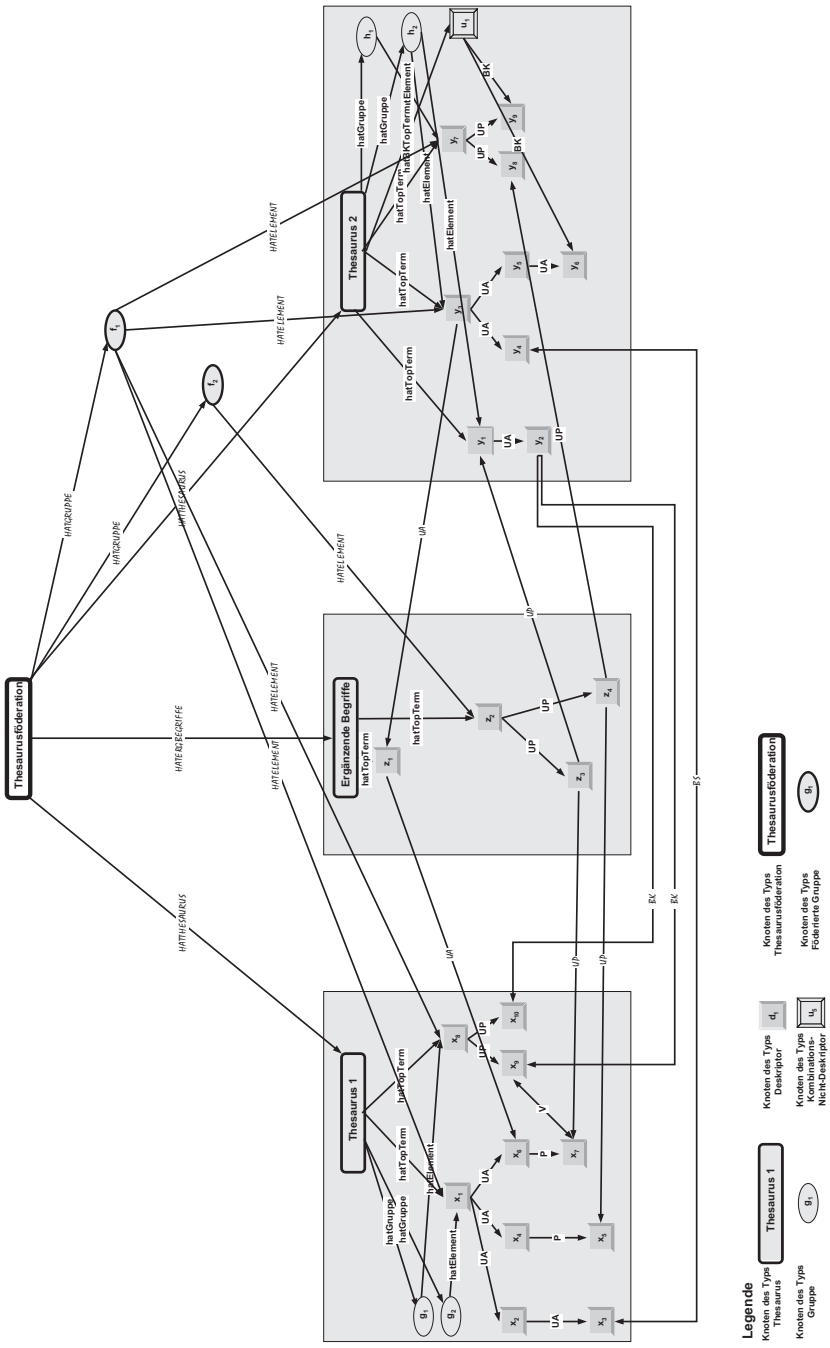

Abbildung 6.15: Vereinfachte Darstellung eines Thesaurusföderationsgraphen

Definition 6.22 *Von einem Deskriptor $x \in \mathbf{D}$ geht eine Kante des Typs $\mathcal{BS}$, $\mathcal{BK}$, $\mathcal{UA}$, $\mathcal{UP}$ bzw. $\mathcal{VB}$ zu einem Deskriptor $y \in \mathbf{D}$ genau dann, wenn es eine Inter-Thesaurus-Äquivalenz-, -Benutze-Kombinations-, -Abstraktions-, -Bestands- bzw. -Assoziationsbeziehung zwischen x und y gibt sowie Selbstbezug und gleichartige Mehrfachbeziehungen zwischen gleichen Knotenpaaren ausgeschlossen sind:*

$$(x, \mathcal{BS}, y) \in E \quad \Leftrightarrow \quad (x, y) \in \mathcal{E} \wedge \star x < \star y \tag{6.33}$$
$$(x, \mathcal{BK}, y) \in E \quad \Leftrightarrow \quad (x, y) \in \mathcal{C} \tag{6.34}$$
$$(x, \mathcal{UA}, y) \in E \quad \Leftrightarrow \quad (y, x) \in \mathcal{A} \tag{6.35}$$
$$(x, \mathcal{UP}, y) \in E \quad \Leftrightarrow \quad (y, x) \in \mathcal{P} \tag{6.36}$$
$$(x, \mathcal{VB}, y) \in E \quad \Leftrightarrow \quad (x, y) \in \mathcal{V} \wedge \star x < \star y \tag{6.37}$$

Der Selbstbezug der Inter-Thesaurus-Äquivalenzbeziehung wird also im Thesaurusföderationsgraphen aufgrund der Zusatzbedingung $\star x < \star y$ nicht explizit ausgedrückt. Durch diese Zusatzbedingung werden ebenfalls die ungerichteten $\mathcal{VB}$-und $\mathcal{BS}$-Kanten – wie auch die ungerichteten VB-Kanten – nur einfach zwischen einem Deskriptorpaar etabliert.

Definition 6.23 *Von einem Deskriptor $x \in \gamma^\star(\star x).D$ geht eine Kante des Typs* UA, UP *bzw.* VB *zu einem Deskriptor $y \in \gamma^\star(\star x).D$ genau dann, wenn es eine implizierte Intra-Thesaurus-Abstraktions-, -Bestands- bzw. -Assoziationsbeziehung zwischen x und y gibt:*

$$(x, \mathsf{UA}, y) \in E \quad \Leftrightarrow \quad (y, x) \in \mathsf{A} \tag{6.38}$$
$$(x, \mathsf{UP}, y) \in E \quad \Leftrightarrow \quad (y, x) \in \mathsf{P} \tag{6.39}$$
$$(x, \mathsf{VB}, y) \in E \quad \Leftrightarrow \quad (x, y) \in \mathsf{V} \wedge \star x < \star y \tag{6.40}$$

Die ungerichtete VB-Kante wird wiederum nur einfach zwischen einem Deskriptorpaar etabliert.

Definition 6.24 *Eine Kante $(n_1, (\alpha_1, \alpha_2), n_2)$ besitzt eine* Konfliktmarkierung *genau dann, wenn diese Kante an einem Konflikt beteiligt ist:*

$$(\star k, t, kr, ks) \in \alpha_2 \Leftrightarrow$$
$$\exists k = (t, v, r_1, r_2, s) \in K : (n_1, n_2) \in v \tag{6.41}$$

$$\textit{wobei } kr = \begin{cases} 1 & \textit{falls } (n_1, n_2) \in r_1 \\ 2 & \textit{falls } (n_1, n_2) \in r_2 \\ 0 & \textit{sonst} \end{cases} \quad \textit{und } ks = \begin{cases} 0 & \textit{falls } (n_1, n_2) \notin s \\ 1 & \textit{falls } (n_1, n_2) \in s \end{cases}$$

Die Menge der Konfliktmarkierungen kann aufgrund der Definition von K nur bei Kanten, die Deskriptorknoten verbinden, nicht leer sein.

Bemerkung 6.4 *Kanten zwischen Deskriptoren werden in der weiteren Arbeit auch durch deren Bezeichner und den Kantentyp notiert. Soll der Herkunftsthesaurus des Deskriptors zusätzlich erkenntlich gemacht werden, notieren wir den ersten Buchstaben des Thesaurusnamens durch einen Punkt getrennt vor dem Bezeichner, z.B.* (G.land conservation, $\mathcal{BS}$, A.land management).

6.4.2 Pfade

Definition 6.25 *Ein* Pfad *p in einem Thesaurusföderationsgraphen $F = (N, E)$ von einem Knoten n_1 zu einem Knoten n_l ($l \geq 2$) ist wie ein Pfad in einem Thesaurusgraphen definiert als $p = n_1.e_1.n_2.e_2...e_{l-1}.n_l$, wobei*

$$\forall i \in [1, l-1] : \exists (n_i, e_i, n_{i+1}) \in E \tag{6.42}$$

Ein solcher Pfad wird auch notiert als $n_1 \overset{\star}{\hookrightarrow} n_l$.

Abstraktions-, Bestands- und Hierarchiepfade in der Thesaurusföderation sind definiert als Pfade über Kanten innerhalb eines Thesaurus und zwischen den Thesauri.

Definition 6.26 *Ein wie in Definition 6.25 definierter Pfad heißt* Abstraktionspfad der Thesaurusföderation*, wenn zusätzlich gilt*

$$\forall i \in [1, l-1] : \delta_1(e_i) \in \{UA, \mathcal{UA}, \mathsf{UA}\} \tag{6.43}$$

Ein Abstraktionspfad wird auch notiert als $n_1 \overset{\mathrm{UA}^\star}{\hookrightarrow} n_l$.

Definition 6.27 *Ein wie in Definition 6.25 definierter Pfad heißt* Bestandspfad der Thesaurusföderation*, wenn zusätzlich gilt*

$$\forall i \in [1, l-1] : \delta_1(e_i) \in \{UP, \mathcal{UP}, \mathsf{UP}\} \tag{6.44}$$

Ein Bestandspfad wird auch notiert als $n_1 \overset{\mathrm{UP}^\star}{\hookrightarrow} n_l$.

Definition 6.28 *Ein wie in Definition 6.25 definierter Pfad heißt* Hierarchiepfad der Thesaurusföderation*, wenn zusätzlich gilt*

$$\forall i \in [1, l-1] : \delta_1(e_i) \in \{UA, UP, \mathcal{UA}, \mathcal{UP}, \mathsf{UA}, \mathsf{UP}\} \tag{6.45}$$

Ein Hierarchiepfad wird auch notiert als $n_1 \overset{\mathrm{U}^\star}{\hookrightarrow} n_l$.

6.4.3 Invarianten

Anhand des formalen Modells für Thesaurusföderationen (vgl. Abschnitt 6.3) können die Invarianten für einen Thesaurusföderationsgraphen hergeleitet werden. Wir führen sie explizit auf, um eine einfache Überprüfung der Erfüllung zu ermöglichen.

6.4.3.1 Identität der Komponententhesauri

Die Komponententhesauri selber werden nicht verändert (in Definition 6.1, S. 81, gehen die Komponententhesauri unverändert in die Föderation ein). Das bedeutet u.a., dass keinerlei Begriffe hinzugefügt oder entfernt werden und dass die Intra-Thesaurus-Relationen innerhalb eines Komponententhesaurus nicht verändert werden.

6.4.3.2 Konsistenz der Komponententhesauri und des Thesaurus der Ergänzenden Begriffe

Es wird vorausgesetzt, dass die Komponententhesauri sowie der Thesaurus der Ergänzenden Begriffe sämtliche in Abschnitt 5.2.3 aufgeführten Invarianten erfüllen.

6.4.3.3 Richtiger Einsatz der Thesaurus-verbindenden Kanten

Alle Kanten des Tys $\mathcal{BS}$, $\mathcal{BK}$, $\mathcal{UA}$, $\mathcal{UP}$ bzw. $\mathcal{VB}$ dürfen nur eine Beziehung zwischen Deskriptoren aus *verschiedenen* Komponententhesauri oder zwischen Ergänzenden Begriffen und Deskriptoren aus Komponententhesauri darstellen. Dies gilt aufgrund der Definition dieser Kanten.

Selbstverweise werden durch diese Invariante ebenfalls untersagt.

Aufgrund Bedingung 6.1 der Definition 6.11, die die paarweise Disjunktheit der Thesauri äquivalenter Begriffe fordert, darf von einem Knoten in einem Thesaurus maximal eine $\mathcal{BS}$-Kante zu jedem anderen Thesaurus führen.

Bei $\mathcal{BK}$-Kanten müssen mindestens jeweils zwei bei der Kombination benutzten Knoten aus *einem* Komponententhesaurus stammen (vgl. Definition 6.11, Bedingung 6.3, die der Inter-Thesaurus-Benutze-Kombination-Relation zugrunde liegt). Sie dürfen aber nicht durch einen Abstraktionspfad verbunden sein (vgl. Bedingung 6.4). Aufgrund von Bedingung 6.5 gilt, dass von einem Deskriptor nicht gleichzeitig $\mathcal{BS}$- und $\mathcal{BK}$-Kanten in *einen* Thesaurus verweisen dürfen.

6.4.3.4 Verbundenheit der Ergänzenden Begriffe

Aufgrund Definition 6.5 muss jeder Ergänzende Begriff entweder

- durch mindestens zwei Kanten mit Begriffen in einem Komponententhesaurus verbunden sein oder
- durch mindestens zwei $\mathcal{UA}$- oder $\mathcal{UP}$-Kanten mit weiteren Ergänzenden Begriffen verbunden sein.

Diese Invariante sichert die Minimalität der Begriffe des Thesaurus der Ergänzenden Begriffe.

6.4.3.5 Markierung bei Verstoß gegen Einzigartigkeit einer Kante

Zwei Knoten dürfen in eine Richtung ausschließlich durch eine einzige Kante verbunden sein. Dies gilt aufgrund der gleichlautenden Invariante für Thesaurusgraphen, der paarweisen Disjunktheit der Relationen (vgl. Abschnitt 6.3.7.1) und den Definitionen der Kanten, die von dem Wurzelknoten der Föderation und von Föderierten Gruppen ausgehen (Definitionen 6.20 und 6.21).

Bei Deskriptorknoten können aufgrund von implizierten Kanten Verstöße gegen diese Bedingung auftreten, die sowohl als Verstoß gegen die Zyklenfreiheit als auch als Beziehungstypdifferenz markiert werden (vgl. Abbildung 6.12). Wir formulieren die Eigenschaft der Markierung bei Verstoß gegen die Einzigartigkeit einer Kante für den zweiten Fall – der erste Fall erfordert nur

wenige Modifikationen dieser Notation – wie folgt:

$$
\begin{aligned}
&\forall x_1, x_2 \in N, x_1 \neq x_2 : \\
&\quad (x_1, (\alpha_{1_1}, \alpha_{1_2}), x_2) \in E \ \wedge \ (x_1, (\alpha_{2_1}, \alpha_{2_2}), x_2) \in E \wedge \\
&\quad \alpha_{1_1} \neq \alpha_{2_1} \wedge \alpha_{1_1}, \alpha_{2_1} \in \{UA, UP, V, \mathsf{UA}, \mathsf{UP}, \mathsf{V}\} \wedge \\
&\quad [\nexists y_1, y_2 \in N : (y_1, (\alpha_{3_1}, \alpha_{3_2}), y_2) \ \wedge \ (y_1, (\alpha_{4_1}, \alpha_{4_2}), y_2) \wedge \\
&\qquad\qquad \alpha_{3_1} \neq \alpha_{4_1} \wedge \alpha_{3_1}, \alpha_{4_1} \in \{UA, UP, V, \mathsf{UA}, \mathsf{UP}, \mathsf{V}\} \wedge \\
&\qquad\qquad (y_1, \mathcal{BS}, x_1) \ \wedge \ (y_2, \mathcal{BS}, x_2)] \\
&\Leftrightarrow \exists \star k \in \mathbb{N} : \\
&\quad (\forall (x_1, (\beta_{t_1}, \beta_{t_2}), x_2) \in E : \\
&\qquad (\star k, \text{„Beziehungstypdifferenz“}, kr_{t_1}, ks_{t_1}) \in \beta_{t_2}) \wedge \\
&\quad [\forall u_i, v_j \in N : ((x_1, (BS, \alpha_{5_{i_2}}), u_i) \in E \vee (u_i, (BS, \alpha_{5_{i_2}}), x_1) \in E)) \wedge \\
&\qquad\qquad ((x_2, (BS, \alpha_{6_{j_2}}), v_j) \in E \vee (v_j, (BS, \alpha_{6_{i_2}}), x_2) \in E)) \\
&\qquad \Rightarrow (\star k, \text{„Beziehungstypdifferenz“}, kr_{5_{i_2}}, ks_{5_{i_2}}) \in \alpha_{5_{i_2}} \wedge \\
&\qquad\quad (\star k, \text{„Beziehungstypdifferenz“}, kr_{6_{j_2}}, ks_{6_{j_2}}) \in \alpha_{6_{j_2}} \wedge \\
&\qquad\quad (\forall (u_i, (\alpha_{u_{i_{m_1}}}, \alpha_{u_{i_{m_2}}}), x_2) \in E : \\
&\qquad\qquad (\star k, \text{„Beziehungstypdifferenz“}, kr_{u_{i_{m_1}}}, ks_{u_{i_{m_1}}}) \in \alpha_{u_{i_{m_2}}}) \wedge \\
&\qquad\quad (\forall (x_1, (\alpha_{v_{i_{g_1}}}, \alpha_{v_{i_{g_2}}}), v_j) \in E : \\
&\qquad\qquad (\star k, \text{„Beziehungstypdifferenz“}, kr_{v_{i_{g_1}}}, ks_{v_{i_{g_1}}}) \in \alpha_{u_{i_{m_2}}}) \wedge \\
&\qquad\quad (\forall (u_i, (\alpha_{u_{i_{n_1}}}, \alpha_{u_{i_{n_2}}}), v_j) \in E : \\
&\qquad\qquad (\star k, \text{„Beziehungstypdifferenz“}, kr_{u_{i_{n_1}}}, ks_{u_{i_{n_1}}}) \in \alpha_{u_{i_{n_2}}})]
\end{aligned}
$$

Der erste Teil dieser Formel spezifiziert zwei Begriffe, zwischen denen verschiedene Typen von Intra- und implizierten Intra-Thesaurus-Beziehungen bestehen. Existieren solche Begriffspaare, ist eine entsprechende Markierung erforderlich. Um den Konflikt nur einfach zu markieren, wird von äquivalenten Begriffspaaren das mit dem niedrigsten Identifikator des Begriffes, von dem die Beziehungen ausgehen, ausgewählt. Markiert sind dann alle Beziehungen zwischen dem spezifizierten Begriffspaar, alle Beziehungen, die zwischen den jeweils äquivalenten Begriffen bestehen, sowie alle Äquivalenzbeziehungen zwischen den beteiligten Begriffen.

6.4.3.6 Markierung von Abstraktionszyklen

Existieren Zyklen innerhalb der Abstraktionspfade der Thesaurusföderation, so sind alle an einem Zyklus beteiligte Kanten entsprechend markiert. Das bedeutet:

$$
\begin{aligned}
\forall n \in N : n \overset{UA^\star}{\hookrightarrow} n \Rightarrow \exists \star k \in \mathbb{N} : \forall (n_i, n_j) \in n \overset{UA^\star}{\hookrightarrow} n, (n_i, \alpha, n_j) \in E : \\
(\star k, \text{„Abstraktionszyklus“}, kr_{i,j}, ks_{i,j}) \in \delta_2(n_i, \alpha, n_j)
\end{aligned}
$$

wobei $kr_{i,j}$ und $ks_{i,j}$ wie in Definition 6.24 beschrieben definiert sind.

Die Markierung von Abstraktionszyklen gilt aufgrund der in Bedingung 6.18 geforderten Markierung bei Verstoß gegen die Zyklenfreiheit der Inter- und Intra-Thesaurus-Abstraktionsrelationen.

6.4.3.7 Weitere Markierungsinvarianten

Entsprechend der Markierung von Abstraktionszyklen können alle in Abschnitten 6.3.6 und 6.3.7 aufgeführten Konfliktmarkierungen auf den Thesaurusföderationsgraphen übertragen werden. Wir verzichten auf die Notation dieser sehr ähnlichen Übertragungen und führen zusammenfassend die weiteren Markierungsinvarianten namentlich auf: Markierung von Abstraktionsredundanzen, Abstraktionsniveaudifferenzen, Bestandszyklen, Hierarchiezyklen, Beziehungstypdifferenzen, Schwesternassoziationen, AP-Schwestern und PA-Schwestern.

6.5 Resümee

Auf der Basis existierender Modelle für Multi-Thesaurus- und Multi-Ontologie-Systeme haben wir ein Informationsmodell für Thesaurusföderationen entwickelt, das die Grundlage für ein skalierbares und flexibles Multi-Thesaurus-System und damit für den weiteren Verlauf dieser Arbeit ist.

Die Entscheidung, Beziehungen zwischen Begriffen durch Inter-Thesaurus-Relationen auszudrücken, deren Semantik aus klassischen Thesauri übertragen wurde, erlaubt, alle notwendigen Beziehungen auszudrücken, und hat die weiteren Vorteile, dass bereits vorhandene Werkzeuge leicht angepasst werden können und Benutzer keine grundsätzlich neue Semantik der Relationen erlernen müssen.

Durch das explizite Aufführen von Invarianten wird eine Mindestqualität der Thesaurusföderation festgelegt, die für die Unterstützung des Information-Retrieval-Prozesses in föderierten Informationssystemen erforderlich ist. Da eine Reihe von Konflikten nicht oder nur mit sehr großem Aufwand a priori aufgelöst werden kann, wurden Invarianten mit Konfliktmarkierungen eingeführt, die erstmalig das situationsabhängige Auflösen der Konflikte im Moment der Benutzeranfrage erlauben. Berücksichtigt werden von uns zudem nicht nur die explizt ausgedrückten Beziehungen, sondern ebenfalls die durch explizite Beziehungen implizierten Beziehungen. Erst so können Verstöße gegen die Invarianten vollständig erkannt werden.

Um den Zugriff auf die verteilten Thesauri und damit die Datenhoheit der Thesaurusanbieter zu unterstützen, wurde auf eine Interlingua zum Herstellen von Beziehungen zwischen den Komponententhesauri verzichtet. Sattdessen wurde die Idee einer Extralingua entwickelt, die ausschließlich zusätzliches Informationswissen enthält, das in den Komponententhesauri nicht enthalten ist. Ein Nebeneffekt dieser Extralingua ist, dass das Wissen, das an zentraler Stelle vorgehalten wird, ein deutlich geringeres Volumen als bei Verwendung einer Interlingua hat.

Kapitel 7

Wissensakquisitionsarchitektur

Das bereits in Kapitel 4 eingeführte Vorgehensmodell bildet den organisatorischen Rahmen der Begriffsintegration. Als technischer Rahmen wird in diesem Kapitel eine flexible und skalierbare Architektur entwickelt, die die Akquise des Integrationswissens innerhalb der verschiedenen Phasen des Vorgehensmodells adäquat unterstützt. Die Einordnung dieser Wissensakquisitionsarchitektur in den Lösungsansatz wird in Abbildung 7.1 dargestellt.

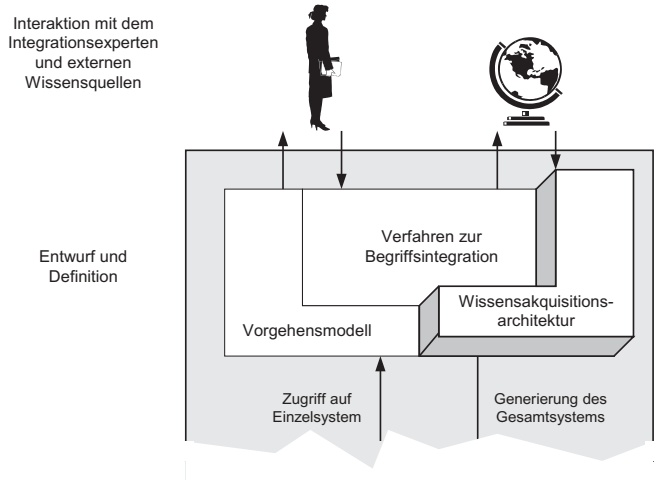

Abbildung 7.1: Einordnung der Wissensakquisitionsarchitektur in den Lösungsansatz

Die Anforderungen an eine solche Wissensakquisitionsarchitektur werden in Abschnitt 7.1 zusammengefasst. Anhand dieses Anforderungskataloges fällt die Entscheidung, die Wissensakquisitionsarchitektur als eine Blackboard-Architektur zu realisieren. Die Prinzipien von Blackboard-Architekturen sowie deren Anwendung in der Praxis werden in Abschnitt 7.2 erläutert. Basierend auf diesem Modell wird eine flexible Architektur zur Akquisition von Integrationswissen in Thesaurusföderation entwickelt (Abschnitt 7.3). Das in Abschnitt 7.4 entwickelte Bewertungsmodell ermöglicht das Einbringen unsicheren Wissens, eine Bewertung der Qualität von Wissensproduzenten sowie eine Aggregierung von Einzelbewertungen einer Aussage zu dessen Gesamtbewertung.

7.1 Anforderungen

Die Forderungen der Skalierbarkeit und Flexibilität bedeuten hinsichtlich einer Architektur zur Unterstützung der Akquise des Integrationswissens für Thesaurusföderationen folgende Anforderungen:

- Wie wir in den Kapiteln 2 und 3 gezeigt haben, existiert für das Problem der Thesaurusintegration kein geschlossener Lösungsansatz. Daher soll die Architektur die *flexible und iterative Lösungssuche* unterstützen.
- Aufgrund der Ungenauigkeit bei der Interpretation der Semantik der Begriffe und ihrer Beziehungen sowie der generellen Vagheit der Bedeutung natürlicher Sprache kann eine optimale Lösung für die Thesaurusintegration nicht angegeben werden. Die Architektur soll daher das Finden *„zufriedenstellender" Lösungen* und von *Lösungsalternativen* für Teilprobleme unterstützen.
- Der Lösungsraum kann in angemessener Zeit nicht vollständig durchsucht werden. Eine *Zerlegung des Problems* in Teilprobleme, wie dies bereits in Kapitel 4.2.2.1 für eine grobgranulare Betrachtungsebene erläutert wurde, soll deshalb von der Architektur ebenso unterstützt werden wie die *Einschränkung des Suchraumes* und die *strategisch geschickte Suche*. Als Basis hierzu sollen die Berücksichtigung des aktuellen Lösungsstandes und von Lösungsalternativen, ein opportunistisches Vorgehen sowie problemspezifische Strategien dienen können.
- Das *Experimentieren mit verschiedenen Akquisitionsalgorithmen* soll ermöglicht werden, um auf die Heterogenität der verschiedenen Thesauri angemessen reagieren zu können.
- Zur Lösungsfindung sollen *unterschiedliche Arten von Wissen und Wissensverarbeitungsverfahren* eingebracht werden können, etwa linguistisches Wissen, strukturbasiertes Wissen und das Wissen menschlicher Integrationsexperten.
- Es soll ermöglicht werden, *unsicheres Wissen* in die Lösungsfindung einzubringen. Damit sollen auch Aussagen getroffen werden können, die nicht vollkommen abgesichert sind, um der unscharfen Semantik der Begriffe und der Komplexität des Problems gerecht werden zu können. Dieses unsichere Wissen soll im weiteren Lösungsprozess angemessen berücksichtigt werden.
- Die Architektur soll einen *modularen Lösungsansatz* unterstützen. Das bedeutet zum einen, dass eine *Trennung der Problemlösungsstrategie von den Problemlösungsverfahren* ermöglicht werden soll. Zum anderen sollen die verschiedenen *Problemlösungsverfahren so gekapselt* werden können, dass die interne Repräsentation der zu verabeitenden Daten sowie die Implementierung der Verfahren verborgen wird und so die Erweiterbarkeit um weitere Lösungsverfahren und deren Austauschbarkeit gewährleistet ist.
- *Interaktion* soll in mehrfacher Hinsicht ermöglicht werden: Bei der inkrementellen Lösungsfindung sollen die Verfahren auf Ergebnissen anderer Verfahren aufbauen können. Schließlich soll der Mensch seine Expertise durch Vorschläge, Bewertungen und Entscheidungen einbringen können.

Ein erster Abgleich dieser Anforderungen mit den Eigenschaften von Blackboard-Architekturen (insbesondere die lose Kopplung unterschiedlicher Repräsentations- und Verarbeitungsmittel, die flexible und opportunistische Lösungssuche, die Trennung von Problemlösungsverfahren von

der Problemlösungsstrategie) legt es nahe, am Blackboard-Modell orientierte Architekturen detaillierter zu betrachten.

7.2 Blackboard-Architekturen

7.2.1 Einführung

7.2.1.1 Blackboard-Modell

Das *Blackboard-Modell* [New62, AN72, Sim77] entstammt dem Bereich der Künstlichen Intelligenz. Es versucht erstmals, den Prozess des verteilten Problemlösens durch geeignete Hilfsmittel zu unterstützen [BZW98]. Die folgende Metapher geht auf das Hersay-Projekt [EHRLR88, LE88] zurück und beschreibt alle Elemente der heutigen Interpretation des Blackboard-Modells (nach [Cra95]): Man stelle sich eine Gruppe von Experten vor, die für ein komplexes Problem mithilfe einer Wandtafel kooperierend (aber auch rivalisierend) eine akzeptable Lösung suchen. Sie können auf die Tafel schreiben und malen, Einträge wegwischen und ersetzen sowie Einträge mit Pfeilen verbinden und dadurch ihren Beitrag zur Lösungsfindung leisten. Dabei darf jeder Experte nur dann Einträge auf die Tafel schreiben oder sie verändern, wenn sie einen Bezug zu seiner Expertise haben und zur Lösungsfindung beitragen. Die Kommunikation der Experten geschieht ausschließlich über das Blackboard, ihnen ist nicht erlaubt, miteinander zu reden. Das Problem wird als gelöst betrachtet, wenn die Experten darin übereinstimmen, dass auf der Tafel eine adäquate Lösung dargestellt ist.

7.2.1.2 Komponenten eines Blackboard-Systems

Einem *Blackboard-System* liegt das Blackboard-Modell als Systemorganisations- und Strukturierungsprinzip zugrunde. Es wird auf Architekturebene durch eine *Blackboard-Architektur* beschrieben.

Die bedeutendsten Komponenten einer Blackboard-Architektur sind somit das *Blackboard*[1] als zentrales Kommunikationsmedium und globaler Informationsträger sowie *Agenten*, die als spezialisierte Experten jeweils Teilprobleme lösen. Zur Steuerung der Agenten wird in vielen Blackboard-Architekturen zusätzlich eine Steuerungskomponente verwendet.

Definition 7.1 *Das* Blackboard *ist eine globale, unter Umständen hierarchisch organisierte Datenbasis. Es speichert alle Daten, Zwischenergebnisse und Ergebnisse, die während des Problemlösungsprozesses entstehen, und stellt sie für die Weiterverarbeitung bereit. Dabei werden als Elemente abstrakte Datentypen auf das Blackboard geschrieben, die ggf. untereinander verknüpft werden. Diese Elemente werden als* Blackboardeinträge *bezeichnet, im Problemlösungskontext auch als* Hypothesen.

Definition 7.2 *Die* Agenten *(die Experten der Metapher) streben mit ihrer Expertise eine geeignete Lösung des Problems an. Das bedeutet, dass die Experten das Problemlösungswissen implementieren. Sie sind jeweils zuständig für Lösungsbeiträge zu einem bestimmten Typ von Teilproblemen. Ihren Beitrag leisten sie, indem sie entweder selbst den Problemlösungsprozess beobachten und sich melden, wenn sie relevante Situationen erkennen, oder aber nachdem sie von der Steuerungskomponente dazu aufgefordert wurden.*

[1] Der deutsche Begriff Wand- oder Schultafel wird in diesem Zusammenhang auch in der deutschsprachigen Fachliteratur nicht verwendet.

Definition 7.3 *Da in der Anwendung häufig nicht alle möglichen Agenten gleichzeitig ihre Beiträge liefern sollen, entscheidet eine* Steuerungskomponente*, welche der möglichen Agenten in welcher Reihenfolge zum Einsatz kommen. Sie bildet somit den Schiedsrichter in Konfliktsituationen. Dabei stützt sich die Steuerungskomponente auf eine Problemlösungsstrategie, die unabhängig vom Problemlösungswissen implementiert ist.*

7.2.1.3 Entwurfsfreiräume

Die Umsetzung des Blackboard-Modelles in Blackboard-Systeme erlaubt eine Reihe von Entwurfsfreiräumen. Tabelle 7.1 stellt diese Entwurfsfreiräume zusammenfassend dar (vgl. [Cor89, Cra95, Kir96, BZW98]).

	Blackboard	Agenten
Architektur	zentral vs. mehrere Blackboards für verschiedene Aufgaben	alle Agenten gleichberechtigt vs. zusätzliche Steuerungsagenten und/oder Management-Module als Strategie und Benachrichtigungsinstrumente
Organisation und Kommunikation	alle Informationen für alle Agenten zugänglich vs. Einteilung in verschiedene Bereiche	Kommunikation ausschließlich über Blackboard vs. direkte Kommunikation vs. Beauftragung weiterer Agenten
Ausführung		sequenziell vs. parallel (und verteilt)

Tabelle 7.1: Entwurfsfreiräume für Blackboard-Systeme

Beim Entwurf eines Blackboard-Systems als Wissensakquisitionssystem für Thesaurusföderationen gilt es, die angemessenen Alternativen auszuwählen.

7.2.1.4 Potenzielle Probleme

Potenzielle Probleme, die allgemein beim Umgang mit Blackboard-Architekturen beachtet werden müssen und bei der Begriffsintegration ebenso auftreten können, sind [Vie97, BZW98, Bor99]:

Blackboard als Flaschenhals: Die zentrale Rolle des Blackboards und die Notwendigkeit der globalen Verfügbarkeit können zu Engpässen und Verzögerungen bei Zugriffen führen. Dies gilt insbesondere, wenn es nur einen globalen Bereich auf dem Blackboard gibt.

Qualität der Blackbord-Einträge: Ursprünglich existiert keine Instanz, die die Qualität der Einträge überprüft. Neuere Ansätze führen daher eine zentrale Managementkomponente ein, den so genannten *Moderator*.

Netzbelastung: Sind die Agenten verteilt, kann es zu einer erhöhten Netzbelastung kommen, da alle Informationen auf das Blackboard zu schreiben sind.

7.2.2 Blackboard-Architekturen in der Praxis

Das Blackboard-Modell wurde inzwischen in einer Reihe von Systemen zur Lösung komplexer Probleme aufgegriffen. Wenn es auch noch nicht als ein Standard-Modell bekannt ist, wird eine

weitere Verbreitung für die Zukunft erwartet. Dies liegt daran, dass zum einen die Anforderungen an die „Lösungskompetenz“ von Software-Systemen immer größer wird und zum anderen das Zusammenspiel von im Netz verteilten Agenten zur Lösungsfindung weiter an Bedeutung gewinnen wird. Nicht zuletzt aus diesen Gründen findet sich das Blackboard-Modell mittlerweile sogar unter den *Entwurfsmustern* (vgl. etwa [BMR+96]).

Beispiele für den Einsatz von Blackboard-Systemen in jüngerer Zeit sind etwa wissensbasierte Assistenzsysteme in der Medizin [Kin96] oder „intelligente“ Tutor-Systeme [Bor99], also Lernumgebungen. Den mit diesen Systeme zu lösenden Problemen ist gemein, dass die Lösungsfindung – das bedeutet in diesen Fällen die intelligente Unterstützung des Menschen – ebenso wie die Integration von Thesauri ein wissensintensiver Prozess ist.

Das zur automatischen Erzeugung von Thesauri entwickelte Blackboard-System von Viegener [Vie97] ist das unserer Arbeit am nächsten stehende System. Es wird zur Kombination von Konstruktionsverfahren verwendet, um eine unabhängige Entwicklung der einzelnen Verfahren zu gewährleisten und dennoch die Grundlage für ein zielgerichtetes Vorgehen des Gesamtsystems zu sein. Da das Modell sich in dieser Arbeit mit ähnlicher Zielsetzung bewährt hat, sehen wir eine Weiterentwicklung der Architektur und eine Anpassung für die Thesaurusintegration als sehr erfolgsversprechend an. Defizite in der Architektur von Viegener sehen wir insbesondere in der Nicht-Berücksichtigung jeder Interaktionen mit menschlichen Experten. Die Problemlösungsstrategie ist ohne Berücksichtigung von Rahmenbedingungen und Lösungsfortschritt bzw. Schwierigkeiten bei der Lösungsfindung wenig flexibel. Des Weiteren fehlt die Möglichkeit, das im Laufe des Lösungsprozesses gewonnene unsichere Wissen im weiteren Prozess mit zu berücksichtigen. Ebenso findet keine klare Trennung einer Entscheidungsfindung sowie der Umsetzung dieser Entscheidung statt.

7.2.3 Anforderungsabgleich

Die Tendenz, aufgrund der positiven Erfahrungen von Viegener bei der Verwendung einer Blackboard-Architektur für die automatische Thesauruserstellung eine solche Architektur auch für die semi-automatische Integration von Thesauri zu verwenden, wird durch folgende Gegenüberstellung wissensbasierter Systeme verstärkt (nach [Sch93]):

	Semantisches Modell	Organisation der Komponenten	Kontrolle
Regelbasiertes System	Zielreduktion; Problemlösen als Beweisen	Zielhierarchie; Wenn-Dann-Regeln	Regelverkettung gemäß Inferenzmechanismus
Fallbasiertes System	Problemlösen durch Analogie; Lernen durch Zuwachs an Erfahrung	Gedächtnis mit früheren Fällen; Anpassungsheuristiken	Ähnlichkeitsmaß bestimmt Auswahl früherer Fälle
Objektorientiertes System	Gliederung nach Objekten, nicht Funktionen; Vererbung von Eigenschaften und Methoden; direkte Kommunikation	Klassenhierarchie	Kontrollübertragung durch Nachrichten

Fortsetzung auf der nächsten Seite ...

... Fortsetzung

	Semantisches Modell	Organisation der Komponenten	Kontrolle
Blackboard-System	opportunistische, inkrementelle Problemlösung; lose Kopplung unterschiedlicher Repräsentations- und Verarbeitungsmittel; indirekte Kommunikation	globaler Mehrebenenspeicher; modulare Spezialisten	Spezialisten als Agenten; explizite Kontrollstrategie; Agenda aktivierter Spezialisten

Tabelle 7.2: Architekturen für wissensbasierte Systeme

In dieser Gegenüberstellung sind *Agentensysteme* als wissensbasierte Systeme nicht enthalten. Das ist darauf zurückzuführen, dass die Begriffe *Agenten* und *Agentensysteme* weitläufig genutzt werden und eine einheitlich Sicht nicht existiert (vgl. z.B. [Min85, GK94, Coe94, HR95, FG96, WJ95]). Semantisches Modell, Organisation der Komponenten und Kontrolle können in unterschiedlichen Arten von Agentensystemen unterschiedlich sein. Da Agentensysteme im Bereich der wissensbasierten Systeme und insbesondere der verteilten künstlichen Intelligenz eine wichtige Rolle spielen, betrachten wir sie jedoch näher.

Zur Unterscheidung von nicht-agentischen Systemen werden eine Reihe von Eigenschaften für Agenten genannt [FG96, WJ95, Fri00]: Autonomie (Kontrollfähigkeit über eigenes Handeln im Rahmen des Handlungsspielraums), Intelligenz (im weitesten Sinne), Interaktivitat (Interaktion durch gleichberechtigte Kommunikation häufig auf Grundlage der Sprechakttheorie [Sea69] auf linguistischer Ebene), Reaktivität (unmittelbare Reaktion auf Ereignisse), Intentionalität und zielgerichtetes Verhalten (explizite Darstellung und Nutzung intentionaler Begriffe) und Rationalität (Nutzenoptimierung). Diese Eigenschaften werden jedoch sehr unterschiedlich interpretiert, daher wird häufig zwischen verschiedenen Agententypen mit spezifischen Schwerpunkten unterschieden (z.B. Kognitive Agenten, Kooperative Agenten, Konkurrierende Agenten, Reaktive Agenten).

Unter Berücksichtigung des üblichen Spielraumes bei der Definition von Agentensystemen sind Blackboard-Systeme durchaus ein Typ von Agentensystem – da mehrere Agenten beteiligt sind sogar eines Multiagentensystems. In verteilten Umgebungen werden aufgrund der indirekten Kommunikation über das zentrale Blackboard häufig andere Typen von Agentensystemen mit direkter sprechaktbasierter Kommunikation bevorzugt. Jedoch spielen im Rahmen unserer Arbeit die Verteiltheit bei der Problemlösung ebenso wie die sprechaktbasierte Kommunikation keine Rolle, daher halten wir die Blackboard-Architektur für die angemessenere Architektur. Zudem sei angemerkt, dass für den Fall einer angestrebten verteilten Problemlösung bereits in [CCWB94] eine Architektur auf Basis hierarchisch verteilter Blackboards entwickelt wurde.

Ein Vergleich der Anforderungen aus Abschnitt 7.1 mit den Eigenschaften von Architekturen für wissensbasierte Systeme bestätigt die Auswahl einer Blackboard-Architektur. Der direkter Abgleich der Anforderungen mit den Eigenschaften unter Berücksichtigung der Entwurfsfreiräume wird in Tabelle 7.3 dargestellt.

Anforderung	Erfüllung mit Blackboard-Architektur
flexible und iterative Lösungssuche	Aufgabe einer geeigneten Steuerungskomponente
Lösungsalternativen und Finden einer zufriedenstellenden Lösung	Aufgabe einer geeigneten Steuerungskomponente
Zerlegung des Problems	für explizte Zerlegung Erweiterung erforderlich, implizite Zerlegung durch für Teilprobleme verantwortliche Experten
Einschränkung des Suchraumes, strategisch geschickte Suche	Aufgabe einer geeigneten Steuerungskomponente
Experimentieren mit verschiedenen Akquisitionsalgorithmen	Unterstützung durch Blackboard als Zwischenergebnisspeicher
Einbringen unterschiedlicher Arten von Wissen und Wissensverarbeitungsverfahren	explizite Unterstützung durch Implementierung von Problemlösungsverfahren als Agenten
Einbringen von unsicherem Wissen	Erweiterungen erforderlich
Trennung von Problemlösungsstrategie und Problemlösungsverfahren	explizite Unterstützung durch Differenzierung zwischen Steuerungsagent und Problemlösungs-Agenten
Interaktionsunterstützung	Blackboard als Interaktionsinfrastruktur und Hypothesen als Interaktionsmittel

Tabelle 7.3: Abgleich der Anforderungen mit einer Blackboard-Architektur

Anhand der Ergebnisse dieses Anforderungsabgleiches schließen wir, dass eine Blackboard-Architektur sich als vielversprechende Grundlage der weitergehenden Arbeiten eignet. Im Folgenden stellen wir daher eine für die Unterstützung des gesamten Prozesses der Thesaurusintegration entwickelte blackboardbasierte Wissensakquisitionsarchitektur vor.

7.3 Blackboardbasierte Wissensakquisitionsarchitektur FA^{2}ITH

7.3.1 Überblick

Unsere Blackboard-Architektur, die wir FA^{2}ITH für *Flexible Architektur zur Akquisition von Integrationswissen in THesaurusföderationen* nennen, ist in Abbildung 7.2 im Überblick dargestellt.

Zentrale Komponente ist das Blackboard, das wir in verschiedene Bereiche unterteilen (vgl. Abschnitt 7.3.2). Über dieses Blackboard kommunizieren unterschiedliche Typen von Agenten:

Experten: Die Experten besitzen das eigentliche Problemlösungswissen und wenden dieses in Verfahren an, um über die Lösung von Teilproblemen zur Lösung des Gesamtproblems beizutragen.

Benutzeragent: Das Integrationswissen kann aufgrund des erforderlichen grundsätzlichen Verständnisses der Semantik nur semi-automatisch erworben werden. Die Mitarbeit eines menschlichen Experten wird von dem Benutzeragenten – im Sinne eines ausgezeichneten Experten – unterstützt.

Planungsagent: Der Planungsagent unterstützt den menschlichen Experten in der Detaillierung der Lösungsstrategie. Dazu identifiziert er Abhangigkeiten und die Erfüllung notwendiger Voraussetzungen.

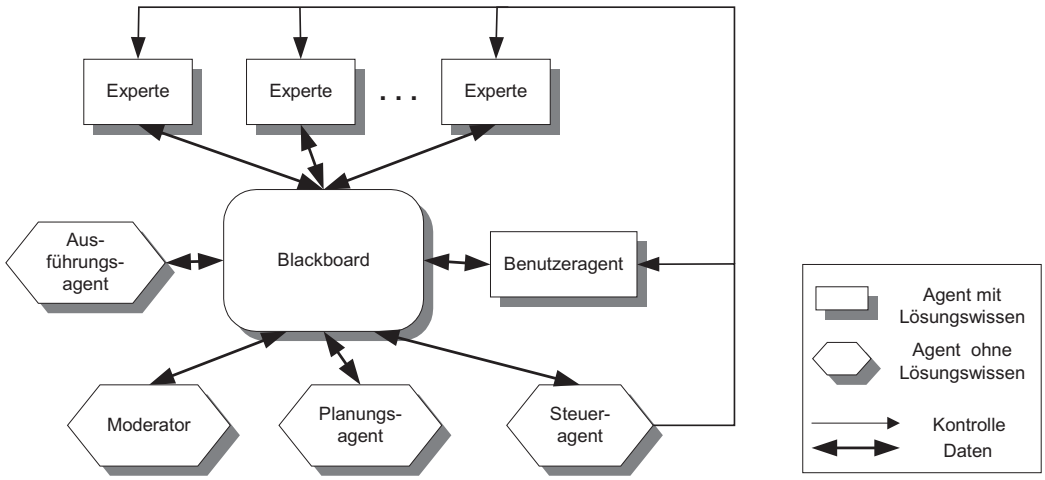

Abbildung 7.2: FA^{2}ITH im Überblick

Steueragent: Der Steueragent kontrolliert die Abarbeitung der Lösungsstrategie, indem er zur Bearbeitung der Teilprobleme opportunistisch und optimal im Sinne einer zügigen Problembearbeitung Experten auswählt.

Moderator: Der Moderator überprüft die Qualität der Einträge des Blackboards und entscheidet anhand eines Bewertungsmodelles, ob Hypothesen akzeptiert oder abgelehnt werden.

Ausführungsagent: Befinden sich auf dem Blackboard Lösungen für Teilprobleme, so ist es die Aufgabe des Ausführungsagenten, diese in die Thesaurusföderation zu übertragen.

Im Gegensatz zu klassischen Blackboard-Architekturen unterteilen wir die Aufgaben der Steuerungskomponente somit in eine Planungsaufgabe und eine Durchführungsaufgabe. Auf diese Weise wird eine explizite Zerlegung des Gesamtproblems in Teilprobleme unterstützt, auf der wiederum eine flexible Lösungssuche aufsetzen kann. Statt für die Teilprobleme und deren Abhängigkeiten sowie für den Steuerungsagenten Beschreibungen, Konzepte und Umsetzungen neu zu entwickeln, setzen wir an dieser Stelle auf Techniken, Methoden und Werkzeuge aus dem Bereich der Workflow-Management-Systeme. Deren Aufgabe ist es, die Definition von Prozessen zu unterstützen und die Abarbeitung von Prozess-Instanzen zu steuern. Somit decken sie zentrale Aufgaben des Planungs- und Steuerungsagenten bereits ab.

Eine prozessorientierten Darstellung des Zusammenspiels der Agenten zeigt Abbildung 7.3. Bei einer solchen Betrachtung ist die Lösungsstrategie ein Prozess, der im Wesentlichen vom Benutzeragenten sowie mit Unterstützung des Planungsagenten definiert und modifiziert wird. Der Steueragent interpretiert diese Prozessdefinition und steuert die Abarbeitung eines dermaßen definierten Prozesses. Für solche Prozessdefinitionen existieren in Workflow-Management-Systemen Standards, die wir aufgreifen. Die Aufgabe des Steueragenten wird in einem Workflow-Management-System auf Basis ebensolcher Prozessdefinitionen von der Workflow-Maschine wahrgenommen [Hol95]. Entsprechend können wir den Steueragenten auf einer solchen Workflow-Maschine aufsetzen.

Die einzelnen Komponenten von FA^{2}ITH werden im Folgenden näher erläutert.

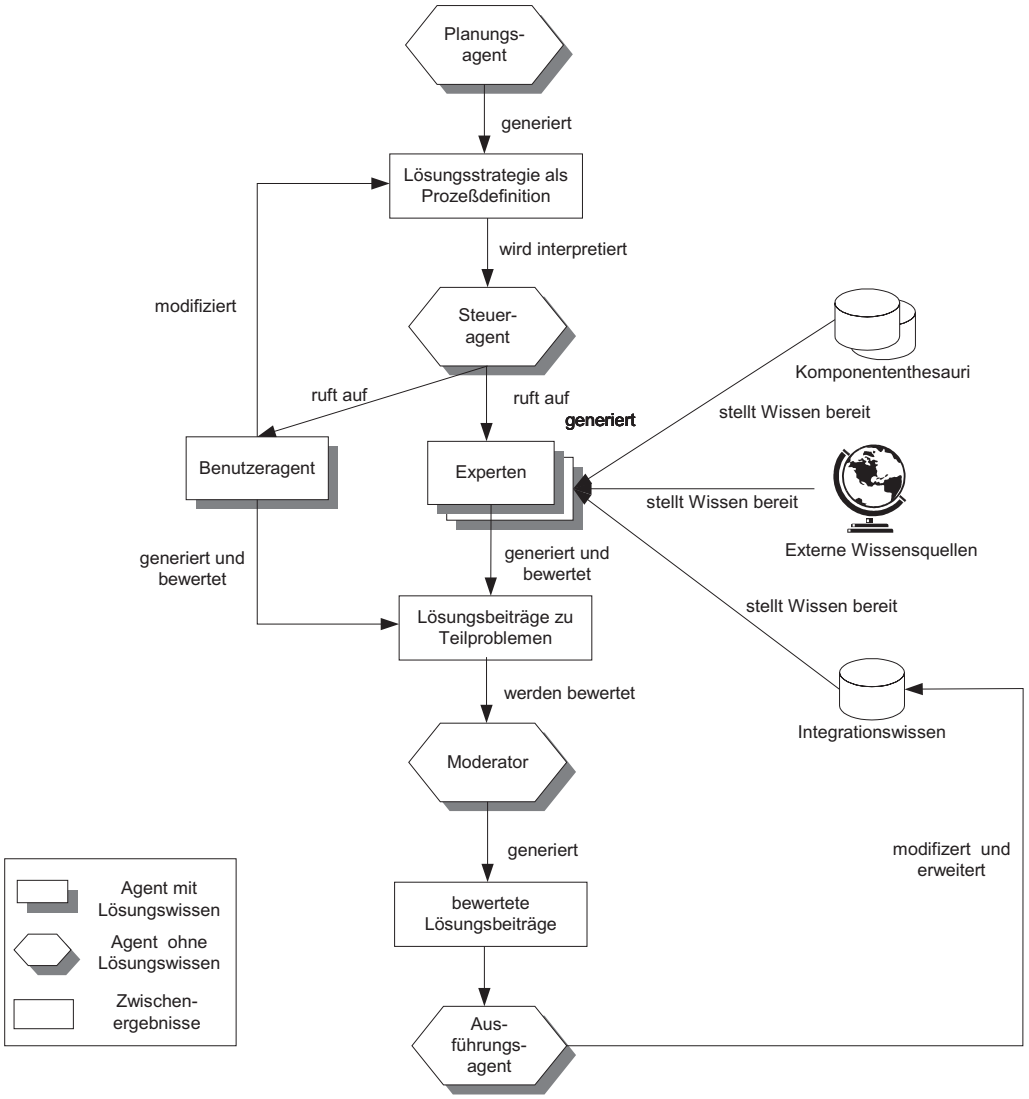

Abbildung 7.3: Prozessorientierte Darstellung des Zusammenspiels der Agenten

7.3.2 Blackboard und Blackboard-Einträge

Das Blackboard ist das zentrale Kommunikationsmedium und globaler Informationsträger in FA^{2}ITH. Wir unterteilen das Blackboard funktional in zwei Bereiche (vgl. Abbildung 7.4), die als eigenständige Blackboards betrachtet werden können:

Domänen-Blackboard: Das Domänen-Blackboard verwaltet alle Informationen, die von den Experten und dem Benutzeragent zur Lösung von Teilproblemen beigetragen werden. Das Domänen-Blackboard kann selbst wiederum in zwei Bereiche unterteilt werden: ein Bereich ist für die Verwaltung von Hypothesen – das sind Vorschläge mit Bewertungen der Experten, über deren Akzeptanz oder Ablehnung noch nicht entschieden ist – verantwortlich. Ein weiterer Bereich verwaltet Fakten, also Hypothesen, die entweder akzeptiert oder abgelehnt worden sind und somit als gesichert gelten und nicht von weiteren Experten bewertet werden müssen.

Kontroll-Blackboard: Das Kontroll-Blackboard nimmt Steuerungs- und Verwaltungsinfor-

mationen auf. Anhand der Informationen dieses Blackboards kann der Steueragent die Experten entsprechend ihrer Eignung mit der Lösung von Teilproblemen beauftragen. Das Kontroll-Blackboard kann in drei Bereiche unterteilt werden: Eine Registratur verwaltet die Definition des vorgesehenen Integrationsprozesses und alle Anmeldungen von Experten, die zur Lösung des Problems beitragen wollen. Weitere Bereiche ermöglichen es, Kontexte und kontextbezogene Bewertungen zu den Experten abzulegen.

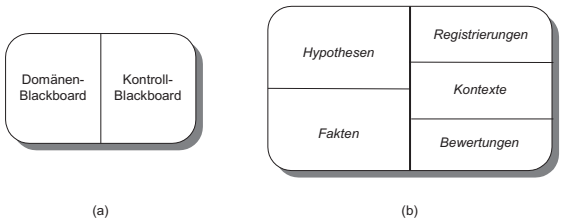

Abbildung 7.4: Struktur des Blackboard von FA^{2}ITH: (a) funktionale Unterscheidung, (b) Blackboard-Einträge

Diese erste Unterteilung des Blackboards ermöglicht die dezentrale Speicherung und bietet somit das Potential, ausreichende Verfügbarkeit und Performanz zu gewährleisten. Eine weitere Unterteilung der Blackboard-Bereiche ist für die – aufgrund des großen Datenvolumens kritischen – Bereiche für die Verwaltung der Hypothesen und Fakten möglich. Dieser Aspekt wird in Abschnitt 7.3.2.1 wieder aufgegriffen.

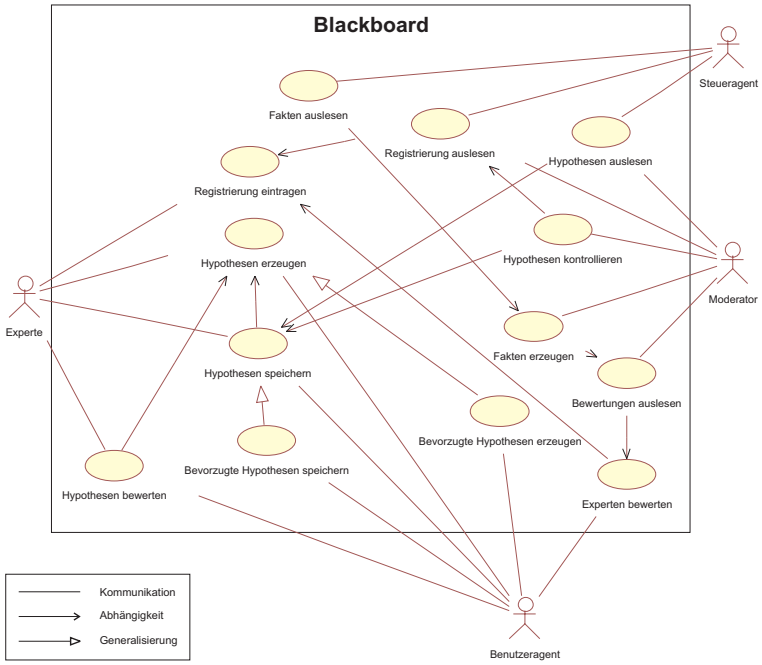

Abbildung 7.5: Anwendungsfall-Diagramm zum System Blackboard

Die Abläufe und Abhängigkeiten, die bei der Verwendung des Blackboards durch die Agenten entstehen, sind mittels UML-Anwendungsfall-Diagramm in Abbildung 7.5 zusammenfassend dargestellt (auf den Ausführungsagenten sowie die Kontexte wird an dieser Stelle aus Gründen der besseren Lesbarkeit verzichtet). So kann sich beispielsweise der Experte beim Blackboard registrieren (Registrierung eintragen), Hypothesen erzeugen und speichern sowie vorhandene

Hypothesen bewerten. Nur dem Benutzeragenten ist es möglich, bevorzugte Hypothesen, also Hypothesen, die höher priorisiert sind, abzulegen.

7.3.2.1 Hypothesen

Durch Hypothesen sollen Veränderungen an einem Komponententhesaurus oder an der Thesaurusföderation bewirkt werden. Solche Veränderungen können für einen Komponententhesaurus beispielsweise die Transformation des Komponententhesaurus-Modells in unser Thesaurus-Modell sein oder auch die Normierung von Benennungen und Definitionen. Auf die Thesaurusföderation bezogen sind solche Veränderungen

- das *Einfügen* oder *Entfernen* einer *Inter-Thesaurus-Beziehung*,
- das *Einfügen* oder *Entfernen* eines *Ergänzenden Begriffes*,
- das *Einfügen* oder *Entfernen* einer *Konfliktmarkierung*.

Modifikations-Operationen können durch die Kombination einer Einfüge- und einer Entferne-Hypothese ausgedrückt werden.

Die Hypothesen werden von den Experten bzw. vom Benutzeragent erzeugt, vom Blackboard verwaltet und von diesem zur Weiterverarbeitung durch die Agenten bereitgestellt. Zentraler Bestandteil einer Hypothese ist die durch die Hypothese dargestellte Veränderung, formuliert als *These* oder *Behauptung* (engl. assertion). Die Information dieser Behauptungen unterscheidet sich je nach dem Typ der Veränderung, die bewirkt werden soll. Beispielsweise enthält eine 1:n-Inter-Thesaurus-Beziehung Verweise auf die in Beziehung stehenden Begriffe sowie eine Spezifikation des Typs der Beziehung.

Jede Hypothese besitzt zudem einen Zeitstempel (Erzeugungszeitpunkt) sowie die Möglichkeit, diese als *bevorzugte Hypothese* zu markieren. Eine solche Markierung darf ausschließlich durch den Benutzeragenten vorgenommen werden. Bevorzugte Hypothesen ermöglichen eine bevorzugte – im Sinne von beschleunigte – Bewertung dieser Hypothesen durch die Experten. Dies ist erforderlich, da der erfolgreiche Fortgang des Integrationsprozesses vom menschlichen Experten abhängt und diesem möglichst rasch die benötigten Informationen für seine Entscheidungsfindung geliefert werden sollen.

Die weiteren Informationen, die Bestandteil einer Hypothese sind, hängen davon ab, ob die Hypothese auf dem Domänen-Blackboard abgelegt oder einem Experten zur Bearbeitung übergeben ist:

Hypothesen: Hypothesen, die von einem Experten resp. dem Benutzeragenten erzeugt oder bewertet werden, nennen wir der Einfachheit halber *Hypothesen*. Diese können durch genau einen Experten bewertet werden und Auskunft über die bisherige Gesamtbewertung geben. Für eine detaillierte Betrachtung der Bewertungsmöglichkeiten sei auf Abschnitt 7.4 verwiesen.

Blackboard-Hypothesen: Hypothesen werden als *Blackboardhypothesen* auf dem Blackboard gespeichert. Dabei werden die Hypothesen, die von den Experten oder dem Benutzeragenten erzeugt und bewertet wurden, derart zusammengefasst, dass bei jeder *neuen* Behauptung, die auf dem Blackboard auftritt, eine neue Blackboardhypothese erzeugt wird. Zu dieser einen Behauptung können nun beliebig viele Hypothesenbewertungen der Experten resp. des Benutzeragenten abgespeichert werden. Somit sieht sich der Moderator in die

Lage versetzt, zu einer Behauptung sofort über all ihre Bewertungen zu verfügen. Zu jeder Hypothesenbewertungen ist der die Bewertung abgebende Experte und der Zeitpunkt der Bewertungsabgabe bekannt. Wird eine Hypothese von einem Experten mehrfach bewertet, wird ausschließlich die letzte Bewertung berücksichtigt.

7.3.2.2 Fakten

Wenn der Moderator zu einer Blackboard-Hypothese aufgrund der verschiedenen durch die Experten resp. den Benutzeragenten zu einer Hypothese abgegebenen Teilbewertungen eine Gesamtbewertung berechnet (vgl. Abschnitt 7.4), die die Behauptung einer Hypothese akzeptiert bzw. ablehnt, wird aus dieser Hypothese ein *Faktum* mit dem Status *akzeptiert* bzw. *abgelehnt*. Fakten gelten als gesichert und müssen nicht mehr bewertet werden. Sie können gleichwertig mit anderen Informationen zur Bewertung offener Hypothesen herangezogen werden. Die Ausführung der aus dem Faktum zu schließenden Aktion in Hinsicht auf die Veränderung von Komponententhesauri oder der Thesaurusföderation wird vom Ausführungsagenten übernommen (vgl. Abschnitt 7.3.8).

7.3.2.3 Registrierungen

Innerhalb eines eigenen Blackboardbereiches, der Registratur, können zwei verschiedene Typen von Registrierungen abgelegt werden:

Aufgaben-Registrierungen: Welche Teilaufgaben bei der Thesaurusintegration zu lösen sind und wie diese Teilaufgaben voneinander abhängen, kann durch Aufgaben-Registrierungen beschrieben werden. Damit bietet die Aufgaben-Registratur die Möglichkeit, den Integrationsprozess in Form einer *Aufgaben-Agenda* zu definieren. Da eine solche Aufgaben-Agenda große Ähnlichkeiten mit einer Workflow-Definition hat, übertragen wir Workflow-Konzepte zur Definition der Aufgaben-Agenda und können dadurch auf fortgeschrittene Methoden und Standards aufsetzen. Die Aufgaben-Agenda kann als eine Menge von Workflows angesehen werden.

Die bedeutendsten Standardisierungsbestrebungen im Workflow-Bereich gehen von der Workflow Management Coalition (WfMC), einem Zusammenschluss namhafter Hersteller, z.B. IBM, SNI, Staffware, Novell, aus [Wor00a]. Innerhalb des Workflow-Referenz-Modelles [Hol95] werden fünf funktionale Schnittstellen zu Workflow-Diensten standardisiert. Relevant für die Prozessdefinition ist die Schnittstelle 1 [Wor99]. Als Bestandteil dieser Schnittstelle wird die Workflow-Prozess-Definitionssprache (engl. Workflow Process Definition Language oder kurz WPDL) als formale Sprache zur Definition und zum Austausch von Prozessdefinitionen spezifiziert. Durch die Standardisierung von WPDL wird der Austausch von Prozessdefinitionen zwischen verschiedenen Systemen möglich, wodurch an dieser Stelle die Offenheit der Wissensakquisitionsarchitektur gewährleistet wird. Um die Mächtigkeit von WPDL zur Definition der Aufgaben-Agenda zu reduzieren, definieren wir eine eigene Aufgaben-Agenda-Definitions-Sprache, AADS, die auf WPDL basiert(vgl. Anhang A).

Aufgaben-Agenden können für Standard-Aufgaben (z.B. Einfügen eines neuen Thesaurus in die Föderation oder Entfernen eines Komponententhesaurus aus der Föderation) vorgegeben werden. Individuelle Anpassungen oder völlig neue Aufgaben-Agenden sind aber möglich. Die Erstellung bzw. Anpassung der Aufgaben-Agenda ist Aufgabe des Pla-

nungsagenten in Zusammenarbeit mit dem menschlichen Experten (vgl. Abschnitt 7.3.6) während der Phase der Festlegung der Integrationsstrategie.

Experten-Registrierungen: Jeder Experte, der an der Problemlösung teilnehmen möchte, muss sich beim Blackboard anmelden. Eine solche Experten-Registrierung beinhaltet einen Identifikator des Experten, eine Beschreibung der erforderlichen Eingaben sowie der möglichen Ausgaben (vgl. Anhang B), Verweise auf Profile, die kontextabhängig (vgl. Abschnitt 7.3.2.5) zur Konfiguration des Experten verwendet werden sollen, sowie Informationen zum Aufruf des Experten.

Die Experten-Registrierungen stellen zusammen mit der Aufgaben-Agenda eine wichtige Basis für den Steueragenten dar, um eine angemessene Problemlösungsstrategie zu implementieren.

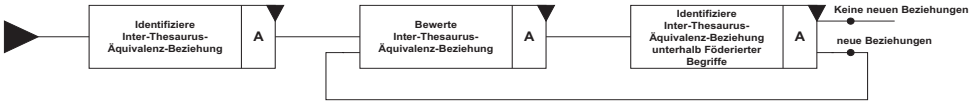

Abbildung 7.6: Ausschnitt einer Aufgaben-Agenda mit Schleife

Beispiel 7.1 *In Abbildung 7.6 wird ein einfacher Ausschnitt aus einer Aufgaben-Agenda dargestellt.*

7.3.2.4 Bewertungen

Ein weiterer Blackboard-Bereich ermöglicht es, kontextbezogene Bewertungen zu den Experten abzulegen. Wie diese Bewertungen an sich aussehen, wird detailliert in Abschnitt 7.4.2 erläutert. Der Bezug zu einem Kontext wird durch Verweis auf einen solchen hergestellt. Ggf. ist ein neuer Kontext zu erzeugen.

7.3.2.5 Kontexte

Ein Experte, der bei der Integration eines Thesaurus in eine Thesaurusföderation hervorragende Arbeit leistet, kann bei der Integration eines anderen Thesaurus in die gleiche Föderation vielleicht nur mittelmäßige Arbeit beitragen[2]. Zudem soll es möglich sein, Experten für verschiedene Kontexte zu konfigurieren. Daher ist es erforderlich, Kontexte explizit zu repräsentieren und Expertenbewertungen sowie die Konfiguration dieser Experten vom Kontext abhängig zu machen.

Unter *Kontext* verstehen wir eine Menge von Fakten oder Gegebenheiten, die eine bestimmte Situation oder ein Ereignis umgeben. Ein Kontext-Modell muss also die Gegebenheiten um eine Situation herum im Prozess der Thesaurusintegration beschreiben. Dazu sieht unser einfaches Kontext-Modell folgende Informationen vor:

Phase: Die Phase innerhalb des Vorgehensmodells.

Aufgabe: Die Aufgabe, die innerhalb der Phase durchgeführt wird.

[2]Ein in der Literatur häufig genanntes Beispiel ist ein Spracherkennungssystem, dass zur Erkennung von amerikanischem Englisch entwickelt wurde und für britische Sprecher wesentlich schlechtere Ergebnisse liefert. Eine explizite Berücksichtigung dieses Kontextes kann zu deutlich besseren Ergebnissen führen, wenn zusätzlich die Unterschiede der Akzente berücksichtigt werden können.

Thesaurusföderation: Den Zustand der Thesaurusföderation. Dieser wiederum setzt sich zusammen aus dem Identifikator der Föderation und für jeden Komponententhesaurus einer Menge von Tupeln, die den Identifikator des Komponententhesaurus, die Version des Komponententhesaurus, den Zustand der Integration (*Integrationsprozess begonnen*, *Integration fertig*, *Integration verbessert*) und den Zeitpunkt, zu dem der Integrationszustand erreicht wurde, beinhaltet. Wenn kein Föderationsidentifikator angegeben wird, bezieht sich der Kontext auf einen einzelnen Komponententhesaurus, der dann als einziges Tupel angegeben ist. Somit besteht die Möglichkeit, diesen Kontext auch in der Vorbereitungsphase zu verwenden.

Darüber hinausgehende Kontext-Informationen, die nicht in direktem Zusammenhang mit der Thesaurusföderation und den Komponententhesauri stehen, sondern mit den einzelnen Agenten (z.B. welche Wissensquellen in welchen Versionen verwendet werden), werden nicht berücksichtigt. Ergeben sich hier grundlegende Veränderungen, wird erwartet, dass der Agent sich abmeldet und neu anmeldet, so dass er dann wie ein neuer Agent berücksichtigt werden kann.

Das Problem des Transfers in einen neuen Kontext würde den Rahmen dieser Arbeit sprengen. Stattdessen sei exemplarisch auf die Arbeiten von Pratt [Pra93] und von Widmer [WK96] verwiesen (beide im Zusammenhang von Kontexten für Neuronale Netze).

7.3.3 Planungsagent

Aufgabe des Planungsagenten ist die Erstellung der Aufgaben-Agenda für die Integration einer vorgegebenen Menge von Thesauri. Da die Zerlegung des Gesamtproblemes in Teilprobleme eine große intellektuelle Herausforderung ist, kann diese Aufgabe nicht ausschließlich von einem Software-Agenten ausgeführt werden. Stattdessen wird die wesentliche Aufgabe, das eigentliche Zerlegen in Teilprobleme, vom menschlichen Integrationsexperten übernommen. Das Gerüst hierfür liefern wir im Rahmen dieser Arbeit. Der Planungsagent interagiert somit – zumindest indirekt – mit den menschlichen Experten (vgl. auch Abschnitt 7.3.5). Die wesentliche Aufgabe, die der Planungsagent übernimmt, ist, die Planung anhand eines von den menschlichen Experten erstellten übergeordneten Gesamtplanes für ein aktuell vorliegendes Szenario der Integration einer gegebenen Menge von Komponententhesauri mit ihren spezifischen Eigenschaften anzupassen.

Die Schnittstellen des Planungsagenten ermöglichen zum einen den Zugriff auf den übergeordneten Gesamtplan, der ebenfalls in AADS ausgedrückt werden kann und Bestandteil der Konfiguration des Planungsagenten ist. Zum anderen erzeugt und modifiziert er die Aufgaben-Agenda auf dem Blackboard für einen vorgegebenen Fall der Integration einer Menge von Komponententhesauri. Dazu erhält der Planungsagent zusätzlich zu dem Gesamtplan Zugriff auf die Ergebnisse der Analysephase.

7.3.4 Experten

Die Experten repräsentieren gleichrangige Verfahren für die Thesaurusintegration. Sie besitzen das Problemlösungswissen, mit dem Teilprobleme gelöst werden können, indem Hypothesen aufgestellt und bewertet werden. Experten, die sich an der Problemlösung beteiligen wollen, müssen sich mit der Beschreibung der Aufgabe, zu deren Lösung sie beitragen können, beim Blackboard anmelden (vgl. Abschnitt 7.3.2.3).

Die interne Repräsentation und die Schlussfolgerungsmechanismen, die ein Experte nutzt, sind gekapselt, also nach außen hin nicht sichtbar. Experten werden somit als Black-Box betrachtet,

die mittels geeigneter Eingaben zu Ergebnissen kommen, die als Hypothesen und Hypothesenbewertungen Teillösungen zur inkrementellen Erstellung der Gesamtlösung beitragen. Experten können daher z.B. als regelbasiertes System, als auf deskriptiven Sprachen basierendes System oder als objektorientiertes System implementiert sein, ohne dass dies Auswirkungen auf seine Integrationsfähigkeit in das Gesamtsystem hat.

Das Interaktionsmodell für die Experten sieht vor, dass diejenigen Experten, die auf dem Blackboard registriert sind, ausschließlich über das Blackboard miteinander kommunizieren können. Es ist jedoch zulässig, dass Experten, die ein Teilproblem zur Bearbeitung übertragen bekommen haben, wiederum andere Agenten beauftragen können, die Aufgabe vollständig bzw. in Teilen in ihrem Namen zu lösen.

Voraussetzung für die Teilnahme eines Experten am Problemlösungsprozess ist das Vorhandensein der folgenden Schnittstellen:

Konfigurationsschnittstelle: Die Konfigurationsschnittstelle muss die kontextbezogene Konfiguration des Experten ermöglichen. Um diese Schnittstelle so allgemein und so einfach wie möglich zu halten, wird über sie nur der Name der entsprechenden Konfigurationsdatei mitgeteilt. Diese wiederum enthält die Experten-spezifischen Einstellungen, die jeweils für einen Kontext gelten, als Namen-Wert-Paare.

Steuerungsschnittstelle: Die Steuerungsschnittstelle ermöglicht es dem Steueragenten, den Kontrollfluss zu beherrschen. Er kann über diese Schnittstelle die Experten beauftragen, im Sinne einer bestimmten Aufgabe neue Hypothesen zu erzeugen oder vorhandene Hypothesen zu bewerten. Im ersten Fall ist es notwendig, dazu den Experten die zu betrachtenden Integrationsobjekte (die Thesaurusföderation oder Komponententhesauri) mitzuteilen. Im zweiten Fall (Hypothesenbewertung) erhält er die zu bewertenden Hypothesen. Als Ergebnis liefert er in beiden Fällen eine Menge von bewerteten Hypothesen.

Damit die Möglichkeit besteht, einen Experten abzubrechen, da z.B. seine Bearbeitungszeit zu lang ist, soll der Experte ebenfalls eine Methode implementieren, die den Abbruch seiner weiteren Arbeit initiiert.

Experten können über die Schnittstellen des Blackboards, der Thesauri, der Thesaurusföderation und weiterer Quellen auf diese zugreifen. Welche Quellen der Experte zur Bearbeitung seiner Aufgabe heranzieht, bleibt ihm weitestgehend überlassen. So ist auch die Berücksichtigung von Zwischenständen auf dem Blackboard möglich, nicht allerdings der Zugriff auf die Registrierungs- und Bewertungsbereiche des Kontroll-Blackboards. Diese Einschränkung verhindert eine Manipulation der Hypothesenbewertungen durch die Experten aufgrund der Kenntnis ihrer eigenen Bewertungen im System.

Aufgrund der großen möglichen Bandbreite bei der Realisierung von Experten ist die Konfiguration eines Experten in der Regel ein individueller Vorgang. Gemeinsam aber sind die Grundlagen, anhand derer die Konfiguration vorgenommen werden kann:

Analyseergebnisse: Anhand der Ergebnisse der Analyse der Komponententhesauri (vgl. Kapitel 9) werden Schlussfolgerungen über die spezifischen Eigenschaften der Thesauri gezogen. Diese Ergebnisse können bei der individuellen Konfiguration berücksichtigt werden.

Testdaten: Eine Teilmenge der zu bewertenden Daten bzw. der bei der Bewertung zu berücksichtigenden Daten wird ausgewählt. Anhand der Ergebnisse, die ein Experte für diese Testdaten erzielt, kann der menschliche Integrationsexperte die Konfiguration optimieren.

Eine weitere Gemeinsamkeit der Expertenkonfiguration ist, dass abhängig von den zur Beurteilung einer Hypothese herangezogenen Informationen festgelegt werden kann, wie sicher sich der Experte bei der Beurteilung ist (vgl. auch Abschnitt 7.4.1).

In den folgenden Kapiteln 8 bis 11 werden eine Reihe von Experten für die Thesaurusintegration vorgestellt.

7.3.5 Benutzeragent

Der *Benutzeragent* ist ein unter allen Agenten ausgezeichneter Agent, da er das Wissen des menschlichen Experten während des Problemlösungsprozesses beisteuert. Aufgrund der großen Bedeutung, die damit dem Benutzeragenten zukommt, widmen wir uns ihm an dieser Stelle ausführlicher.

7.3.5.1 Anforderungen

An den Benutzeragenten werden folgende Anforderungen gestellt:

Einbringen des Expertenwissens: Über den Benutzeragenten soll der menschliche Experte am Problemlösungsprozess aktiv teilnehmen können. Das Spektrum des Wissens, das der menschliche Experte dazu einbringen können soll, geht dabei über das der Software-Experten hinaus. Es kann wie folgt klassifiziert werden:

Hypothesen und Hypothesenbewertungen: Wie die Software-Experten kann der menschliche Experte Hypothesen erzeugen bzw. bewerten. Diesen Hypothesen kommt aber eine besondere Bedeutung zu: Ist sich der menschliche Experte sicher, dass eine Hypothese richtig oder falsch ist, wird diese, ohne dass eine Bewertung durch andere Experten erfolgt, zum Faktum. Ist sich der Experte nicht sicher über die Gültigkeit der Behauptung seiner Hypothese, soll diese vorrangig bearbeitet werden.

Die Formulierung der Behauptungen der Hypothesen soll dabei für den menschlichen Experten intuitiv möglich sein.

Einwirkung auf die Prozesssteuerung: Dem menschlichen Experten soll es möglich sein, aufgrund seiner Bewertung der aktuellen Situation steuernd in den Prozess einzugreifen. Dies soll durch die Entscheidung über die Terminierung oder das Wiederaufsetzen der Bearbeitung einer Teilaufgabe, durch die Modifikation der Aufgaben-Agenda oder – indirekter – durch die Bewertung eines Software-Experten für einen Kontext geschehen können.

Die Konfiguration der Software-Experten (vgl. S. 123) für einen bestimmten Kontext, die ebenfalls durch einen menschlichen Experten angepasst wird, wird an dieser Stelle als eine Aufgabe angesehen, die nicht zur Zuständigkeit des Benutzeragenten gehört. Stattdessen wird angenommen, dass es für die unterschiedlichen Experten individuelle Konfigurationsverfahren gibt.

Der menschliche Integrationsexperte kann somit sowohl sein Problemlösungswissen einbringen als auch in die Problemlösungsstrategie eingreifen. Während des Integrationsprozesses kann damit möglichst umfassend von seiner Expertise profitiert werden.

Darstellen des aktuellen Zustandes der Thesaurusföderation: Damit der menschliche Experte sein Wissen angemessen einbringen kann, ist es erforderlich, dass er sich stets einen Überblick über den aktuellen Zustand der Thesaurusföderation verschaffen kann.

Dazu soll es möglich sein, innerhalb der Föderation nach Benennungen und Begriffen zu suchen und das Begriffsnetz explorativ zu erkunden. Berücksichtigt werden sollen sowohl die Komponententhesauri als auch das in Form von Fakten eingebrachte zusätzliche Integrationswissen der Thesaurusföderation über Inter-Thesaurus-Beziehungen, Ergänzende Begriffe und Konflikte.

Darstellen des Integrationsfortganges: Über den aktuellen Zustand der Thesaurusföderation hinaus soll auch der Fortgang des Integrationsprozesses dargestellt und analysiert werden können. Das bedeutet zum einen, dass eine Darstellung der Aufgaben-Agenda und deren Bearbeitungszustand ersichtlich sein soll. Zum anderen sollen die aktuellen Hypothesen und deren Auswirkungen bis hin zu den durch die Hypothesen verursachten Konflikte dargestellt werden können.

7.3.5.2 Lösungsansatz

Um diese Anforderungen zu erfüllen, wird der Benutzeragent mit einer Darstellungs- und einer Erfassungskomponente ausgestattet. Beide bestehen wiederum aus weiteren Teilkomponenten:

Darstellungskomponente: Die Darstellungskomponente kann gemäß der Anforderungen in zwei Teilkomponenten zerlegt werden:

Föderationsdarstellungskomponente: Die Föderationsdarstellungskomponente ermöglicht über eine einfache Suche, Begriffe auszuwählen. Das Netzwerk von Intra- und Inter-Thesaurus-Beziehungen um einen ausgewählten Begriff herum wird angezeigt; innerhalb des Begriffsnetzes kann navigiert werden. Um die große Menge an Informationen angemessen darstellen zu können, ist der zentrale Bestandteil der Föderationsdarstellungskomponente eine interaktive Fischaugendarstellung.

Innerhalb der Föderationsdarstellungskomponente kann ausgewählt werden, welche Arten von Hypothesen (mit den zugehörigen Auswirkungen) zusätzlich angezeigt werden können. Eine farbliche und grafische Codierung markiert diese Hypothesen in der Darstellung.

Aufgabendarstellungskomponente: Die Aufgaben-Agenda mit den einzelnen Aufgaben wird von einer eigenen Darstellungskomponente als Graph angezeigt. Sie ermöglicht es dem menschlichen Experten, den aktuellen Stand im Integrationsprozess zu beurteilen. Da die Aufgaben-Agenda eine Menge von Workflows in einem für Workflow-Management-Systeme gängigen Format spezifiziert (vgl. Abschnitt 7.3.2.3, Aufgaben-Registrierungen), kann dabei auf Monitoring-Werkzeuge von Workflow-Management-Systemen zurückgegriffen werden.

Erfassungskomponente: Auch die Erfassungskomponente wird in mehrere Teilkomponenten zerlegt:

Erfassungskomponente für Hypothesen: Diese Teilkomponente ermöglicht die Formulierung von Hypothesen einschließlich ihrer Bewertung. Dazu werden die unterschiedlichen Typen von Behauptungen innerhalb der Hypothesen und deren entsprechende Struktur der Erfassungskomponente bekannt gemacht. Da sich die Behauptungen außer im Behauptungstyp im Wesentlichen durch die Anzahl der Felder unterscheiden, die ein oder mehrere Begriffe bzw. ein oder mehrere Beziehungen aufnehmen, kann diese Bekanntmachung über wenige Metadaten geschehen. Bei Ergänzenden Begriffen gilt zusätzlich zu berücksichtigen, dass der Ergänzende Begriff an sich

zu erfassen ist. Der Ansatz der Beschreibung der Behauptungstypen über Metadaten eröffnet die Möglichkeit der einfachen Erweiterung um weitere Behauptungstypen.

Durch die Integration mit der Darstellungskomponente kann eine intuitive Spezifikation der relevanten Beziehungen und Begriffe durch Auswahl von Objekten in der graphischen Darstellung angeboten werden.

Steuerungskomponente: Die Steuerungskomponente beinhaltet die Funktionalität einer einfachen Modellierungskomponente eines Workflow-Management-Systems (Prozess-Definitions-Werkzeug), um direkt in die Aufgaben-Agenda eingreifen zu können. Da die Aufgaben-Agenda als Menge von Workflows definiert ist, kann an dieser Stelle wiederum auf Standards, die durch die Workflow Managament Coalition (WfMC) [Wor00a] spezifiziert und die inzwischen von verschiedenen Herstellern implementiert wurden, zurückgegriffen werden. Dies gilt ebenso für das Kontrollieren der Prozesssteuerung im Sinne der Terminierung einer Teilaufgabe bzw. des Wiederaufsetzens bei einer bereits zuvor bearbeiteten Teilaufgabe – beides klassische Aufgaben eines menschlichen Überwachers eines Workflow-Ausführungsdienstes [Hol95].

Auch der Benutzeragent kann vom Steueragenten zur Erzeugung oder Bewertung von Hypothesen aufgefordert werden. Daher implementiert er wie die anderen Experten eine Steuerungsschnittstelle.

7.3.6 Steueragent

Die Hauptaufgaben des *Steueragenten* bestehen darin, den Prozessablauf zu beherrschen und steuernd auf ihn einzuwirken, um den Problemlösungsprozess opportunistisch voranzubringen. Dazu beauftragt er auf dem Blackboard registrierte Experten mit der Lösung von Teilproblemen aus deren Aufgabengebiet. Zur Auswahl der geeignetsten Experten interpretiert er die Aufgaben-Agenda und greift auf die Registrierungen der Experten und deren Bewertungen auf dem Kontroll-Blackboard zu. Im Besonderen hat der Steueragent die Verantwortung, zu erkennen, ob die auf dem Blackboard registrierten Experten auch alle Aufgabenbereiche, die zur Abarbeitung der Aufgaben-Agenda notwendig sind, abdecken.

Der Steueragent hält intern Steuerungsdaten, die u.a. den Zustand der unterschiedlichen Prozesse und Aktivitäten darstellen. Diese Steuerungsdaten werden, insofern sie die Aufgaben-Agenda betreffen, bei Zustandsänderungen ebenfalls auf das Blackboard geschrieben. So können andere Agenten – insbesondere der Benutzeragent – Informationen über den aktuellen Bearbeitungsstand abrufen.

7.3.7 Moderator

Die Hauptaufgaben des *Moderators* bestehen darin, die qualitative Überprüfung von Blackboardeinträgen vorzunehmen und Hypothesen zu Fakten werden zu lassen. Dazu benötigt er Zugriff auf die Komponententhesauri, das Integrationswissen, die Bewertungen der Hypothesen sowie die Bewertungen der Experten. Das Bewertungsmodell legt dabei fest, wie eine Hypothese zum Faktum wird (vgl. Abschnitt 7.4). Zur qualitativen Überprüfung kann eine Menge von Regeln vorgegeben werden, die die Qualitätskriterien definieren. Beispielsweise kann eine Regel lauten, dass sich widersprechende Fakten (z.B. soll zwischen zwei Begriffen eine Äquivalenzbeziehung und eine Abstraktionsbeziehung etabliert werden) nicht den Qualitätskriterien genügen.

Kann auch nach Ausführung aller Agenten über eine Hypothese nicht entschieden werden, ob sie zum Faktum wird, gibt es verschiedene Möglichkeiten zu deren weiterer Behandlung:

- Die Hypothese wird vom Blackboard gelöscht und in weiteren Verfahrensschritten nicht mehr berücksichtigt. Dies ist z.B. sinnvoll, wenn sehr unsichere Hypothesen über Inter-Thesaurus-Beziehungen aufgestellt wurden, die für die weitere Betrachtung nicht mehr sinnvoll sind.

- Die Hypothese geht als unsicheres Wissen in die Bewertung anderer Hypothesen ein. Dabei kann die Gesamtbewertung der Hypothesen berücksichtigt werden.

- Über die Hypothese soll auf jeden Fall entschieden werden, da eine solche Entscheidung für den weiteren Verlauf erforderlich ist und Fehlentscheidungen hingenommen werden können (Beispiel: Klassifikation einer Hierarchiebeziehung als Abstraktions- oder Bestandsbeziehung). Ausschlaggebend ist im Zweifelsfalle ein ausgezeichneter Standardagent, dessen Bewertung, die nicht unentschieden sein darf, zu einer Entscheidung führt.

Unser Blackboard unterstützt alle drei Möglichkeiten, um die erforderliche Flexibilität zu gewährleisten, dass der menschliche Experte (oder auch der Moderator) in unterschiedlichen Situationen unterschiedliche Entscheidungen über das weitere Vorgehen trifft. Bei Vorhandensein eines Standardagenten wird generell die dritte Alternative ausgeführt.

7.3.8 Ausführungsagent

Die anhand der Fakten festgestellten Veränderungen an der Thesaurusföderation werden vom Ausführungsagenten durchgeführt, d.h., der Ausführungsagent macht die Implikationen dieser Fakten persistent. Dazu müssen dem Ausführungsagenten alle Behauptungstypen und die entsprechenden Implikationen einer Akzeptanz bzw. Ablehnung bekannt sein.

Die Bekanntmachung kann durch eine Menge einfacher Regeln geschehen, die zur Durchführung der Veränderungen an der Thesaurusföderation interpretiert werden. Der Einfluss der Veränderung auf Konflikte (Treten weitere Konflikte auf oder können Konfliktmarkierungen entfernt werden?) ist zu überprüfen. Nach der Interpretation eines Faktums wird dieses vom Blackboard entfernt.

Beispiel 7.2 *Bei Inter-Thesaurus-Äquivalenzbehauptungen sind, abhängig davon, ob die Behauptung akzeptiert oder abgelehnt wurde und ob eine solche Beziehung bereits in der Thesaurusföderation vorhanden ist, drei unterschiedliche Aktionen für den Ausführungsagenten erforderlich. Die entsprechenden Regeln sind in Tabelle 7.4 aufgeführt (Bedingungen in den ersten drei Spalten, Aktionsteil in der vierten Spalte).*

Behauptungstyp	akzeptiert	bereits vorhanden	Aktion
ITÄ-Beh.	ja	ja	-
ITÄ-Beh.	ja	nein	ITÄ-Bez. einfügen, Konfliktbewertung
ITÄ-Beh.	nein	ja	ITÄ-Bez. entfernen, Konfliktbewertung

Fortsetzung auf der nächsten Seite ...

... *Fortsetzung*

Behauptungstyp	akzeptiert	bereits vorhanden	Aktion
ITÄ-Beh.	nein	nein	-

Erläuterung:
ITÄ-Beh. = Behauptung über Inter-Thesaurus-Äquivalenzbeziehung
ITÄ-Bez. = Inter-Thesaurus-Äquivalenzbeziehung

Tabelle 7.4: Regelbasis des Ausführungsagenten (Ausschnitt)

7.4 Bewertungsmodell

Der Grad an Vertrauen, mit dem ein Experte seine Hypothesenbewertung abgibt, variiert je nach vorliegenden Indizien, anhand derer er seine Entscheidungen trifft. So kann eine Hypothese über eine Inter-Thesaurus-Äquivalenzbeziehung bei Thesauri mit vielen Quasi-Synonymen deutlicher bewertet werden, wenn dies anhand der Deskriptor-Benennungen, die die zentrale Bedeutung festhält, geschieht und nicht anhand der Nicht-Deskriptor-Benennungen, die eher im Sinne von Bedeutungserweiterungen zu verstehen sind. Ebenso können verschiedene Regeln, die ein Experte implementiert, unterschiedlich verlässlich sein. Die Experten wiederum als eine Lösungskomponente im System können in unterschiedlichen Kontexten unterschiedlich gute Ergebnisse liefern. Daher können zusätzlich die Experten in einem bestimmten Kontext bewertet werden.

Schließlich ist eine aggregierte Gesamtbewertung erforderlich, um anhand der einzelnen gewichteten Bewertungen, die von den Experten abgegeben wurden, sowie der Expertenbewertungen selbst zu einer Entscheidung zu gelangen, ob eine Hypothese akzeptiert oder abgelehnt werden soll.

Unsere Ansätze zur Bewertung der Experten und Hypothesen sowie zur Aggregation der Einzelbewertungen zu einer Gesamtbewertung werden in den folgenden Abschnitten vorgestellt.

7.4.1 Bewertung der Hypothesen

Eine Gewichtung der Akzeptanz bzw. Ablehnung einer Hypothese durch einen Experten dient folgenden Zielen:

- Der Experte wird nicht gefordert, eine Ja-/Nein-Aussage zu treffen, sondern kann die Sicherheit seiner Aussage durch einen Konfidenzfaktor ausdrücken. Dies ist insofern von Bedeutung, als in der Praxis eine wirklich definitive Aussage über die Zustimmung bzw. Ablehnung einer Hypothese aufgrund der Komplexität des Problemes und der unscharfen Semantik der Begriffe quasi unmöglich ist.
- Der menschliche Experte kann unsicheres Wissen bereitstellen, indem er eine Hypothese einbringt und dieser eine gewichtete Bewertung mitgibt. Das System versucht mithilfe der Software-Experten, diese Hypothesen zu bewerten, mit dem Ziel, sie schließlich zu akzeptieren oder abzulehnen.

Eine solche Hypothesenbewertung kann durch einen *Konfidenzfaktor* (engl. *confidence factor* oder *certainty factor*) CNF_{hyp}, der einen Wert aus dem Intervall $[-1, 1] \in \mathbb{R}$ annehmen kann, ausgedrückt werden. Dabei bedeutet eine Bewertung im Intervall $[-1, 0)$ eine Ablehnung der

Hypothese, die umso stärker ist, je kleiner die Bewertung ist. Eine Bewertung im Intervall $(0, 1]$ bedeutet eine Akzeptanz der Hypothese, die umso größer ist, je größer der Wert ist. 0 schließlich ist die neutrale Bewertung, d.h. es kann weder eine ablehnende noch eine zustimmende Aussage gemacht werden (vgl. Abb. 7.7).

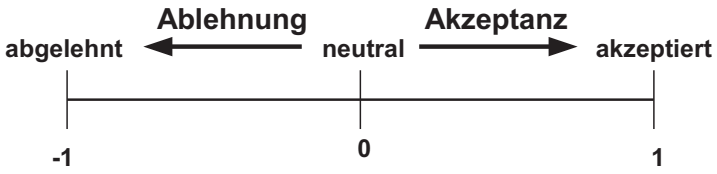

Abbildung 7.7: Wertebereich und Bedeutung der Konfidenzfaktoren für Hypothesen

Es sei darauf hingewiesen, dass Konfidenzfaktoren keine Wahrscheinlichkeiten sind. Sie sind informale Maße des Vertrauens in eine These. Sie repräsentieren den Grad, zu dem wir glauben, diese These sei tatsächlich wahr (vgl. [HK85, S. 42]).

Die Ermittlung der Konfidenzfaktoren für Hypothesen kann auf zwei Arten geschehen:

Manuell: Die Bewertung der Hypothesen durch den menschlichen Experten geschieht intuitiv. Der menschliche Experte kann dabei auf sein Wissen und seine Erfahrungen zurückgreifen. Die Abstufung der Konfidenzfaktoren braucht dabei nicht besonders fein zu sein. Untersuchungen (vgl. [Dav86, S. 958]) haben gezeigt, dass eine grobe Abstufung, z.B. definitiv (1.0), fast sicher (0.7), wahrscheinlich (0.3), vielleicht (0.0), wahrscheinlich nicht (-0.3), fast sicher nicht (-0.7) und definitiv nicht (-1.0) genügen, um Situationen einzuschätzen. Dies ist mit Rücksicht auf den Anwender solcher Systeme ein nicht zu unterschätzender Aspekt. Darüber hinaus kann so das ohne Zweifel vorhandene Risiko der Zuordnung unterschiedlicher Konfidenzfaktoren verschiedener Personen bei gleichen Sachverhalten vermindert werden.

Berechnung: Software-Experten müssen den Konfidenzfaktor berechnen. Dazu sind grundlegende Hinweise in der Wissensbasis dieser Experten notwendig.

Die Berechnung und der Umgang mit Konfidenzfaktoren ist eine Form des Unsicheren Schließens (engl. *uncertain reasoning*). Aufgrund der sehr guten Ergebnisse, die in der Vergangenheit mit regelbasierten Expertensystemen erzielt wurden, sowie der einfachen Berechnung und Anwendbarkeit, wählen wir das Mycin-Modell. Alternativ, aber mit größerem Aufwand, könnten auch Bayessche Netzwerke angewandt werden. Vorteil wäre ein mathematisch fundierteres Modell. Allerdings wird bei zum Teil nur groben Schätzungen der Konfidenzfaktoren bzw. der Probabilitäten durch Anwendung der Bayesschen Netzwerke kein Vorteil erzielt. Hinzu kommt die Komplexität des Ansatzes, wenn mehrere bzw. viele Ereignisse ein und dasselbe Ereignis begründen (Analyse verschiedener Tatsachen zur Begründung einer Aussage). Beides aber sind wesentliche Eigenschaften, wenn die Hypothesenbewertung durch die Experten stattfindet.

Im Mycin-Modell werden Konfidenzfaktoren – obwohl sie im Sinne des Begriffs keine Wahrscheinlichkeiten darstellen – nach den Gesetzmäßigkeiten des Bayes-Theorems (Konzept der bedingten Wahrscheinlichkeiten) berechnet (vgl. [SB85, S. 235] und [Pup88, S. 52], zum Bayes-Theorem selbst [Bor79, S. 72ff]). Prinzipiell treten in regelbasierten Systemen, die wir im Rahmen dieser Arbeit als Grundlage der Software-Experten annehmen, drei Fälle auf, die jeweils unterschiedliche Wege der Berechnung von Konfidenzfaktoren erfordern (vgl. [HK85, S. 51]):

- In zusammengesetzten Prämissen können unter mehreren Klauseln, die jeweils mit logischen Operatoren (UND oder ODER) verbunden sind, unsichere Klauseln sein, z.B. bereits bewertete Hypothesen; eine unsichere Prämisse führt zu einer unsicheren Schlussfolgerung. Bei UND-Verknüpfungen ist der Konfidenzfaktor der Prämisse CNF_p das Minimum der Konfidenzfaktoren der Klauseln, bei ODER-Verknüpfungen das Maximum.

- Eine Regel selbst kann zu weniger als 100 % sicher sein. Der Konfidenzfaktor einer einzelnen Regel CNF_r kann dabei entsprechend dem Konfidenzfaktor für Experten berechnet werden. Er wird in der Regel kontextabhängig sein und muss entsprechend konfiguriert werden können. Sind CNF_p und CNF_r bekannt, berechnet sich $CNF_{hyp} = CNF_p \cdot CNF_r$.

- Identische Folgerungen können durch mehr als eine Regel gezogen werden, d.h. eine Hypothese wird aufgrund verschiedener Regeln bewertet. Regeln mit der gleichen Tendenz (Ablehnung bzw. Akzeptanz) sollen die Bewertung verstärken, Widersprüche sollen zu einer schwächeren Bewertung führen. Diese Anforderungen erfüllt folgende Funktion des Mycin-Modells zur Berechnung des kombinierten Konfidenzfaktors z:

$$z = \begin{cases} x + y - x \cdot y & \text{falls } x, y \geq 0 \\ \frac{x+y}{1-min(|x|,|y|)} & \text{falls } x, y \text{ unterschiedliche Vorzeichen besitzen} \\ x + y + x \cdot y & \text{falls } x, y < 0 \end{cases}$$

  wobei x und y die aufgrund von zwei verschiedenen Regeln berechneten Konfidenzfaktoren für eine Hypothese sind. Die Berechnungsvorschrift nennen wir *Aggregationsformel*. Sie kann sequenziell ausgeführt werden. Für die Extremfälle, bei denen Konfidenzfaktoren von -1 oder 1 angegeben werden, sind Sonderregeln angegeben. Wir betrachten diese nicht weiter, da wir davon ausgehen, dass diese Fälle für die Thesaurusintegration uninteressant sind und wir mit Konfidenzfaktoren, die nahe der Extremfälle gewählt werden, genügend Spielräume haben.

Beispiel 7.3 *Anhand zweier verschiedener und voneinander unabhängiger Regeln bewertet ein Experte eine Hypothese zweimal zustimmend, anhand Regel 1 mit einer Hypothesenbewertung von 0.7 und anhand Regel 2 mit einer Hypothesenbewertung von 0.8. Nach obiger Formel ergibt sich für die Gesamtbewertung durch diesen Experten der Wert 0.94.*

7.4.2 Bewertung der Experten

Eine Bewertung der Experten dient zwei Zielen:

- Die Hypothesenbewertungen der Experten können gewichtet in eine Gesamtbewertung eingehen. Somit kann der unterschiedliche Grad des Vertrauens, der dem Experten in einem bestimmten Kontext entgegengebracht wird, ausgedrückt und berücksichtigt werden.

- Der Steueragent kann eine Auswahl von Agenten anhand der Expertenbewertungen treffen und erhält somit eine zusätzliche Entscheidungsgrundlage für eine opportunistische Strategie.

Wie bereits bei der Hypothesenbewertung kann die Expertenbewertung durch einen Konfidenzfaktor CNF_{exp} geschehen. Der Konfidenzfaktor CNF_{exp} kann einen Wert aus dem Intervall $[0, 1] \in \mathbb{R}$ annehmen. Je größer die Bewertung ist, desto größer ist das Vertrauen, das dem Experten entgegengebracht wird (vgl. Abb. 7.8).

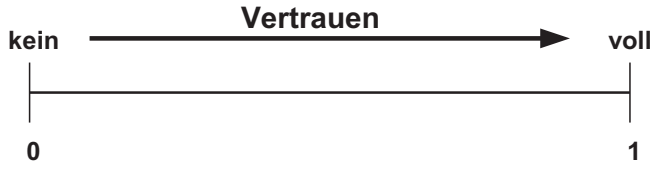

Abbildung 7.8: Wertebereich und Bedeutung der Konfidenzfaktoren für Experten

Der Experte wird für die Qualität seiner erbrachten Leistung bewertet. Diese Bewertung kann auf folgende Arten, die auch kombiniert werden können, geschehen:

- Ein menschlicher Experte betrachtet (stichprobenweise) die Ergebnisse eines Experten und bewertet ihn daraufhin.
- Die Ergebnisse des Experten werden anhand von Testdaten bewertet, bei denen die „richtigen“ Hypothesen bekannt sind.
- Die Ergebnisse des Experten werden mit den Ergebnissen anderer, bereits bewerteter Experten verglichen.
- Wurde der Experte bereits in einem ähnlichen Kontext bewertet, kann diese Bewertung auf den neuen Kontext anhand von Übertragungskriterien übertragen werden. Diese wird im Allgemeinen durch den menschlichen Experten geschehen, der die Ähnlichkeit der Kontexte beurteilen kann.

Der erste und der zweite Fall können beide darauf zurückgeführt werden, dass die Aussagen der Software-Experten verglichen werden mit Aussagen menschlicher Integrationsexperten, die a posteriori (Fall 1) oder a priori (Fall 2) getroffen werden. Da dies in der Regel die Grundlage der Expertenbewertungen sein wird, vertiefen wir an dieser Stelle die Betrachtungen.

Die einfachste Form der Berechnung der Expertenbewertung besteht darin, die Anzahl der Aussagen mit der korrekten Tendenz $n_{korrekt}$ (Ablehnung oder Akzeptanz) zu dividieren durch die Summe der Aussagen mit der korrekten Tendenz und der Aussagen mit der falschen Tendenz n_{falsch}[3]:

$$CNF_{exp} = \frac{n_{korrekt}}{n_{korrekt} + n_{falsch}}$$

Bei einer solchen Berechnung bleibt aber unberücksichtigt, wie sicher sich die Experten bei einer korrekten bzw. falschen Aussage waren (vgl. Abschnitt 7.4.1). Ist die Konfidenz der Experten bei richtigen Aussagen durchschnittlich besser als bei falschen Aussagen, sollte deren Bewertung besser, ansonsten schlechter ausfallen. Um dies zusätzlich zu berücksichtigen, können die durchschnittlichen Konfidenzfaktoren der korrekten ($CNF_{korrekt}^{Mittelwert}$) bzw. falschen Aussagen ($CNF_{falsch}^{Mittelwert}$) berechnet werden und in eine Bewertung einfließen. Dies kann wiederum mit der Formel von Mycin erreicht werden:

$$CNF_{exp} = \begin{cases} x + y - x \cdot y & \text{falls } y \geq 0 \\ \frac{x+y}{1-min(|x|,|y|)} & \text{falls } y < 0 \end{cases}$$

[3] Im Information Retrieval wird das entsprechende Maß Precision genannt (vgl. Anhang D.2, S. 273). Für die von uns betrachtete Verlässlichkeit der Bewertung eines Experten ist dies das entscheidende Maß. Das im Information Retrieval ergänzend verwendete Maß des Recall für die Bewertung der Vollständigkeit spielt bei der Verlässlichkeit keine Rolle.

wobei $x = \frac{n_{korrekt}}{n_{korrekt}+n_{falsch}}$ und immer $x \geq 0$ gilt sowie $y = CNF^{Mittelwert}_{korrekt} - CNF^{Mittelwert}_{falsch}$.

Beispiel 7.4 *In der Vorbereitungsphase bewerten die Experten die Hypothese, dass es sich bei einer Hierarchierelation in AGROVOC um eine Abstraktionsrelation handelt. Die Richtigkeit dieser Bewertung kann für Testdaten anhand deren manueller Klassifikation überprüft werden. Ein Experte bewertet in* 55.7% *der Fälle die Hypothesen korrekt, in* 3.0% *falsch und in* 41.3% *wird keine Bewertung abgegeben. Der gemittelte absolute Konfidenzfaktor bei korrekten Bewertungen ist 0.8, bei falschen Bewertungen 0.5. Der Konfidenzfaktor für diesen Experten berechnet sich wie folgt:*

$$\begin{aligned} CNF_{exp} &= \frac{55.7}{55.7+3.0} + (0.8-0.5) - \frac{55.7}{55.7+3.0} \cdot (0.8-0.5) \\ &= 0.964 \end{aligned}$$

Als Kontext werden die Phase (Vorbereitung), die Aufgabe (Klassifikation Hierarchierelation) und der beteiligte Thesaurus (AGROVOC, Version 3.0) festgehalten.

Da der menschliche Experte schließlich die Entscheidungshoheit besitzt, wird der Benutzeragent immer mit 1 bewertet. In einem kooperativen Szenario, das über den Rahmen dieser Arbeit hinausgeht, kann davon abgewichen werden, wenn auch den menschlichen Experten (kontextabhängig) unterschiedliches Vertrauen in ihre Entscheidung entgegengebracht wird.

7.4.3 Berechnungsmodell für eine aggregierte Gesamtbewertung

Hypothesen können von einem oder auch – und dieser Fall wird angestrebt - mehreren Experten bewertet worden sein. Für die Entscheidung, ob eine Hypothese akzeptiert, abgelehnt oder noch keine Aussage getroffen werden kann, sind die Hypothesenbewertungen der einzelnen Experten sowie Expertenbewertungen zu berücksichtigen. An die Aggregation werden folgende Anforderungen gestellt:

- Neutrale Hypothesenbewertungen (Wert = 0) beinflussen die Gesamtbewertung nicht, wenn für dieselbe Hypothese auch nicht-neutrale Bewertungen vorliegen. Ansonsten ist auch die Gesamtbewertung neutral.
- Die Hypothesenbewertungen von Experten, deren Bewertungen vom Betrag her unterhalb eines Mindestmaßes (z.B. 0.5) sind, werden ebenfalls nicht berücksichtigt.
- Hypothesenbewertungen von Experten mit einem größeren Vertrauensgrad soll ein größeres Vertrauen eingeräumt werden als Bewertungen von Experten mit einem kleineren Vertrauensgrad.
- Hypothesenbewertungen mit der gleichen Tendenz (Ablehnung bzw. Akzeptanz) sollen die Bewertung verstärken, Widersprüche sollen zu einer schwächeren Bewertung führen.

Eine derart aggregierte Gesamtbewertung kann wie folgt erreicht werden: Für jedes Paar Experten-/Hypothesenbewertung wird durch Multiplikation der Konfidenzfaktoren ein gewichteter Bewertungswert bestimmt. Diese gewichteten Bewertungswerte gehen – falls die Expertenbewertung das definierte Mindestmaß überschreitet – in eine Gesamthypothesenbewertung ein, die entsprechend den Ausführungen im vorangegangenen Abschnitt auf der Aggregationsformel basiert und sequenziell berechnet werden kann (vgl. S. 130).

Unterschreitet die Gesamthypothesenbewertung g_t eine untere Schranke (z.B. 0.35), wird die Hypothese abgelehnt. Wird eine obere Schranke (z.B. 0.65) überschritten, wird die Hypothese angenommen. Ansonsten kann keine Entscheidung über die Ablehnung bzw. Akzeptanz dieser Hypothese gemacht werden. Sie verbleibt dann als offene Hypothese auf dem Blackboard.

Beispiel 7.5 *Von Experten mit den Expertenbewertungen* $\{0.9, 0.95, 0.65, 0.4\}$ *wurden die zugehörigen Hypothesenbewertungen* $\{0.0, 0.8, -0.7, 0.2\}$ *erzeugt. Das erste und das vierte Paar werden nicht weiter berücksichtigt, da die Hypothesenbewertung neutral ist bzw. die Expertenbewertung unter dem Mindestmaß von 0.5. Für das zweite und das dritte Paar ergeben sich die gewichteten Hypothesenbewertungen 0.76 und -0.455. Diese Werte gehen in die Aggregationsformel ein und liefern als Gesamtbewertung 0.56. Wenn eine obere Schranke von 0.5 vorgegeben wurde, ist diese Hypothese somit zu akzeptieren.*

7.5 Resümee

Die Kombination einer Blackboard-Architektur mit Ansätzen aus dem Workflow-Managament-System-Bereich sowie einem auf Konfidenzfaktoren basierenden Bewertungsmodell bei strikter Trennung der Problemlösungsstrategie von den Problemlösungsverfahren führt zu einer mächtigen Wissensakquisitionsarchitektur, die die für die Thesaurusintegration erforderliche Flexibilität und Skalierbarkeit besitzt.

Insbesondere zeichnet sich unsere Architektur FA^{2}ITH durch die einfache Erweiterbarkeit und Anpassbarkeit bezüglich veränderter und weiterentwickelter Problemlösungsstrategien sowie der einfachen Integration von Integrationsverfahren, die mit unterschiedlichen Methoden, Techniken und Werkzeugen entwickelt werden können, aus. Das Einbringen des menschlichen Expertenwissens in den Integrationsprozess wird – im Gegensatz z.B. zu der Blackboard-Architektur von Viegener [Vie97] – auf vielfältige Art und Weise unterstützt. Der menschliche Experte kann sein Wissen sowohl in die Problemlösungsstrategie einbringen als auch in das eigentliche Problemlösungsverfahren.

Den potenziellen Engpässen bei Blackboard-Architekturen wurde bereits beim Entwurf von FA^{2}ITH durch die Aufteilung des Blackboards in verschiedene Bereiche sowie die Einführung des Moderators und der Möglichkeit der Agentenbewertungen entgegengewirkt.

Mit der in diesem Kapitel vorgestellten Architektur, die der zentrale Bestandteil unseres Rahmenwerkes für Thesaurusföderation ist, haben wir somit eine vielversprechende Ausgangslage für den Thesaurusintegrationsprozess.

Kapitel 8

Vorbereitungsphase

Trotz existierender Standards und Normen für Thesauri unterscheiden sich die Informationsmodelle von Thesauri erheblich. Für die eigentliche Integrationsphase aber sind vergleichbare Informationsmodelle – und damit eine vergleichbare Semantik der Begriffsgraphen – sowie eine vergleichbare Syntax der Benennungen erforderlich. In diesem Kapitel werden Verfahren zur Identifikation von Informationsmodellabweichungen und deren Behandlung vorgestellt (zur Einordnung vgl. Abbildung 8.1).

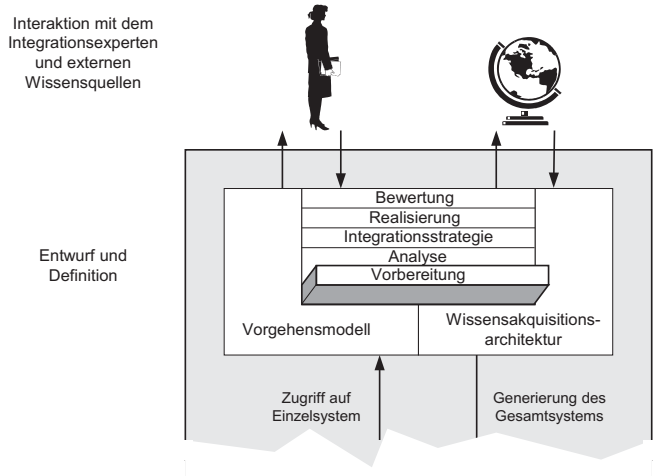

Abbildung 8.1: In der Vorbereitungsphase als erster Phase der Begriffsintegration wird durch Angleichen der Informationsmodelle die Voraussetzung für die Integrationsverfahren der folgenden Phasen geschaffen

Der Schwerpunkt dieses Kapitels liegt auf der Herstellung der Konformität der Informationsmodelle (Abschnitt 8.1), zusätzlich werden Verfahren zum Herstellen normierter Bezeichner und Definitionen entwickelt (Abschnitte 8.2 und 8.3).

8.1 Herstellung der Konformität der Informationsmodelle

Um die Konformität der Informationsmodelle der Komponententhesauri mit dem von uns definierten Informationsmodell (vgl. Kapitel 5 und hier insbesondere Abschnitt 5.2.3) zu gewährleisten, müssen Abweichungen identifiziert und Maßnahmen zur Herstellung der Konformität

ergriffen werden. In allen folgenden Phasen kann anschließend von einem einheitlichen Mindeststandard der Informationsmodelle ausgegangen werden. Es sei angemerkt, dass zusätzliche Informationen, die im Rahmen der Vorbereitungsphase zu einem Komponententhesaurus erzeugt werden, aufgrund der Autonomie der Komponententhesauri separat von diesen gehalten werden müssen.

In Tabelle 8.1 sind die möglichen Typen von Informationsmodellabweichungen und deren Behandlung dargestellt.

Abweichung	Behandlung
Begriffe und Benennungen	
Keine Ausweisung von Vorzugsbenennungen	Auswahl einer Vorzugsbenennung, entsprechende Konfiguration der Planungsverfahren
Keine Trennung Benennung - Annotation	Trennung herbeiführen
Identische BK-Verweismengen von verschiedenen Nicht-Deskriptorknoten	reduzieren
Beziehungen zwischen Nicht-Deskriptoren	ignorieren
Unverbundene Nicht-Deskriptoren	ignorieren
Keine Disjunktheit der BS- und BK-Verweise	BK-Kanten ignorieren
Semantische Relationen	
Selbstverweise	ignorieren
Unterschiedliche Kanten in eine Richtung zwischen identischen Knoten	reduzieren
Keine inverse Kante	inverse Kanten erzeugen
Keine Symmetrie der Assoziationskanten	Symmetrie herstellen
Keine Differenzierung der Hierarchierelation in Abstraktions- und Bestandsrelation	Differenzierung herstellen
Differenzierung der Hierarchierelation über Abstraktions- und Bestandsrelation hinaus	reduzieren
Abstraktionspfade mit Zyklen	reduzieren
Bestandspfade mit Zyklen	reduzieren
Hierarchiepfade mit Zyklen	reduzieren
Redundante Abstraktionspfade	reduzieren
Gruppen	
Keine Gruppen vorhanden	Gruppenzuordnung durchführen
Gruppenzuordnung von Nicht-Toptermen	reduzieren
Erläuterung: ignorieren = Informationen werden ignoriert reduzieren = Informationsmodellkompatible Reduktion der Informationen	

Tabelle 8.1: Typen von Informationsmodellabweichungen bei Komponententhesauri und deren Behandlung

In den folgenden Abschnitten 8.1.1 bis 8.1.3 werden Verfahren zur Identifikation und Behandlung solcher Informationsmodellabweichungen detaillierter beschrieben. Es würde den Rahmen dieser Arbeit sprengen, für alle aufgeführten Abweichungen Transfer-, Reduktions- bzw. Anreicherungsverfahren ausführlich vorzustellen. Wir beschränken uns daher auf solche Fälle, die in der Praxis häufig vorgefunden werden. Es sei darauf hingewiesen, dass alle Informationen, die zusätzlich gewonnen werden, aufgrund der Autonomie der Komponententhesauri in der Föderation gespeichert werden.

8.1.1 Begriffe und Benennungen

8.1.1.1 Keine Ausweisung von Vorzugsbenennungen

8.1.1.1.1 Identifikation der Informationsmodellabweichung Es gibt Thesauri, die nicht zwischen Deskriptoren und Nicht-Deskriptoren unterscheiden, sondern alle Benennungen gleichwertig behandeln, also beispielsweise für die Indexierung zulassen. Diese Thesauri werden als Thesauri ohne Vorzugsbenennungen bezeichnet. Wenn WordNet [Mil98] als Thesaurus betrachtet wird, ist durch die Zusammenfassung gleichbedeutender Benennungen zu so genannten Synsets (Synonymmengen) ohne Kennzeichnung einer ausgezeichneten Benennung ein solcher Thesaurus ohne Vorzugsbenennungen gegeben.

Zu erkennen ist ein Thesaurus ohne Vorzugsbenennungen daran, dass alle Benennungen entweder als Deskriptoren oder als Nicht-Deskriptoren betrachtet werden und zwischen diesen Benennungen Beziehungen existieren, die in unserem Thesaurusmodell nicht vorgesehen sind. Formal bedeutet dies:

$$\begin{aligned}((\forall n \in N : \lambda_n^{typ} \neq Deskriptor) \quad &\wedge \quad (\exists x, y \in B - D : (x, \alpha, y) \in E)) \vee \\ ((B - D = \emptyset) \quad &\wedge \quad (\exists x, y \in D : (x, \alpha, y) \in E))\end{aligned}$$

wobei $\alpha \notin \{BS, BF, BK, KB, OA, UA, OP, UP, VB\}$

Ein Algorithmus zum Erkennen eines Thesaurus ohne Vorzugsbenennungen ist nicht erforderlich, da dieses Faktum von einem Integrationsexperten einfach erkannt werden kann.

8.1.1.1.2 Ausweisung von Vorzugsbenennungen Ist eine Entscheidung für die Integration eines Thesaurus ohne Vorzugsbenennungen in die Thesaurusföderation gefallen – wir betrachten einen solchen Fall aufgrund der wenigen Thesauri ohne Vorzugsbenennungen als Ausnahme – behandeln wir diesen Thesaurus aus Gründen der einfacheren Handhabung wie einen Thesaurus mit Vorzugsbenennungen. Daher wählen wir aus der Menge der gleichwertigen Benennungen, die einen Begriff repräsentieren, eine Benennung als Vorzugsbenennung aus. Falls es Hinweise auf eine ausgezeichnete Benennung gibt (z.B. eine nicht-alphabetische Reihenfolge), können diese Hinweise verwendet werden. Ansonsten kann ein Vergleich mit den Benennungen anderer Komponententhesauri, die Vorzugsbenennungen besitzen, Kriterien für eine Auswahl liefern. Schließlich kann eine zufällige Auswahl stattfinden.

Wir vermerken in der Föderation, dass es sich ursprünglich um einen Thesaurus ohne Vorzugsbenennungen handelt. Diese Information wird bei der Festlegung der Intergarionsstrategie berücksichtigt.

8.1.1.2 Keine Trennung Benennung – Annotationen

8.1.1.2.1 Identifikation der Informationsmodellabweichung Unser Modell sieht eine strikte Trennung von Benennungen und Annotationen, z.B. Polysem-/Homonymauflösungen, vor (vgl. Definition 5.1). Unterstützt wird dies von der für Datenbanken üblichen Forderung nach erster Normalform [Dat95]. Diese verlangt, dass die Domäne aller Attribute atomar sein soll.

In einer Vielzahl von Thesauri kann hingegen beobachtet werden, dass eine Trennung der eigentlichen Benennungen von Annotationen unterschiedlichen Typs (z.B. Polysem-/ Homonyma
eine Anmerkung, ob der Deskriptor zur Indexierung erlaubt ist oder nicht, Beschreibung der

Quelle der Benennung) zwar für den menschlichen Benutzer ersichtlich wird, sie explizit aber nicht stattfindet (Beispiele aus GEMET: *fallout (chemicals)*, *PEL (permissible exposure limit)* und *public (n.)*). Stattdessen wird der Benennung die Annotation angehängt, so dass *eine* Zeichenkette entsteht. Dabei werden bestimmte Sonderzeichen zu einer rein optischen Trennung von Benennung und Annotation verwandt, die es aber nicht ersichtlich werden lassen, um welchen Typ von Annotation es sich handelt. Für eine maschinelle Weiterverarbeitung ist ein Transfer in unser Modell erforderlich.

Der Integrationsexperte erkennt die fehlende Trennung von Benennung und Annotation bei einer Durchsicht der Benennungen und kann das oder die verwendeten Sonderzeichen zur Kennzeichnung der Annotation bestimmen. Da es sich häufig um Klammern zur Kennzeichnung der Annotationen handelt, kann bereits folgende Feststellung als Indikator für eine fehlende Trennung von Benennung und Annotation angesehen werden:

$$\exists n \in B \subset N : „(“ \in \lambda_n^s$$

8.1.1.2.2 Separieren von Annotationen und Benennungen Es wird angenommen, dass die Klammer als Sonderzeichen die Annotation von der Benennung trennt, andere Zeichen oder Zeichenfolgen können vom Integrator angegeben werden. Somit kann die Annotation von der Benennung separiert werden.

Um den Typ der Annotation zu erkennen, können Muster vorgegeben werden, die jeweils für einen bestimmten Annotationstypen gültig sind. Diese Muster können initial vom System vorgegeben sein oder vom Integrator – auch zur Laufzeit – angegeben werden. Je nach Annotationstyp wird die Annotation behandelt. Das Tupel (Muster, Aktion) bildet somit eine einfache Regel. Die Regelreihenfolge bestimmt deren Auswertungsreihenfolge, so dass spezifischere Regeln vor unspezifischeren Regeln aufgeführt werden. Tritt der Fall auf, dass eine Benennung/Annotation keinem Muster zugeordnet werden kann, wird der Integrator gebeten, eine neue Regel anzugeben.

Beispiel 8.1 *Für die Separierung der Annotation von den Benennungen wurde für GEMET „(*)“ als Separierungsmuster erkannt.* $(\lambda_n^s)^{benennung}$ *und* $(\lambda_n^s)^{annotation}$ *bezeichnen dann die Benennung bzw. die Annotation. Tabelle 8.2 zeigt die Annotationsmuster, die hier der Verständlichkeit wegen in natürlicher Sprache beschrieben werden, mit Beispielen, zugehörigen Annotationstypen und deren Behandlung. Die unterschiedlichen Behandlungen belegen die Notwendigkeit der Unterscheidung der Annotationstypen.*

8.1.1.3 Identische BK-Verweismengen von verschiedenen Nicht-Deskriptorknoten

8.1.1.3.1 Identifikation der Informationsmodellabweichung Existieren von verschiedenen Nicht-Deskriptorknoten identische BK-Verweismengen, bedeutet dies, dass die Nicht-Deskriptoren als synonym zu betrachten wären. Solch eine Beziehung widerspricht aber dem Sinne von Kombinations-Nicht-Deskriptoren, die zugelassen sind, um die Menge der Deskriptoren überschaubar zu halten. Daher ist eine Reduktion auf unser Informationsmodell erforderlich.

Identische BK-Verweismengen und die entsprechenden Nicht-Deskriptoren können durch einen paarweisen Vergleich aller Mengen von BK-Verweisen identifiziert werden:

$$\begin{aligned} \exists m, n \in N : \quad & \#m \neq \#n \wedge \\ & \lambda_m^{typ} = \text{Kombinations-Nicht-Deskriptor} = \lambda_n^{typ} \wedge \\ & \{c : (m, BK, c) \in E\} = \{d : (n, BK, d) \in E\} \end{aligned}$$

$(\lambda_n^s)^{benennung}$	$(\lambda_n^s)^{annotation}$	Beispiel	Annotationstyp	Behandlung
Folge von Großbuchstaben	Für jeden Großbuchstaben aus $(\lambda_n^s)^{benennung}$ existiert in gleicher Reihenfolge ein mit diesem Buchstaben beginnendes Wort	PEL (permissible exposure limit)	Akronym-/ Grundwörter	$(\lambda_n^s)^{annotation}$ als Synonym aufnehmen, „acronym for $(\lambda_n^s)^{annotation}$“ als Homonym-/Polysem-Auflösung aufnehmen
-	Zeichenfolge “n.”	public (n.)	Wortartkennzeichnung	“noun” als Homonym-/Polysem-Auflösung aufnehmen
-	1 oder 2 Großbuchstaben	cantonal law (CH)	Kennzeichnung des Herkunftslandes	$(\lambda_n^s)^{annotation}$ als Erläuterung aufnehmen
-	$\geq$ 3 Großbuchstaben	vulnerable species (IUCN)	Verweis auf detaillierte Definition	$(\lambda_n^s)^{annotation}$ als Erläuterung aufnehmen
Folge von Kleinbuchstaben, -, /, Leerzeichen	Folge von Kleinbuchstaben, -, /, Leerzeichen	plant (industry)	Homonym-/ Polysem-Auflösung	$(\lambda_n^s)^{annotation}$ als Homonym-/Polysem-Auflösung aufnehmen

Tabelle 8.2: Annotationsmuster – geordnet nach abfallender Priorität – und deren Behandlung bei GEMET

8.1.1.3.2 Reduktion auf eindeutige BK-Verweismengen Eine Reduktion auf unser Informationsmodell bedeutet die Auswahl höchstens einer der identischen Verweismengen. Eine solche Auswahl kann nicht automatisiert werden, sondern wird dem Integrationsexperten überlassen. Die zu den nicht-ausgewählten Verweismengen gehörenden Nicht-Deskriptoren werden ignoriert.

8.1.1.4 Weitere Informationsmodellabweichungen

Die vertiefte Behandlung weiterer Informationsmodellabweichungen würde den Rahmen dieser Arbeit sprengen. Zumal eine Identifikation oft trivial ist und auch die Behandlung zu wenig neuen Erkenntnissen führt. Mindestens überprüft werden sollte zusätzlich, ob Beziehungen zwischen Nicht-Deskriptoren existieren (wenn dies ausgeschlossen wird, werden für Nicht-Deskriptoren auch Selbstverweise ausgeschlossen), ob unverbundene Nicht-Deskriptoren existieren und dass die Disjunktheit der BS- und BK-Verweise garaniert wird.

8.1.2 Semantische Relationen

Auch bei der vertieften Betrachtung von Informationsmodellabweichungen, die die semantischen Relationen (Abstraktions-, Bestands-, Hierarchie- und Verwandtschaftsrelation) betreffen, beschränken wir uns auf exemplarische Fälle, die wir nach Dringlichkeit einer Behandlung und Vorhandensein einer nicht unerheblichen Komplexität auswählen.

8.1.2.1 Keine Differenzierung der Hierarchierelation

8.1.2.1.1 Identifikation der Informationsmodellabweichung Aufgrund der grundsätzlich unterschiedlichen Semantik der Abstraktions- und Bestandsrelation und auch deren unterschiedlichen Eigenschaften ist für das weitere Vorgehen der Integration eine Differenzierung der Hierarchierelation in Abstraktions- und Bestandsrelation erforderlich. Erst durch einen dadurch geschaffenen *differenzierten relationalen Kontext* können weitere Integrationsverfahren die Bedeutung und Verwendung des Begriffs im Zusammenhang mit weiteren Thesaurusbegriffen ausreichend berücksichtigen.

Sieht das Informationsmodell des Komponententhesaurus nur eine Art von hierarchischer Relation vor, ist zu überprüfen, ob es sich ausschließlich um Abstraktions- bzw. Bestandsbeziehungen handelt. Eine solche Prüfung kann stichprobenartig manuell geschehen oder aber es werden die im folgenden Abschnitt beschriebenen Mechanismen zur Klassifizierung hierarchischer Relationen angewandt.

8.1.2.1.2 Klassifizierung hierarchischer Relationen Ergibt die Überprüfung, dass im Komponententhesaurus sowohl Abstraktions- als auch Bestandsbeziehungen vorkommen, diese aber nicht differenziert werden, ist jede hierarchische Beziehung als Abstraktions- oder als Bestandsbeziehung zu klassifizieren.

Grundlage für eine solche in der Literatur bisher nach unserem Wissen vollkommen vernachlässigte maschinelle Klassifizierung sind die Benennungen der Ober-/Unterbegriffspaare. Aufgrund der erforderlichen Analyse dieser Benennungen stellt die Linguistik das bedeutenste Hilfsmittel für eine maschinelle Klassifizierung dar. Zusätzlich können externe Wissensquellen einbezogen werden, die entweder bereits hierarchische Beziehungen in Abstraktions- und Bestandsbeziehungen unterscheiden oder aber andere Klassifizierungen für Benennungen anbieten. Wir setzen voraus, dass eine Benennungsnormierung (vgl. Abschnitt 8.2) bereits stattgefunden hat.

Diese Teilphase der Vorbereitungsphase wird auf Grundlage der Blackboard-Architektur durchgeführt. Da wir davon ausgehen, dass eine Hierarchiebeziehung, die keine Abstraktionsbeziehung ist, eine Bestandsbeziehung ist, stellen wir für alle Ober-/Unterbegriffspaare die Hypothese auf, dass diese Begriffe in einer Abstraktionsbeziehung stehen. Die Ablehnung einer solchen Hypothese bedeutet, dass es sich um eine Bestandsbeziehung handelt.

Die regelbasierte semi-automatische Hypothesenbewertung findet in folgenden Schritten statt:

Bereitstellen von Testdaten: Sowohl zu einer Gewinnung der Klassifizierungsregeln als auch zu einer Bewertung dieser werden bereits klassifizierte Beispiele (Testdaten) benötigt. Diese Beispiele werden durch eine zufällige Auswahl von Hierarchiebeziehungen (zufällige Begriffe jeweils mit all ihren Unterbegriffen, um die gesamte Aufteilung des Begriffes zu untersuchen) und die von einem Benutzeragenten unterstützte manuelle Klassifizierung durch den Integrationsexperten gewonnen.

Aufstellen von Hypothesenbewertungsregeln: Die Regeln stellen unser Problemlösungswissen dar. Daher kann dieser Schritt auch als Wissensbasisinitialisierung betrachtet werden. Die Regeln können anhand der manuell klassifizierten Beispiele gelernt oder aber manuell aufgestellt werden. Da die Regelmenge verhältnismäßig klein ist, haben wir auf die Implementierung von Lernverfahren verzichtet und die Regeln manuell aufgestellt. Diese in Tabelle 8.3 dargestellten Regeln bieten für englischsprachige Benennungen bereits eine gute Ausgangsbasis und können leicht auf beliebige Thesauri übertragen werden. Es ist

zu beachten, dass der Übersichtlichkeit halber nur die Regeln für die jeweiligen Deskriptoren aufgeführt sind. Die Regeln werden ebenso für Deskriptor-/Nicht-Deskriptor- wie für Nicht-Deskriptor-/Nicht-Deskriptorvergleiche aufgestellt.

Falls Hypothesen von keiner der unabhängigen Regeln bewertet werden können, können Regeln ergänzt werden, die aufgrund des gesamten Bewertungsergebnisses der anderen Regeln Bewertungen liefern. In diesem Fall sind das die Regeln 7 und 8, die davon ausgehen, dass nicht-klassifizierte Schwesterknoten in einer gleichartigen Beziehung mit dem Vaterknoten stehen wie bereits klassifizierte Schwesterknoten, wenn diese alle einheitlich klassifiziert wurden.

Wurde eine Beziehungen anhand keiner der Regeln 1 bis 8 klassifiziert, kann entweder eine manuelle Klassifizierung erfolgen oder aber es greifen Standardregeln: Eine nicht-klassifizierte Beziehung wird als Bestandsbeziehung klassifiziert, wenn die bisher klassifizierten Beziehungen überwiegend Bestandsbeziehungen sind, ansonsten als Abstraktionsbeziehung. Somit wird schließlich ohne menschlichen Eingriff eine Entscheidung getroffen.

Unterschiedliche Konfidenzfaktoren können für die einzelnen Regeln, aber auch innerhalb der einzelnen Regeln (z.B. abhängig davon, ob die Bedingung bei Betrachtung der Deskriptoren oder der Nicht-Deskriptoren erfüllt ist) festgelegt werden. Diese Konfidenzfaktoren können im Rahmen der Regelbewertung modifiziert werden.

Nr.	Bedingung	Beispiel	Bewertung
WordNet-basierte Regeln zum Allgemeinen Auffinden von Hierarchiebeziehungen			
1	$\lambda^s_{d_1}$ ist (direkter oder indirekter) Abstraktionsoberbegriff von $\lambda^s_{d_2}$ in WordNet	additive – anticaking agent	Hypothesen-bestätigung
2	$\lambda^s_{d_1}$ ist (direkter oder indirekter) Bestands-oberbegriff von $\lambda^s_{d_2}$ in WordNet	flower – stamen	Hypothesen-ablehnung
WordNet-basierte Regeln zur Analyse von Mehrwortbenennungen			
3	$\lambda^s_{d_1}$ oder $\lambda^s_{d_2}$ ist Mehrwortbenennung und letztes Wort von $\lambda^s_{d_1}$ ist (direkter oder indirekter) Abstraktionsoberbegriff von letztem Wort von $\lambda^s_{d_2}$ in WordNet	chemical processes - softening	Hypothesen-bestätigung
4	$\lambda^s_{d_1}$ oder $\lambda^s_{d_2}$ ist Mehrwortbenennung und letztes Wort von $\lambda^s_{d_1}$ ist (direkter oder indirekter) Bestandsoberbegriff von letztem Wort von $\lambda^s_{d_2}$ in WordNet	steam boiler - whistle	Hypothesen-ablehnung
Regeln zur linguistischen Analyse von Mehrwortbenennungen			
5	letztes Wort von $\lambda^s_{d_1}$ ist identisch mit letztem Wort von $\lambda^s_{d_2}$ (d_2 ist Spezialisierung von d_1)	bird - breeding bird	Hypothesen-bestätigung
6	$\lambda^s_{d_2}$ besteht aus mindestens zwei Wörtern und erstes Wort von $\lambda^s_{d_2}$ ist Wortform von letztem Wort von $\lambda^s_{d_1}$ (d_2 ist Aufteilung von d_1)	accidents - accident source	Hypothesen-ablehnung

Fortsetzung auf der nächsten Seite ...

... Fortsetzung

Nr.	Bedingung	Beispiel	Bewertung
Zwischenergebnisbasierte Regeln			
7	die Beziehung zwischen d_1 und d_2 wurde bisher nicht klassifiziert und alle klassifizierten Unterbegriffe von d_1 wurden als Abstraktionsunterbegriffe klassifiziert	arthropods - crustaceans; chelicerates	Hypothesenannahme
8	die Beziehung zwischen d_1 und d_2 wurde bisher nicht klassifiziert und alle klassifizierten Unterbegriffe von d_1 wurden als Bestandsunterbegriffe klassifiziert	government - executive; officialdom	Hypothesenablehnung
Standard-Regeln			
9	Mehrheit der klassifizierten Hierarchiebeziehungen sind Abstraktionsbeziehungen		Hypothesenannahme
10	Mehrheit der klassifizierten Hierarchiebeziehungen sind Bestandsbeziehungen		Hypothesenablehnung
11	Anzahl der klassifizierten Bestandsbeziehungen ist identisch mit Anzahl der Abstraktionsbeziehungen		Hypothesenannahme

Tabelle 8.3: Regeln zur Bewertung der Hypothese $(d_1, UA, d_2) \in E$ falls d_1 Oberbegriff einer nichtklassifizierten Hierarchiebeziehung zu d_2 ist

Zuordnen von Regeln zu Experten: Theoretisch kann jede Regel einem eigenen Experten zugeordnet werden. Dies verursacht aber einen hohen Kommunikationsaufwand. Daher werden ähnliche Regelmengen gruppiert und jeweils als Gruppe einem Experten zugewiesen. Kriterium für eine solche Gruppierung ist beispielsweise, welche linguistischen Verfahren angewandt werden und auf welche externen Wissensquellen zugegriffen wird.

Bewertung der Regeln: Eine Bewertung der Regeln findet anhand der Ergebnisse der maschinellen Klassifikation angewandt auf die bereitgestellten Beispiele statt. Die Auswertung der korrekten/falschen bzw. nicht klassifizierten Beziehungen erlaubt Rückschlüsse auf die Performanz der verschiedenen Regeln. Anhand dieser Rückschlüsse werden Wissensbasismodifikationen und -erweiterungen sowie Agentenbewertungen durchgeführt. Nicht-performante Regeln (Anteil der falschen Klassifizierungen überschreitet einen Grenzwert) können ganz entfernt werden oder aber die zugehörigen Agenten werden entprechend niedriger bewertet. Für nicht-klassifizierte Hierarchiebeziehungen kann versucht werden, neue Regeln aufzustellen. Nach einer Modifikation der Wissensbasis können solange erneut Bewertungen der Regeln und weitere Modifikationen stattfinden, bis das Ergebnis der maschinellen Klassifikation für die Testdatenmenge zufriedenstellend ist.

Ausführung der Regeln: Aufgrund der Testdaten-basierten Lern- und Konfigurationsphase kann das Regelsystem nun gegen den Gesamtdatenbestand laufen. Die eigentliche Ausführung übernimmt die Blackboard-basierte Wissensakquisitionsarchitektur.

Plausibilitätsprüfungen: Nach einer Klassifikation des Gesamtdatenbestandes, also aller Hierarchiebeziehungen des Komponententhesaurus, können Plausibilitätsprüfungen durchgeführt werden. Dies kann entweder auf der Basis von Stichproben geschehen oder auf der Überprüfung häufiger Muster. Etwa können bei Auftreten von Polydimensionalität, d.h. ein Begriff enthält sowohl Abstraktions- als auch Bestandsunterbegriffe, die entsprechenden Beziehungen dem Experten zur Beurteilung vorgelegt werden.

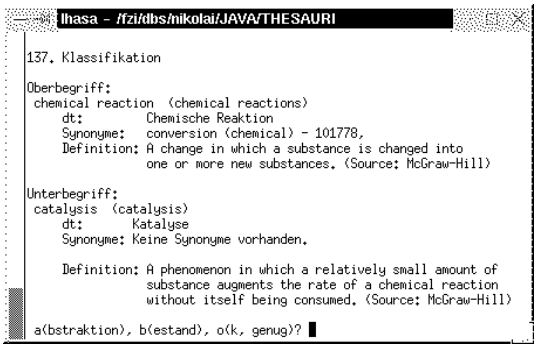

Abbildung 8.2: Benutzeragent zur manuellen Klassifizierung von Hierarchiebeziehungen

Beispiel 8.2 *Die Thesauri GEMET und AGROVOC unterscheiden die Hierarchierelation nicht in Abstraktions- und Bestandsrelation. Mithilfe des in Abbildung 8.2 dargestellten Benutzeragenten wurden jeweils ca. 150 Hierarchiebeziehungen manuell klassifiziert, um Testdaten zu erhalten.*

Für die in Tabelle 8.3 dargestellten Regeln wurden entsprechend der Gruppierung in WordNet-basierte Regeln zum Allgemeinen Auffinden von Hierarchiebeziehungen, WordNet-basierte Regeln für Mehrwortbenennungen, Regeln zur linguistischen Analyse von Mehrwortbenennungen, zwischenergebnisbasierten und Standardregeln fünf Agenten implementiert, von denen die ersten drei Agenten jeweils die Deskriptoren und die Nicht-Deskriptoren betrachten. Diese Agenten erzeugen bzw. bewerten die Hypothese, dass eine Hierarchiebeziehung eine Abstraktionsbeziehung ist. Eine Ablehnung bedeutet, dass es sich um eine Bestandsbeziehung handelt.

Angewandt auf die Testdaten erzielen die Agenten die in Tabelle 8.4 dargestellten Ergebnisse. Die ersten drei Agenten bewerten, wenn sie Bewertungen abgegeben, über 95% der Hypothesen korrekt.[1]

Eine detailliertere Auswertung der einzelnen Regeln ergibt, dass Regel 6 in allen Fällen zu einer korrekten Bewertung führt und Agenten-intern entsprechend hoch bewertet werden kann. Allerdings trifft diese Regel nur in ca. 1% aller Fälle zu. Ob die Bewertung anhand der Deskriptoren oder der Nicht-Deskriptoren getroffen wird, spielt bei beiden Thesauri nur eine untergeordnete Rolle. Mit Abstand am meisten Hypothesen werden anhand der Regeln 5 und 6 bewertet. Diese zeigt die zentrale Bedeutung, die die Mehrwortanalyse bei der Hierarchieklassifikation hat.

Die ca. 45% der Hypothesen, die nicht anhand der ersten drei Agenten bewertet werden, können anhand der Schwesternanalyse und der Mehrheitsverhältnisse überraschend häufig korrekt bewertet werden (93% bzw. 86%), wobei erwartungsgemäß die Schwesternanalyse etwas bessere Ergebnisse produziert.

Das Gesamtergebnis bereits mit diesem Standardsatz an Regeln ist mit über 90% korrekten Entscheidungen hervorragend. Der manuelle Aufwand des Integrationsexperten beschränkt sich auf das werkzeuggestützte Erstellen von Testdaten sowie einer Kontrolle der Ergebnisse, um die Gewichtung der einzelnen Regeln zu optimieren. Da bei der Erstellung dieser Testdaten bereits die Polydimensionalität beobachtet wurde, werden keine auf einer Vermeidung der Polydimensio-

[1] Eine Auswertung der nicht mit dem Integrationsexperten übereinstimmenden Bewertungen ergibt mehrere diskussionswürdige Entscheidungen (z.B. hat der Experte Anseriformes (eine biologische Ordnung) als Abstraktionsunterbegriff von aves (eine biologische Klasse) klassifiziert, die Agenten hingegen als Bestandsunterbegriff) sowie einige Spezialfälle (z.B. geographische Regionen wie *Africa south of Sahara*, die andere Regionen wie *Sudano Sahelian Region* als Bestandsunterbegriffe enthalten), die im Rahmen dieser Arbeit nicht betrachtet werden, für die bei einem größerem Integrationsprojekt bei mehreren Thesauri mit ähnlichen Fällen aber entsprechende Regeln gefunden werden können.

nalität basierenden Plausibilitätsprüfungen durchgeführt. Weitere manuelle Eingriffe erscheinen aufgrund der hervorragenden Ergebnisse nicht notwendig. Allerdings können im Laufe des weiteren Integrationsverfahrens festgestellte Fehl-Klassifikationen korrigiert werden, um die Ergebnisse der Integration weiter zu verbessern.

Es sei bereits an dieser Stelle ein erheblicher Unterschied der beiden untersuchten Thesauri angemerkt: Angewandt auf alle Hierarchiebeziehungen werden in GEMET nur sehr wenige Bestandsbeziehungen erkannt (< 1%), während in AGROVOC die Anzahl der als Bestandsbeziehungen klassifizierten Hierarchiebeziehungen die Anzahl der als Abstraktionsbeziehungen klassifizierten Hierarchiebeziehungen um etwa den Faktor zwei übertrifft. Der überwiegende Teil dieser Bestandsbeziehungen basiert auf umfangreichen biologischen Begriffstaxonomien. Z.B. ist die Familie Cupressaceae Bestandsoberbegriff der Gattung Libocedrus, die wiederum Bestandsoberbegriff der Art Libocedrus Decurrens ist. Da Arten immer einen Doppelnamen besitzen, deren erster die Gattung bezeichnet, trifft insbesondere Regel 6 sehr häufig zu. Eine derart umfangreiche fachliche Begriffstaxonomie ist in GEMET nicht enthalten.

Agent	korrekt (%)		falsch (%)		unbewertet (%)	
	AV	GT	AV	GT	AV	GT
WordNet (Regeln 1-2)	32.3 (100.0)	17.7 (96.7)	0.0 (0.0)	0.6 (3.3)	67.7	81.7
WordNet (Regeln 3-4)	56.3 (95.9)	45.7 (97.4)	2.4 (4.1)	1.2 (2.6)	41.3	53.0
Mehrwortanalyse (Regeln 5-6)	13.8 (85.2)	14.6 (100.0)	2.4 (14.8)	0.0 (0.0)	83.8	85.4
Zwischenergebnis	55.7 (95.9)	51.2 (98.8)	2.4 (4.1)	0.6 (1.2)	41.9	47.6
Schwesternanalyse (Regeln 7-8)	68.6 (94.1)	69.2 (93.1)	4.3 (5.9)	5.1 (6.9)	27.1	25.6
Mehrheit (Regeln 9-11)	92.9	85.9	7.1	14.1	0.0	0.0
Zwischenergebnis	92.9	87.2	7.1	12.8	0.0	0.0
Gesamt	94.6	93.3	5.4	6.7	0.0	0.0

Erläuterung:
AV = AGROVOC
GT = GEMET
Die geklammerten Werte geben an, wie groß der Anteil der korrekten bzw. falschen Bewertungen an der Gesamtzahl nicht-neutraler Bewertungen ist.

Tabelle 8.4: Auswertung der Agenten-Bewertungen der Hypothese $(d_1, UA, d_2) \in E$ falls d_1 Oberbegriff einer nicht-klassifizierten Hierarchiebeziehung zu d_2 ist (Testdaten). Beispiele für korrekte Klassifikationen wurden bereits mit den einzelnen Regeln in Tabelle 8.3 dargestellt.

8.1.3 Gruppen

Da wir über die Zuordnung von Begriffen zu Gruppen semantische Informationen über die Begriffe erhalten, die den Experten bei der Integration wichtige Hinweise auf Integrationsstellen liefern können, ist es eine Aufgabe der Vorbereitungsphase, die Topterme der Komponententhesauri – und nur diese, da die Gruppenzuordnung entlang der Hierarchierelation vererbt wird – einer föderationsweit einheitlichen Menge an Gruppen zuzuordnen. Wir unterscheiden dabei drei verschiedene Ausgangssituationen, die wir in den folgenden Abschnitten kurz diskutieren.

8.1.3.1 Gruppenzuordnung ohne Gruppen in den Komponententhesauri

Falls keiner der Komponententhesauri Gruppen besitzt, wird durch eine nachträgliche Spezifikation von Gruppen und der Zuordnung der Topterme aller Thesauri ein erheblicher Aufwand erforderlich. Das semantische Wissen muss überwiegend manuell eingebracht werden, so dass kein Mehrwert gegenüber einem Einbringen während der weiteren Phasen des Integrationsprozesses entsteht. Daher widmen wir uns diesem Fall nicht weiter.

8.1.3.2 Gruppenzuordnung mit einer Menge an Gruppen

Besitzt genau ein Komponententhesaurus eine Menge an Gruppen, existiert im Vergleich zum vorangegangenen Fall bereits semantisches Zusatzwissen. Um die Vergleichbarkeit zwischen den Thesauri zu verbessern, bietet es sich an, die Topterme der anderen Thesauri ebenfalls diesen Gruppen zuzuordnen. Wir betrachten diesen Fall daher näher, beschränken uns aber der Einfachheit halber auf einen Thesaurus mit Gruppen und einen Thesaurus ohne Gruppen.

Betrachtet werden alle Paare von Toptermen aus den verschiedenen Thesauri. Wiederum sind Regeln die Grundlage des Verfahrens. Für Paare von Toptermen d_1, d_2 wird jeweils eine positive Hypothese aufgestellt, wenn die Bedingung einer Regel erfüllt ist, nach der d_2 aus dem Thesaurus ohne Gruppen zur Gruppe von d_1 zugeordnet werden kann. Auch Regeln für ablehnende Hypothesen können aufgestellt werden, sind jedoch unserer Erfahrung nach nicht notwendig, da die Regeln für zustimmende Hypothesen bereits qualitativ hochwertige Ergebnisse liefern. Diese positiven Regeln sind in Tabelle 8.5 zusammengefasst.

Nr.	Bedingung für Hypothesenannahme
Lexikalische Identität	
1	$\lambda^s_{d_1}$ und $\lambda^s_{d_2}$ sind lexikalisch identisch
2	d_1 besitzt lexikalisch mit $\lambda^s_{d_2}$ identische Nicht-Deskriptoren (oder umgekehrt) oder d_1 und d_2 besitzen lexikalisch identische Nicht-Deksriptoren
Lexikalische Identität von Unterbegriffen	
3	d_1 und d_2 besitzen lexikalisch identische Unterbegriffe (Deskriptoren oder Nicht-Deskriptoren)
Mehrwortanalysen	
4	$\lambda^s_{d_2}$ besteht aus mindestens zwei Wörtern und letztes Wort von $\lambda^s_{d_2}$ ist identisch mit letztem Wort von $\lambda^s_{d_1}$
5	$\lambda^s_{d_2}$ besteht aus mindestens zwei Wörtern und erstes Wort von $\lambda^s_{d_2}$ ist identisch mit $\lambda^s_{d_1}$ oder Wortform von $\lambda^s_{d_1}$
Lexikalische Identität von WordNet-Begriffen	
6	$\lambda^s_{d_1}$ und $\lambda^s_{d_1}$ treten gemeinsam in einem WordNet-Synset auf
7	$\lambda^s_{d_1}$ und $\lambda^s_{d_1}$ treten in WordNet-Synsets auf, die gemeinsame Oberbegriffe besitzen
8	$\lambda^s_{d_1}$ und $\lambda^s_{d_1}$ treten in WordNet-Synsets auf, die gemeinsame Unterbegriffe besitzen
Berücksichtigung der Assoziationsbeziehungen	
9	d_2 ist assoziiert mit einem Begriff, dessen Deskriptor oder Nicht-Deskriptor lexikalisch identisch mit $\lambda^s_{d_1}$ oder dessen Nicht-Deskriptoren ist
10	d_2 ist assoziiert mit einem Begriff, dessen Topterm bereits der Gruppe von d_1 zugeordnet wurde

Tabelle 8.5: Bedingungen für die Zuordnung von Toptermen zu Gruppen eines anderen Thesaurus

Die Gruppen des einzigen Komponententhesaurus mit Gruppen sollten zumindest in einem ersten Schritt eine Teilmenge der Gruppen in der Föderation sein. Weitere Gruppen sowie eine

eventuelle Aufteilung der Gruppen sollten durch den menschlichen Experten inklusive der Zuordnung der Topterme geschehen.

Beispiel 8.3 *Da AGROVOC keine Gruppen besitzt, wurden die Topterme von AGROVOC mittels der vorgestellten Regeln den Gruppen von GEMET zugeordnet. Den Regeln wurden dabei absteigende Konfidenzfaktoren zugewiesen. Von den insgesamt 1515 AGROVOC-Toptermen konnten so 1213 Topterme GEMET-Gruppen zugeordnet werden. Von den nicht zugeordneten Toptermen konnten 257 als lateinische Fachtermini identifiziert werden. Weitere 45 Topterme konnten nicht zugeordnet werden.*

Von den 1213 zugeordneten Toptermen wurden 1063 eindeutig einer Gruppe, 118 2 Gruppen, 26 3 Gruppen und 6 Topterme mehr als 3 Gruppen zugeordnet. Insbesondere die mehr als 3 Gruppen zugeordneten Topterme sollten manuell untersucht werden.

Die lexikalischen Vergleiche innerhalb der Thesauri ergaben mit Abstand die meisten Vorschläge für Gruppenzuordnungen (ca. 500). Die Mehrwortanalyse lieferte annähernd 300, WordNet-Analysen und Assoziationsanalysen jeweils deutlich über 200 Vorschläge.

100 Stichproben ergaben, dass über 85 % der Topterme korrekt Gruppen zugeordnet werden konnten.

8.1.3.3 Gruppenzuordnung mit verschiedenen Mengen an Gruppen

Besitzen die Komponententhesauri verschiedene Mengen von Gruppen, kann angenommen werden, dass nur ein Komponententhesaurus Gruppen besitzt und entsprechend des vorangegangenen Abschnitts können dann dieser Menge von Gruppen die Topterme zugewiesen werden. Wird dies jeweils für die Gruppen der verschiedenen Thesauri durchgeführt, kann die Feststellung der Gruppen mit einem großen Anteil gemeinsamer Topterme bereits wichtige Hinweise auf die zu verwendenden Gruppen in der Föderation liefern. Diese Föderationsgruppen werden vom menschlichen Experten definiert. Dieser kann ebenfalls ausdrücken, welche Gruppen der Komponententhesauri den Föderierten Gruppen entsprechen. Mit diesen Informationen sowie der im vorangegebenen Abschnitt beschriebenen Gruppenzuordnung kann eine Zuordnung der Topterme der Komponententhesauri zu den Föderierten Gruppen geschehen.

8.2 Herstellen normierter Benennungen

Neben der Herstellung konformer Informationsmodelle ist es eine weitere Aufgabe der Vorbereitungsphase sicherzustellen, dass in den weiteren Phasen vergleichbare Benennungen und Definitionen zur Verfügung stehen. Dies wird durch eine *Normierung*, also einer Überführung in eine zu spezifizierende Normalform, erreicht.

Für *Benennungen* wird eine solche Normalform für Einwortbenennungen und Mehrwortbenennungen benötigt. *Definitionen* für Begriffe werden in Form von Deskriptoren und Nicht-Deskriptoren, über die Hierarchie und verwandte Begriffe, über Zusätze zur Polysem- bzw. Homonymauflösung sowie über Definitions- und Erläuterungstexte angegeben. Die Darstellung der Deskriptoren und Nicht-Deskriptoren wird durch die Benennungsnormierung bereits vereinheitlicht. Die übereinstimmende Semantik der hierarchischen Beziehungen und der Assoziationsbeziehungen wird bei dem Transfer in ein einheitliches Informationsmodell zumindest grundsätzlich berücksichtigt. Die Zusätze zur Polysem- bzw. Homonymauflösung werden durch die Überführung in ein gemeinsames Informationsmodell sowie die Benennungsnormierung in

eine Normalform überführt. Die Analyse natürlichsprachiger Texte, die zur Angabe von Definitionen und Erläuterungen verwendet werden, ist aufgrund der Komplexität ein eigenes Forschungsthema. Sie wird im Rahmen dieser Arbeit nicht weiter behandelt, d.h., wir verzichten auf die Herstellung einer Normalform für Definitionen und Erläuterungen. Daraus resultierende Einschränkungen bei der Informationsverarbeitung werden in Kauf genommen.

Mit dem Separieren von Benennung und Annotation (vgl. Abschnitt 8.1.1.2.2) wurde bereits ein erster Schritt in Richtung normierter Benennungen unternommen. Die weiteren Schritte betreffen nun den Umgang mit der Benennung selbst und werden in den folgenden Abschnitten erläutert.

8.2.1 Allgemeine Benennungsnormierung

Zur Normierung der englischsprachigen Benennungen werden linguistische Verfahren in folgenden aufeinander aufbauenden Schritten angewandt:

1. Die Grundform aller Wörter der Benennung wird gebildet.

2. Wörter mit Bindestrichen werden zu einer Schreibweise ohne Bindestrich transformiert, wenn WordNet eine gleichbedeutende Benennung ohne Bindestrich enthält (Beispiel: aus re-afforestation wird reforestation).

3. Die in einer Liste enthaltenen zu löschenden Wörter (Artikel und Präpositionen: *a*, *the*, *as*) werden gelöscht, wenn das Wort nicht das letzte Wort der Benennung ist (Beispiel: vitamin a). Zusätzlich wird bei den Artikeln *a* und *the* die Reihenfolge der Wörter vor und hinter dem zu löschenden Wort getauscht, wenn das Wort davor auf die Silbe *ing* endet, also vermutlich ein Verb ist (Beispiel: aus *chewing the cud* wird *cud chewing*, aus *composition of the population* wird *composition of population*).

4. Die in einer Liste enthaltenen Aufzählungswörter (*and* und *or*) werden jeweils durch ein Komma ersetzt bzw. gelöscht, wenn dies zu zweifachem Kommata führen würde (Beispiel: *swan, goose, and duck* wird zu *swan, goose, duck*).

5. Von hinten beginnend werden die in einer Liste aufgeführten bei einer Wortreihenfolge zu vertauschenden Wörter gelöscht (*after*, *for*, *from*, *in*, *of*, *on*, *per*, *to* und *with*). Dabei werden die Wörter vor dem gelöschten Wort (ggf. bis zu dem nächsten Wort dieser Liste) mit den Wörtern nach dem Wort vertauscht (Beispiel: aus *report on state of environment* wird erst *report on environment state* und schließlich *environment state report*).

   Mit diesem Schritt wird eine „sinnerhaltende Normierung“ erreicht. Diese wird vermerkt. Für einige Integrationsverfahren jedoch ist eine weitergehende Normierung erforderlich, die in den folgenden Schritten beschrieben wird.

6. Bei den folgenden Wörtern wird wie im vorangegangenen Schritt beschrieben vorgegangen: *against*, *at*, *by*, *into*, *off*, *out*, *through* und *to*. (Beispiel: *offence against environment* wird zu *environment offence*.

7. Römische und griechische Zahlwörter werden gelöscht, wenn diese nicht am Anfang stehen (Beispiel aus *Agenda 21* wird *Agenda* aber *x ray* wird nicht modifiziert).

Die Grundformbildung erfolgt mit der entsprechenden WordNet-Funktionalität. Um Nomen, die im Plural eine andere Bedeutung haben als im Singular, im Plural zu belassen (Beispiele: AIDS

und aid, acoustics und acoustic), werden solche Wörter, die im Plural mit eigener Bedeutung in WordNet gefunden werden, nicht in die Singularform transformiert.

Eine weitere Analyse von Schreibvarianten kann zusätzlich durchgeführt werden. An dieser Stelle gehen wir aber davon aus, dass die Thesauri bei wichtigen Schreibvarianten diese bereits als Synonyme aufführen.

Selbstverständlich müssen auch die oben aufgeführten Listen auf Vollständigkeit (Werden weitere Präpositionen innerhalb der Mehrwortbenennungen verwendet?) und Gültigkeit überprüft werden. Wir haben sie als Beispiel für die Thesauri AGROVOC und GEMET aufgeführt.

Bereits mit dieser allgemeinen Benennungsnormierung ist ein erheblicher Schritt in Richtung der Vergleichbarkeit der Benennungen getan.

Beispiel 8.4 *Vor der Benennungsnormierung stimmen in AGROVOC und GEMET 1156 englischsprachige Benennungen lexikalisch überein. Nach der Benennungsnormierung steigt die Zahl der lexikalisch identischen Benennungen um* 65.7% *auf 1916 an.*

8.2.2 Normierung von Eigennamen

Einige Thesauri enthalten Eigennamen als Benennungen. Gerade bei Eigennamen aber tritt das Problem der Schreibvarianten besonders häufig auf. Im Rahmen der Benennungsnormierung ist es daher sinnvoll, diese Eigennamen ebenfalls auf eine normierte Form zurückzuführen. Als normierte Form bietet sich dabei die in DIN- oder ISO-Normen oder in standardisierten Nomenklatura-Listen aufgeführte Form oder Kurzform an. Die Erläuterung wird bei einer Transformation um einen entsprechenden Hinweis ergänzt.

Beispiel 8.5 *In GEMET werden als Nationenbezeichnungen (innerhalb der Annotationen) 2-stellige Buchstabenkürzel verwendet. AGROVOC hingegen verwendet die britisch-englische Nationenbezeichnung. Anhand der ISO-Norm 3166 können die Bezeichnungen auf das normierte 2-stellige Buchstabenkürzel abgebildet und so direkt lexikalisch miteinander verglichen werden.*

8.3 Herstellen normierter Definitionen

Insofern Begriffe über Deskriptoren und Nicht-Deskriptoren bzw. über die semantischen Relationen definiert werden, wurde eine Normierung der Definitionen bereits über die Benennungsnormierung und das Herstellen einheitlicher Informationsmodelle erreicht. Die natürlichsprachigen Definitions- und Erläuterungstexte zu normieren, haben wir bereits in Abschnitt 8.2 ausgeklammert. Zusätzlich betrachtet hingegen werden Annotationen in Form von Polysem-/ Homonymauflösungen, da diese häufig die Form von Benennungen aufweisen (Beispiel: *water consumption (plants)*). Im Sinne einer Normierung der Begriffsdefinitionen werden daher Annotationen wie Benennungen behandelt und auf eine Grundform zurückgeführt (vgl. Abschnitt 8.2).

8.4 Herstellen von Zugriffsschnittstellen

Grundvoraussetzung für die Begriffsintegration ist ein Zugriff auf die Komponententhesauri. Die Schaffung solcher Zugriffsmöglichkeiten wird aber nicht an dieser Stelle erläutert, da das zentrale

Anliegen der Begriffsintegration eine semantische Integration ist und die technische Integration als gegeben vorausgesetzt wird. Stattdessen sei auf Kapitel 13 verwiesen, in dem Lösungen für den einheitlichen Zugriff auf die Komponententhesauri im laufenden Betrieb vorgestellt werden. Diese Lösungen können leicht auf die erforderlichen Schnittstellen bei der Begriffsintegration übertragen und entsprechend erweitert werden. Um die Performanz der Begriffsintegration zu verbessern – es finden sehr viele Zugriffe auf die Thesaurusbegriffe statt – können die Thesauri zum ausschließlichen Zwecke der Integration über ein Thesaurusaustauschformat (vgl. z.B. [ND98]) auch in ein *gemeinsames* Datenhaltungssystem übertragen werden. Dieser Vorgang ist aber nicht Bestandteil unserer Betrachtungen.

8.5 Resümee

In diesem Kapitel haben wir Verfahren und Hilfsmittel zur Identifikation von Informationsmodellabweichungen und deren Behandlung auf semantischer und syntaktischer Ebene entwickelt. Wir haben gezeigt, dass basierend auf externen linguistischen Wissensquellen wie WordNet und einfachen linguistischen Verfahren die erforderlichen Informationsreduktionen und -anreicherungen auch bei schwierigen Problemen wie der nachträglichen Differenzierung der Hierarchierelation in Abstraktions- und Bestandsrelation weitgehend automatisiert werden kann. Damit haben wir sowohl die Flexibilität des Systems bei unterschiedlichen Thesauri als auch die Skalierbarkeit der entwickelten Verfahren bewiesen. Insbesondere erweist sich bereits in der Vorbereitungsphase die regel- und agentenbasierte Modularisierung der Problemlösung basierend auf der Wissensakquisitionsarchitektur FA^{2}ITH als leistungsfähig, konfigurierbar und erweiterbar.

Das gemeinsame Informationsmodell, in das die Komponententhesauri tranferiert wurden, ist Grundlage jedes weiteren Integrationsschrittes.

Kapitel 9

Analyse von Thesauri

Die eingehende Analyse komplexer Komponentensysteme ist unverzichtbare Grundlage jedes Verfahrens, das zum Ziel hat, diese Komponentensysteme nicht nur oberflächlich und willkürlich zu verbinden, sondern eine bestmögliche Integration ihrer Elemente und Strukturen zu erreichen. Alle aus der Literatur bekannten Ansätze der Thesaurusintegration vernachlässigen diesen Aspekt jedoch beinahe vollkommen. Explizit werden über die beteiligten Komponententhesauri nur sehr wenige Kennzahlen genannt. Inwiefern diese in die Verfahren einfließen, wird nicht kenntlich gemacht. Da aber in einigen Fällen die Ersteller der Komponententhesauri an der Entwicklung der Integrationsverfahren direkt beteiligt sind, kann davon ausgegangen werden, dass deren Kenntnisse der Komponententhesauri implizit in die Verfahren eingegangen sind.

Als Grund für die fehlende Analyse der Komponententhesauri wird von uns die Schwierigkeit angenommen, objektive Aussagen über die Qualität eines Thesaurus machen zu können. Eine algorithmische Auswertung etwa kann die Semantik der Begriffe, deren Vollständigkeit, Aktualität und Adäquatheit sowie die Semantik der Beziehungen zwischen den Begriffen eines Thesaurus nur sehr unzureichend überprüfen. Eine Analyse durch einen Experten ist jedoch zeitaufwendig und teuer.

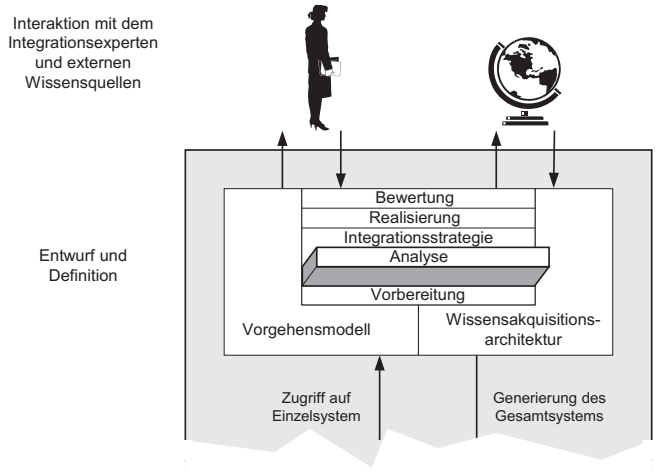

Abbildung 9.1: Erkenntnisse, die in der Analysephase über die Komponententhesauri gewonnen werden, sind Basis des weiteren Vorgehens der Begriffsintegration

In diesem Kapitel setzen wir uns daher mit der Frage auseinander, wie eine Analyse von Kom-

ponententhesauri mit möglichst geringem menschlichen Aufwand durchgeführt werden kann und welche Schlüsse aus diesen Ergebnissen für die Begriffsintegration gezogen werden können (zur Einordnung vgl. Abbildung 9.1). Dazu unterteilen wir die Analyse in Teilbereiche (Abschnitt 9.1), stellen aus der Literatur bekannte Kriterien mit dort genannten Idealwerten zusammen, leiten weitere ab und erarbeiten Schlussfolgerungsmöglichkeiten hinsichtlich der Begriffsintegration (Abschnitt 9.2). Exemplarisch wenden wir als relevant betrachtete Kriterien auf reale Thesauri an, um an konkreten Beispielen die Nützlichkeit der Kriterien zu zeigen (Abschnitt 9.3).

9.1 Teilbereiche einer Analyse

Die Analyse eines Thesaurus soll, um Folgerungen für die Begriffsintegration ziehen zu können, objektive Aussagen über Inhalte, Qualität, Funktionalität etc. des Thesaurus unabhängig von seiner aktuellen Verwendung machen.

In der Literatur gibt es eine Vielzahl von Beiträgen, die Kriterien für eine singuläre Analyse nennen [Lan72, Sly76, Sly77, Koc74, Lan77, Wer85, Lan86, Gan90, Mil91, Rec99]. Bei der automatischen Generierung von Thesauri benutzt Viegener [Vie97] einen Teil dieser Kriterien mit der klassischen Zielsetzung einer Bewertung des (generierten) Thesaurus. Die Berücksichtigung solcher Analyseergebnisse bei der Begriffsintegration ist uns jedoch nicht bekannt. Explizit werden über die beteiligten Thesauri keine (z.B. im SKC-Projekt [MWJ99, JMN+99, JW99] sowie in [SC97]) oder nur sehr wenige Kenndaten genannt (z.B. Anzahl der Bezeichnungen, Begriffe und Zyklen in der Hierarchierelation im ONIONS-Projekt [GPG99, GPS98, PAG99], Anzahl der Bezeichnungen und Begriffe, generelle Struktur der Hierarchie in [MR88]). Falls Kenndaten analysiert wurden, wird in keinem der uns bekannten Ansätze dargestellt, wie diese Analyseergebnisse in die Begriffsintegration einfließen. Dies führen wir auf mangelhaften Austausch zwischen den Wissenschaftsgemeinden der Bibliothekare und Informationswissenschaftler, die diese Analysekriterien entwickelt haben, und den Informatikern, Medizinern und anderen anwendungsnahen Wissenschaftlern zurück, die sich über 20 Jahre später mit der Begriffsintegration auseinander setzen.

Als Teilbereiche einer Analyse unterscheiden wir zwischen der Betrachtung allgemeiner Gesichtspunkte, die unabhängig von Inhalt und Funktionalität des Thesaurus sind, der Betrachtung funktionaler Kriterien, die sich auf die Erarbeitung und Handhabung der Thesauri beziehen, und der Analyse des eigentlichen Inhalts eines Thesaurus. Die in der Literatur genannten Kriterien lassen sich sämtlich diesen drei Kategorien zuordnen (vgl. hierzu auch die Diplomarbeit [Vas99], in der die Kriterien anhand einer Literaturliste zusammengetragen und geordnet wurden):

Allgemeine Analyse: Eine allgemeine Analyse betrifft etwa die Normenkonformität des Thesaurus (die wir in der Vorbereitungsphase (vgl. Kapitel 8) bereits sichergestellt haben), die Kosten für Erstellung und Pflege, die Vollständigkeit der thematischen Abdeckung durch das Thesaurusvokabular sowie Datenformate und Hard- und Software-Kompatibilität [Lan72, Koc74, Lan77, Gan90, Mil91, Rec99]. Diese Kriterien sind bei der Auswahl eines Thesaurus für die Integration in ein Informationssystem relevant, nicht jedoch für die Begriffsintegration nach bereits erfolgter Durchführung der Vorbereitungsphase.

Funktionale Analyse: Kriterien für die funktionale Analyse betreffen die Unterstützung verschiedener Darstellungsformen (dieser Aspekt wurde in der Literatur bereits ausführlich behandelt, s. z.B. [Wer85, FFW91, Pol93, JC95, NKS+99]), die Untersuchung, inwiefern ein Thesaurus überprüft und verbessert wird (zur Thesauruspflege vgl. [DIN87, Cim98]), die Verwendung von Begriffsidentifikatoren, die auch nach Bezeichnungsänderungen weiter

mit dem gleichen Begriff assoziiert werden (vgl. Desideratum III: Concept Permanence in [Cim98]), und die Beurteilung der Software und Software-Schnittstellen zum Suchen und Browsen innerhalb des Thesaurus [Lan77, Gan90, Mil91, NKS+99]. Unter technischem Blickwinkel sind diese Kriterien durchaus relevant für die Integration (z.B. Frequenz der erforderlichen Re-Integration bei Modifikationen, Zugriff auf die Thesauri für die Integration, Spezifikation von ein- oder mehrattributigen Schlüsseln). Inhaltlich können daraus jedoch keine Schlüsse für die Begriffsintegration konkret vorliegender Thesauri gezogen werden.

Inhaltliche Analyse: Die inhaltliche Analyse betrifft sowohl die *Begriffe und Benennungen* als auch die *Relationen* zwischen den Begriffen (Relationen beschreiben Begriffe näher und ordnen sie an). Zudem bilden die Begriffe mit ihren Relationen als Ganzes *Strukturen* mit bestimmten Eigenschaften (Strukturen stellen Zusammenhänge in Form von Begriffsgruppen dar). Von der Analyse der Begriffe und Benennungen, der Relationen und der Strukturen sowie eines Vergleichs der Analyseergebnisse für verschiedene Komponententhesauri erwarten wir direkte Hinweise für die Begriffsintegration.

Da wir in der inhaltlichen Analyse den Hauptansatzpunkt für die Vorbereitung einer adäquaten Begriffsintegration sehen, betrachten wir deren Teilaspekte in den folgenden Abschnitten näher. Wir gehen dabei davon aus, dass die zu untersuchenden Thesauri bereits in der Vorbereitungsphase auf die Konsistenz mit dem in Kapitel 5 entwickelten Thesaurusmodell geprüft und ggf. entsprechend angepasst wurden. Bedingungen wie die Zyklen- und Redundanzfreiheit oder der eindeutigen Benennungen können daher an dieser Stelle als erfüllt vorausgesetzt werden.

9.2 Kriterien zur Komponententhesaurusanalyse

Generell unterschieden werden können Analysen, die direkt die Qualität des Inhaltes eines Thesaurus bewerten, und Analysen, deren Ergebnis quantitative Aussagen über einen Thesaurus sind.

9.2.1 Qualitative Analysen

Qualitative Aspekte eines Thesaurus beinhalten die Güte eines Thesaurus hinsichtlich der Erfüllung allgemeiner Erwartungen an die Auswahl der Begriffe, die Form und Auswahl der diese Begriffe repräsentierenden Benennungen sowie die Etablierung von Beziehungen und Strukturen. Diese Erwartungen sind in den DIN-Normen (insbesonders [DIN87, DIN93b, DIN92]), ebenso wie in weiterführenden Thesaurus-Richtlinien (insbesonders [Wer85]) und der wissenschaftlichen Literatur (z.B. [Lan86, Sly76, Sly77, KT68]) aufgeführt.

Zur Beurteilung der Qualität der Benennungen und der durch sie repräsentierten Begriffe werden etwa die Übereinstimmung des Informationswertes eines Deskriptors mit dem repräsentierten Begriff (Wiedergabetreue), die knappe und treffende Darstellung des Begriffs durch die Benennungen, die Gebräuchlichkeit der Benennungen und deren möglichst detaillierte Repräsentation eines Begriffs genannt (vgl. [Wer85, Lan86]).

Zum Verständnis der Organisation der Begriffe innerhalb eines Thesaurus dienen Relationsanalysekriterien wie die Erfüllung der Invarianten (vgl. Abschnitt 5.2.3) sowie die Richtigkeit und Vollständigkeit der Relationen.

Werden die Relationen im größeren Zusammenhang oder die durch die Gruppen eines Thesaurus

gebildeten Begriffsmengen betrachtet, kann die Struktur des Thesaurus analysiert werden. Eine qualitative Analyse beinhaltet Aussagen über die Auswahl der Gruppen, die Vollständigkeit und Granularität der Gruppen, die korrekte und vollständige Zuordnung der Topterme zu Gruppen, eine der thematischen Abdeckung genügende Tiefe von Teilstrukturen etc.

Sämtliche qualitative Kriterien, die über die Erfüllung der Invarianten hinausgehen (Überprüfung in der Vorbereitungsphase, vgl. Kapitel 8; im Folgenden können die Invarianten also bereits als erfüllt angenommen werden), können nur aufwendig von Domänenexperten überprüft werden. Sie sind daher eher als Hinweise für den menschlichen Experten bei der Thesauruserstellung und übertragen auch bei der Föderationserstellung zu verstehen. Für den Prozess der Begriffsintegration aber, der weitestgehend automatisiert werden soll, werden einfachere Kriterien in Form von berechenbaren Kennzahlen benötigt. Anhand dieser Kennzahlen sollen Rückschlüsse über die Qualität der Thesauri sowie das weitere Vorgehen der Begriffsintegration gezogen werden können. In den folgenden Abschnitten werden daher solche Kennzahlen und deren Bedeutung erarbeitet.

9.2.2 Quantitative Analyse der Benennungen

Zusätzlich zu qualitativen Analysen durch Experten wurden statistische Tests für die Bewertung von Thesauri entwickelt. Zu den Pionieren zählen hier Kochen und Tagliacozzo [KT68], die eine Reihe von Thesauri anhand eines Konnektivitätsverhältnisses (connectedness ratio) und eines Zugangsmaßes (accessibility measure) bewerteten. Höhere Werte wurden als Indiz für einen besseren Thesaurus betrachtet.

Die von Kochen und Tagliacozzo entwickelten Maße wurden von Van Slype [Sly76] wesentlich weiterentwickelt. Neue Maßzahlen wurden ergänzt und Idealwerte anhand der Auswertung von mehreren als „gut“ angesehenen Thesauri ermittelt. Einige der Kriterien wurden wiederum für einzelne Anwendungsbereiche verfeinert (vgl. z.B. [Mil92, Ham89]).

Zusätzlich zu der singulären Bewertung der Güte eines Thesaurus besteht die Möglichkeit einer vergleichenden Bewertung anhand erster Integrationsstellen. Unter der Voraussetzung, dass äquivalente Begriffe in zumindest zwei verschiedenen Thesauri identifiziert werden können, bietet sich der Vergleich der Strukturen dieser Thesauri an. Solch eine vergleichende Bewertung ermöglicht beispielsweise das Auffinden fehlender oder falscher Beziehungen [RPR+98, RWP98][1]. Da eine erste Begriffsintegration jedoch bereits erfolgt sein muss, betrachten wir eine in diesem Sinne vergleichende Analyse nicht als Bestandteil der Analyse der Komponententhesauri sondern als Möglichkeit der iterativen Verbesserung des Integrationsergebnisses (vgl. Kapitel 10).

Wir betrachten die Kriterien der singulären Analyse als Ausgangsbasis und untersuchen ihre Bedeutung für die Begriffsintegration sowohl hinsichtlich der Verfahren (Aussagen durch **V** gekennzeichnet) als auch hinsichtlich Schlussfolgerungen über das erwartete Ergebnis der Begriffsintegration (Aussagen durch **E** gekennzeichnet):

Anzahl der Deskriptoren: Die naheliegendste Kennzahl, die man über einen Thesaurus gewinnen kann, ist die Anzahl der Deskriptoren.

Isoliert betrachtet, kann anhand der Anzahl der Deskriptoren noch keine Aussage getroffen werden, ob ein Fachgebiet besonders umfassend abgedeckt wird. Dazu ist zusätzliches

[1] Innerhalb des SKC-Ansatzes (vgl. hierzu insbesondere [MWJ99]) wird der Begriff Ähnlichkeitsmaß (similarity measure) irreführend verwendet, da kein qualitatives Maß für die Kompatibilität der Komponententhesauri gemeint ist. Stattdessen handelt es sich um eine Folge von Paaren zusammenpassender Objekte oder Begriffe, also um eine Abbildung. Eine Analyse dieser Abbildung erfolgt nicht.

Wissen über die Intention des Thesaurus sowie eine Analyse der Relationen und Strukturen erforderlich. Jedoch kann eine erste Aussage über den Umfang des Begriffsnetzes einen Hinweis auf die zu erwartende Komplexität der Begriffsintegration liefern. Thesauri mit wenigen hundert Deskriptoren ermöglichen es, aufwendige Verfahren mit umfassender Interaktion mit dem menschlichen Experten auf den gesamten Thesaurus anzuwenden, während die Anwendung dieser Verfahren bei Thesauri mit mehreren zehntausend Deskriptoren auf Ausschnitte beschränkt werden sollte. (**V**)

Äquivalenzverhältnis: Als Äquivalenzverhältnis wird das Verhältnis der Anzahl der Nicht-Deskriptoren zur Anzahl der Deskriptoren bezeichnet [Lan86, Sly77]. Diese Kennzahl, in [Wer85] terminologischer Reichtum genannt, beschreibt, wie gut die Bedeutung der Deskriptoren durch Nicht-Deskriptoren näher spezifiziert ist.

Ein kleiner Wert (< 0.4) deutet auf eine hohe Zuverlässigkeit der Nicht-Deskriptoren hin (wenig Quasi-Synonyme), erfordert jedoch zugleich die intensivere Auswertung der weiteren Relationen und Definitionen bzw. das Hinzuziehen von Zusatzwissen, wenn bei Betrachtung der Benennungen keine Beziehungen etabliert werden können. Bei einem größeren Wert ist die Gefahr größer, dass es sich bei den Nicht-Deskriptoren um Quasi-Synonyme handelt, insbesonders wenn zu einem Deskriptor mehr als drei Nicht-Deskriptoren angegeben sind [Wer85]. (**V**)

Bei der Begriffsintegration sollten Nicht-Deskriptoren aus Thesauri mit kleinem Äquivalenzverhältnis bei linguistischen Verfahren ein ähnliches Vertrauen entgegengebracht werden wie Deskriptoren. Bei Thesauri mit großem Äquivalenzverhältnis sollte das Vertrauen geringer sein. (**V**)

Es sollte überprüft werden, ob die Nicht-Deskriptoren in einem Thesaurus mit großem Äquivalenzverhältnis als eigenständige Deskriptoren in einem Thesaurus mit kleinem Äquivalenzverhältnis aufgeführt sind (**V**). Je nach Integrationsziel wird durch die Begriffsintegration der Begriffsumfang von Begriffen mit Quasi-Synonymen verkleinert oder es werden viele Benutze-Kombinationsbeziehungen etabliert. (**E**)

Benutze-Kombination-Beziehungs-Anteil: Der Benutze-Kombination-Beziehungs-Anteil ist die Anzahl der Benutze-Kombination-Beziehungen im Verhältnis zu der Anzahl der Äquivalenzbeziehungen.

Bei einem Benutze-Kombination-Beziehungs-Anteil ungleich Null gilt es, bei der Begriffsintegration die entsprechenden Nicht-Deskriptoren gesondert zu berücksichtigen. (**V**)

Fremdwörteranteil: Für multilinguale Thesauri besteht die Möglichkeit, einen Richtwert für den Fremdwörteranteil zu erhalten. Der Fremdwörteranteil wird als Verhältnis der Kardinalität der Schnittmenge der Benennungen in einer Sprache mit den Benennungen in einer anderen Sprache zu der Gesamtzahl der Benennungen in der ersten Sprache berechnet.

Bei großem Fremdsprachenanteil in einem Thesaurus ist dieser näher zu untersuchen. Prinzipiell ist eine solche „eingebettete Fremdsprache“ ein Indiz dafür, dass es sich um einen Spezial- oder Fachgebietsthesaurus handelt, der eine große Anzahl fremdsprachiger Fachbezeichnungen beinhaltet: insbesondere in naturwissenschaftlichen Bereichen wie Medizin, Biologie und Chemie haben sich etwa lateinische Fachbezeichnungen durchgesetzt, die in verschiedenen Sprachen identisch sind und die Kommunikation der Wissenschaftler erleichtert.

Sollen solche Spezial-Vokabulare mit weniger spezialisierten Vokabularen integriert werden, sind besondere Hilfsmittel wie Wörterbücher für fremdsprachige Bezeichnungen erforder-

lich. Enthalten die zu integrierenden Thesauri jedoch dasselbe fremdsprachige Spezialvokabular, wird eine hohe Erfolgsquote beim Finden von Inter-Thesaurus-Beziehungen durch lexikalische Verfahren erwartet (größere Zuverlässigkeit lexikalischer Analysen). (**V**)

Präkoordinationsgrad: Es gilt festzustellen, inwiefern der Thesaurus bereits präkoordinierte Begriffe enthält, also Begriffe, die in sich bereits eine Verknüpfung mehrerer begrifflicher Einheiten sind (z.B. in Form von Komposita, Mehrwortbenennungen und adjektivischen Wortgruppen) oder aber überwiegend atomare Begriffe, die vom Benutzer bei Bedarf kombiniert werden können, um komplexere begriffliche Einheiten auszudrücken (Postkoordination; vgl. auch [Wer85, S. 56] und Desideratum I: Content in [Cim98]).

In [Lan86, Sly76, Sly77] wird als Maß der Präkoordination die durchschnittliche Anzahl der signifikanten Termini je Deskriptor (Anzahl der Wörter pro Deskriptor ohne Berücksichtigung von Stoppwörtern wie *the, of, from*) vorgeschlagen. Seien a, b, c, d die Anzahl der einzel-, zwei-, drei- bzw. vierwortigen Deskriptoren. Dann bedeutet dies

$$\text{Präkoordinationsgrad} = \frac{(a \times 1) + (b \times 2) + (c \times 3) + (d \times 4)}{a + b + c + d}$$

Abhängig von der Sprache (deutsche Komposita werden als Einzelwort-Deskriptoren betrachtet) bedeuten nach [Sly77, Sly76] Werte zwischen 1.5 und 2 für Englisch und Französisch bzw. zwischen 1.1 und 1.2 für Deutsch einen guten Wert für einen präkoordinierten Thesaurus.

Aussagen für die Begriffsintegration können erst bei einer komparativen Analyse getroffen werden. Es gilt jedoch als wahrscheinlich, dass ein Deskriptor spezifischer ist, je mehr signifikante Termini er enthält. Ein hoher Präkoordinationsgrad lässt somit auf eine größere Tiefe der Abdeckung eines Fachgebietes schließen, ein niedriger auf eine größere Breite. Es ist somit zu erwarten, dass später die Begriffe des spezifischeren Thesaurus in der Regel unter die des breiteren Thesaurus geordnet werden. (**E**)

Es sei angemerkt, dass Integrationsverfahren für präkoordinierte Thesauri aufgrund der Berücksichtigung der Zusammensetzung der Begriffe komplexer sind als solche für ausschließlich postkoordinierte Thesauri. Es werden zusätzlich Verfahren zur Begriffszerlegung benötigt. (**V**)

Flexibilität: Nach [Lan86, Sly77, Sly76] bezeichnet die Flexibilität den durchschnittlichen Kehrwert des Verhältnisses der Anzahl der signifikanten Wörter einer Mehrwort-Benennung zu der Anzahl der einwortigen Benennungen, die eines dieser Wörter sind, also eigenständig als Deskriptoren oder Nicht-Deskriptoren im Thesaurus auftreten:

$$\text{Flexibilität} = \frac{\sum_{i=1}^{n} \frac{|\{b \in B : b \in sig(b_i)\}|}{|sig(b_i)|}}{n}$$

wobei n die Anzahl der Benennungen und sig die Funktion ist, die eine Benennung auf die Menge der signifikanten Wörter der Benennungen abbildet.

Es werden Werte zwischen 0.6 und 1 als Idealwerte genannt, d.h. für jeden zweiwortigen Deskriptor muss es im Durchschnitt mindestens 1.2 Benennungen geben, die aus jeweils einem Wort des Deskriptors bestehen, für jeden dreiwortigen Deskriptor mindestens 1.8 Benennungen. Je größer der Wert ist, desto größer ist die Möglichkeit, dass ein Begriff in der Hierarchierelation näher spezifiziert wird.

Da spezifischere Begriffe nicht unbedingt gleiche Wörter enthalten, ist die Flexibilität im Zusammenhang mit einer Analyse der Hierarchierelation zu betrachten. Steht etwa

eine große Flexibilität bei nur wenigen Hierarchiebeziehungen einer großen Flexibilität mit vielen Hierarchierelationen gegenüber, lässt dies auf einen unterschiedlichen Gebrauch der Hierarchierelationen schließen. (**V**)

Des weiteren lässt eine große Flexibilität, berechnet über alle zu integrierenden Komponententhesauri, auf eine größere Anzahl von zu etablierenden Hierarchiebeziehungen schließen als eine geringe Flexibilität. (**E**)

Definitionsanteil: Die Präzision der Bedeutung eines Deskriptors kann durch vorhandene Definitionen bzw. Erläuterungen verbessert werden. Als Definitionsanteil wird in [Lan86, Sly77, Sly76] das Verhältnis der Deskriptoren mit Definitionen oder Erläuterungen zu der Gesamtzahl der Deskriptoren definiert.

Ein großer Definitionsanteil verspricht zumindest für den menschlichen Experten eine präzise Interpretation der Bedeutung. Auch Integrationsverfahren können ggf. die Definitionen auswerten. Ein geringer Definitionsanteil erschwert die Integration durch die weniger präzise wiedergegebene Bedeutung. Hier werden im Verlauf des Integrationsprozesses häufigere Korrekturen auch der Entscheidungen des menschlichen Experten erwartet. (**V**)

Mittlere Länge der Deskriptoren: Als mittlere Länge der Deskriptoren wird die mittlere Anzahl von Buchstaben eines Deskriptors bezeichnet. In [Sly76] wird als optimale Länge für englische Deskriptoren 30-50 Zeichen genannt, in [Wer85] als maximale Länge 48 Zeichen für deutsche Deskriptoren. Die erste Zahl erscheint jedoch sehr hoch[2], die zweite recht willkürlich an technischen Einschränkungen festgelegt.

Eine geringe mittlere Länge lässt sich auf einen geringen Präkoordinationsgrad zurückführen, eine große mittlere Länge wird mit einem hohen Präkoordinationsgrad einhergehen. Daher können aus dieser Kennzahl ähnliche Folgerungen wie aus dem Präkoordinationsgrad gezogen werden.

Anteil identischer Benennungen: Thesaurusübergreifend kann festgestellt werden, wie groß der Anteil identischer Benennungen in den verschiedenen Thesauri ist.

Obwohl identische Benennungen aufgrund von Polysemie und Homonymie keine eindeutigen Indikatoren für identische Begriffe sind, kann doch eine erste Aussage über die Ähnlichkeit der Vokabulare getroffen werden. Forschungsergebnisse belegen eine Wahrscheinlichkeit größer als 65 % dafür, dass identische Benennungen identische Begriffe repräsentieren (vgl. z.B. [BKK+98]), wenn die Thesauri aus ähnlichen Fachgebieten stammen. Unter Berücksichtigung des Äquivalenzverhältnisses kann als Erwartungswert für die minimale Anzahl der Inter-Thesaurus-Äquivalenzbeziehungen zwischen zwei Thesauri angegeben werden:

$$IT\ddot{A}_{min} = 0.65 \times \frac{|\theta_1.B \cap \theta_2.B|}{max(\ddot{A}Q_{\theta_1}, \ddot{A}Q_{\theta_2}) + 1}$$

wobei $|\theta_1.B \cap \theta_2.B|$ die Anzahl der identischen Benennungen und $\ddot{A}Q_{\theta_i}$ das Äquivalenzverhältnis von Thesaurus θ_i sind. (**E**)

[2]Selbst die längsten Deskriptoren in GEMET sind innerhalb dieses Intervalls, z.B. *integrated environmental protection technology* mit 47 Zeichen, die meisten Benennungen, auch wenn es sich um Mehrwortbenennungen handelt, von kürzerer Länge, z.B. *area of potential pollution* mit 28 Zeichen.

9.2.3 Quantitative Analyse der Relationen

Folgende Kennzahlen können zu einer quantitativen Analyse der Relationen herangezogen werden:

Konnektivität: Die Konnektivität (engl. connectivity oder connectedness ratio) bezeichnet nach [Sly77, Lan86, Sly76, KT68] das Verhältnis der Anzahl der Deskriptoren, die mindestens eine Beziehung zu einem anderen Deskriptor eingehen, zu der Gesamtzahl der Deskriptoren. Als idealer Wert wird 1 angesehen, d.h. es gibt keine unverbundenen Deskriptoren.

Sollte die Konnektivität kleiner als 1 sein, gilt den unverbundenen Deskriptoren besonderes Augenmerk, um sie zu integrieren und damit für die Nutzung der Thesaurusföderation besser zugänglich zu machen. (**V**)

Zugänglichkeit: In [Lan86, KT68, Sly76] wird als Zugänglichkeit (engl. accessibility ratio) die durchschnittliche Anzahl der (ein- und ausgehenden) Beziehungen bezeichnet, die ein Deskriptor mit anderen Deskriptoren eingeht[3]. Als Idealwerte betrachtet über Hierachie- und Assoziationsbeziehungen wird das Intervall von 2 bis 5 vorgeschlagen. So kann eine gute Unterstützung für das Information Retrieval gegeben werden, ohne dass die durch zu viele Beziehungen entstehende Komplexität hindert.

Ein sehr unterschiedlicher Wert der Zugänglichkeit lässt auf eine unterschiedliche Verwendung der Relationen schließen. Entsprechend kann das Gewicht, das während der Begriffsintegration den vorhandenen Beziehungen gegeben wird, variiert werden. (**V**)

Bei Erstellung der Thesaurusföderation soll ein guter Wert angestrebt werden. Das bedeutet, für Deskriptoren mit sehr wenigen Verbindungen weitere Verbindungen zu suchen. Bei Deskriptoren mit sehr vielen Beziehungen kann untersucht werden, ob Beziehungen entfernt werden können. (**E**)

Hierarchie-/Assoziationsrelationsverhältnis: Wiederum einen guten Hinweis auf die Verwendung der Relationen erhält man durch Betrachtung des Verhältnisses der Anzahl der Hierarchiebeziehungen zur Anzahl der Assoziationsbeziehungen.

Ein ähnliches Hierarchie-/Assoziationsrelationsverhältnis lässt eine ähnliche Verwendung der Relationen erwarten. Ansonsten sollte die Herkunft der Unterschiede näher untersucht werden und das Ergebnis von den Integrationsverfahren berücksichtigt werden. (**V**)

Das Hierarchie-/Assoziationsrelationsverhältnis ist zudem im Zusammenhang mit der Zugänglichkeit zu betrachten. Ein guter Zugänglichkeitswert wird bei einem großen Anteil von Assoziationen relativiert, da die Semantik einer Assoziationsbeziehung vager als die einer Hierarchiebeziehung ist.

Abstraktions-/Bestandsverhältnis: Das Abstraktions-/Bestandsverhältnis bezeichnet das Verhältnis der Anzahl der Abstraktionsbeziehungen zur Anzahl der Bestandsbeziehungen.

Gemeinsam mit dem Hierarchie-/Assoziationsrelationsverhältnis ist das Abstraktions-/Bestandsverhältnis ein wichtiger Indikator zur Klassifikation der Thesauri (vgl. Abschnitt 9.2.5).

Anteil polyhierarchischer Begriffe: Thesauri erlauben es, für einen Begriff mehr als einen Oberbegriff anzugeben. Somit kann der Unterbegriff besser definiert werden, andererseits

[3] Hierarchiebeziehungen werden dabei nur in eine Richtung betrachtet, etwa in Richtung der Unterbegriffe.

kann jedoch die Spezifität des Begriffs gestört werden [Mil97]. Der Anteil polyhierarchischer Begriffe bezeichnet den Anteil, den Begriffe mit mehr als einem Oberbegriff an der Gesamtzahl der Begriffe ausmacht.

Wiederum deuten sehr unterschiedliche Werte auf einen unterschiedlichen Gebrauch der Relationen, in diesem Fall der Hierarchierelation hin. Eine weitere Unterscheidung der Polyhierarchie innerhalb der Abstraktions- und der Bestandsrelation ist möglich. Aufgrund der Semantik der Relationen wird prinzipiell erwartet, innerhalb der Bestandsrelation einen größeren Anteil polyhierarchischer Begriffe aufzufinden [MBF90].

Begriffe mit mehr als zwei Oberbegriffen sollten während der Begriffsintegration gesondert untersucht werden (**V**). Zielsetzung der Föderation ist es, auch dort maximal zwei Abstraktionsoberbegriffe pro Begriff zu haben. (**E**)

9.2.4 Quantitative Analyse der Struktur

Die Analyse der Struktur betrifft sowohl die Gruppen eines Thesaurus (falls vorhanden) als auch die Graphenstruktur des Thesaurus, die durch die Relationen gegeben ist:

Anzahl der Gruppen: Die Anzahl der Gruppen ist die einfachste Kennzahl der Strukturanalyse [Sly77, Sly76]. Eine Interpretation ist jedoch schwierig, da die Anzahl der sinnvollen Gruppen von der Breite des abgedeckten Fachgebietes abhängt.

Gruppengröße: Die mittlere Gruppengröße wird als mittlere Anzahl der Deskriptoren pro Gruppe definiert [Lan86, Sly77, Sly76]. Aufgrund der polyhierarchischen Struktur sowie der Definition der Gruppen kann ein Deskriptor in mehr als einer Gruppe vorkommen, die Summe der Größen aller Gruppen also größer als die Anzahl aller Deskriptoren sein.

Um eine erste Einschätzung über die Spannbreite zu bekommen, sind zusätzlich die minimale und maximale Gruppengröße relevant. Der fachliche Schwerpunkt eines Thesaurus wird durch große Gruppen bestimmt.

Homogenität der Gruppen: Die Homogenität der Gruppen kann durch die Varianz, berechnet über die Gruppengrößen g_i, ausgedrückt werden:

$$\text{Gruppengrößenvarianz} = \frac{\sum_{i=1}^{n} (|g_i| - \overline{g})^2}{n}$$

wobei $\overline{g}$ die mittlere Gruppengröße und n die Anzahl der Gruppen sind.

Die Varianz ist ein Hinweis auf den gleichmäßigen Aufbau des Thesaurus. Bei großer Varianz sind die Ursachen zu untersuchen.

Bei Erstellung einer Thesaurusföderation kann es das Ziel sein, eine geringere Varianz zu erzielen als die eines oder mehrerer Komponententhesauri bzw. einer bereits vorhandenen Föderation. Wenn eine solche homogenere thematische Abdeckung erreicht werden soll, müssen die Komponententhesauri mit entsprechender fachlicher Ausrichtung ausgewählt werden. Konnten entsprechende Thesauri gefunden werden, wird durch die Integration eine verbesserte Homogenität der Gruppen erwartet. (**E**)

Größendifferenzen innerhalb Föderierter Gruppen: Wurden die Deksriptoren der zu vergleichenden Thesauri bereits einer identischen Menge von Gruppen der Föderation zugeordnet, können die Größendifferenzen der Föderierten Gruppen bezogen auf die einzelnen Thesauri berechnet werden.

Anhand der Größendifferenzen kann festgestellt werden, welche Teilbereiche des Fachgebiets ähnlich intensiv und welche sehr unterschiedlich abgedeckt werden.

Bei größeren Gruppen und geringer Größendifferenz wird ein höherer Anteil an Inter-Thesaurus-Beziehungen erwartet als in kleinen Gruppen oder bei großer Größendifferenz. (**E**)

Verbindungseinheit: Die Verbindungseinheit ist eine Kennzahl für die Anzahl isolierter Strukturen, also Teilgraphen, die durch keinerlei andere Beziehung mit anderen Teilgraphen verbunden sind, im Verhältnis zu der Gesamtzahl der Deskriptoren:

$$\text{Verbindungseinheit} = 1 - \frac{\text{Anzahl isolierter Strukturen} - 1}{|D| - 1}$$

Bei Thesauri wird ein Wert von 1 normalerweise nicht erreicht, da ein solcher aus mehreren Teilgraphen, die einzelne Bereiche des Fachgebiets repräsentieren, besteht. An einigen Stellen bestehen Verbindungen (z.B. durch Polyhierarchie oder Assoziationen), andere Teilgraphen bleiben isoliert, vgl. [Vie97, S. 128].

Einzelne Deskriptoren oder isolierte Teilgraphen mit wenigen Deskriptoren sollen innerhalb einer Thesaurusföderation möglichst vermieden werden. Daher können solche Deskriptoren bzw. Teilgraphen gesondert auf Integrationsstellen untersucht werden. (**V**)

Mittlere Höhe: Als mittlere Höhe wird die mittlere Länge der Hierarchiepfade von den Toptermen zu Deskriptoren ohne weitere Unterbegriffe plus Eins bezeichnet [Ham89]:

$$\overline{h} = \frac{\sum_{i=1}^{n}(l_i + 1)}{n}$$

mit n der Anzahl aller Pfade von Toptermen zu Deskriptoren ohne weitere Unterbegriffe und l_i der Länge des i-ten Pfades.

Eine geringe mittlere Höhe ist Indikator für einen eher breit angelegten Thesaurus, der Begriffe nicht oder selten stärker differenziert. Eine größere mittlere Höhe spiegelt die stärkere Zielsetzung eines Klassifikationssystems wider, den Hierarchierelationen muss bei der Begriffsintegration besondere Beachtung geschenkt werden. (**V**)

Höhenvarianz: Die Höhenvarianz wird berechnet als:

$$\text{Höhenvarianz} = \frac{\sum_{i=1}^{n}(l_i + 1 - \overline{h})^2}{n}$$

wobei $\overline{h}$ die mittlere Höhe und n die Anzahl der Pfade und h_i die Länge des i-ten Pfades sind.

Je größer die Varianz ist, desto inhomogener ist der Thesaurus hinsichtlich des Detaillierungsgrades, mit dem Begriffe durch Hierarchiebeziehungen verfeinert werden. Eine große Varianz deutet auf eine sehr unterschiedliche Bedeutung einzelner Begriffe hinsichtlich des Gesamtvokabulars hin. Die Integration der „bedeutenderen Begriffe“, also solcher, die Oberbegriffe einer großen Hierarchie sind, wird angestrebt. (**E**)

9.2.5 Klassifikation der Thesauri

Neben den einfachen hierarchischen Klassifikationssystemen, die durch vollkommenen oder mindestens überwiegenden Verzicht auf Definitionen und Erläuterungen, Synonyme und Assoziationsbeziehungen eine vereinfachte Thesaurusausprägung sind, können in der Praxis zwei zentrale Formen von Thesauri unterschieden werden:

Definitionswörterbuchartige Thesauri: Definitionswörterbuchartige Thesauri versuchen, die Begriffe eines Fachgebiets möglichst exakt zu repräsentieren, besitzen somit also die Intention eines Wörterbuches. Um dieses Ziel zu erreichen, werden beinahe ausschließlich strenge Synonyme aufgeführt, es werden möglichst viele Definitionen aufgenommen, aufgrund des stärkeren Definitionscharakters überwiegen Abstraktionsbeziehungen, Bestands- und Assoziationsbeziehungen werden weniger häufig verwendet.

Information-Retrieval-Thesauri: Bei Information-Retrieval-Thesauri steht weniger die Exaktheit der Begriffsdefinitionen im Vordergrund, sondern das Ziel, die Retrievalqualität hinsichtlich der Vollständigkeit der Ergebnisse bei Suchanfragen zu optimieren. Um dies zu erreichen, werden Begriffe zusammengefasst (großer Anteil an Quasi-Synonymen) und es existieren möglichst viele Querbezüge in Form von Assoziationsbeziehungen (großer Anteil von Assoziationsbeziehungen), die den Assoziationen des menschlichen Informationssuchenden entsprechen. Definitionen bzw. Erläuterungen werden in der Regel nur dann aufgeführt, wenn die Gefahr einer Mehrdeutigkeit besteht.

Es wird ersichtlich, dass die Ergebnisse der quantitativen Analyse mit zusätzlichen Stichprobenuntersuchungen der Synonyme (zur Unterscheidung von strengen Synonymen und Quasi-Synonymen) eine Klassifikation der untersuchten Thesauri ermöglichen.

Um den unterschiedlichen Zielsetzungen der Thesaurustypen gerecht zu werden, muss eine Entscheidung für einen Aufbau eher im Sinne eines Wörterbuches oder eines Information-Retrieval-Thesaurus zu Anfang der Begriffsintegration getroffen werden. Entsprechend dieser Entscheidung, die eine Erwartung an das Integrationsergebnis ausdrückt (**E**), gewichten die Begriffsintegrationsverfahren die Berücksichtigung der Beziehungen in den entsprechend klassifizierten Komponententhesauri (**V**).

9.3 Evaluierung ausgewählter Thesauri

Zur Demonstration der möglichen Auswertung der Analyseergebnisse betrachten wir die beschriebenen Maßzahlen für die beiden „großen“ Thesauri AGROVOC und GEMET. Die GCMD Parameter Valids werden aufgrund fehlender Nicht-Deskriptoren, fehlender Assoziationsbeziehungen, einer per Definition identischen Länge für alle Pfade sowie der wesentlich geringeren Anzahl an Deskriptoren (906 unterschiedliche Benennungen) in unserem Beispielszenario als eine besondere Form der Schlagwortliste erkannt. Als solche unterscheiden die GCMD Parameter Valids sich offensichtlich von den reichhaltigeren Thesauri AGROVOC und GEMET und werden nicht in die Analyse einbezogen.

9.3.1 Analyse der Benennungen

Tabelle 9.1 zeigt den Vergleich der Kennzahlen für die Analyse der englischsprachigen Benennungen nach Durchführung einer wie in Abschnitt 8.2 beschriebenen Benennungsnormierung.

Anhand dieser Maßzahlen wird ersichtlich, dass AGROVOC mehr als die dreifache Anzahl an Begriffen beinhaltet (repräsentiert durch Deskriptoren sowie Nicht-Deskriptoren, die in Benutze-Kombination-Beziehungen mit Deskriptoren stehen). Jedoch enthält AGROVOC mit 53 % einen sehr großen Fremdwörteranteil. Eine Durchsicht der identifizierten Fremdwörter ergibt, dass es sich beinahe ausschließlich um lateinische Bezeichnungen von Tieren oder Pflanzen bzw. Tier- oder Pflanzenkrankheiten handelt (z.B. *tithonia rotundifolia, cronartium quercuum*). Der große

Kennzahl	GEMET	AGROVOC
Anzahl der Deskriptoren	5398	16394
Äquivalenzverhältnis	0.22	0.63
Benutze-Kombination-Beziehungs-Anteil	0 %	7.17 %
Fremdwörteranteil	3.56 %	53 %
Präkoordinationsgrad	1.93	1.52
Flexibilität	0.27	0.25
Definitionsanteil	99 %	7 %
Mittlere Länge der Deskriptoren	14.65	12.51
Anteil identischer Benennungen	28.77 % (1896 von 6590)	7.09 % (1896 von 26750)

Tabelle 9.1: Kennzahlen der Analyse der Benennungen

Anteil solcher fremdsprachigen Fachbezeichnungen belegt, dass es sich bei AGROVOC um einen Fachgebietsthesaurus handelt.

Da der Schwerpunkt der von uns betrachteten Begriffsintegration nicht auf der Integration von sehr speziellen Fachvokabularen in allgemeinere Vokabulare sein soll, steht die Betrachtung des entsprechenden AGROVOC-Ausschnittes nicht im Vordergrund. Größere Priorität erhalten die verbleibenden 47 % der AGROVOC Bezeichnungen.

Weiterhin ist festzustellen, dass in GEMET durchschnittlich nur zu jedem fünften Deskriptor ein Nicht-Deskriptor angegeben ist, in AGROVOC hingegen zu zwei von drei Deskriptoren. Jedoch ergeben Stichproben, dass in GEMET die Nicht-Deskriptoren in der Regel Synonyme sind (z.B. Deskriptor *subterranean water* und Nicht-Deskriptor *groundwater*), in AGROVOC häufig jedoch Quasi-Synonyme (z.B. Deskriptor *eyes* und Nicht-Deskriptoren wie *eyelids, pupils, eye lens*), die wohl im Fachgebiet, nicht aber im Thesaurus voneinander unterschieden werden. Dies ist ein starkes Indiz dafür, dass GEMET eher ein Definitionsthesaurus ist, der die Begriffe jeweils möglichst exakt zu definieren versucht, während AGROVOC ein Thesaurus zum Zwecke des Information Retrieval ist, der Begriffe zusammenfasst, um vor allem die Retrievalqualität hinsichtlich der Vollständigkeit zu optimieren. Der große Anteil an Quasi-Synonymen ist bei der Begriffsintegration entsprechend zu berücksichtigen, insbesondere hinsichtlich der Verlässlichkeit lexikalischer Vergleiche.

Nur in AGROVOC wird die Benutze-Kombination-Relation angewandt. Algorithmen zur Begriffsintegration müssen also für AGROVOC diese Beziehung und die relevanten Nicht-Deskriptoren entsprechend behandeln.

Aufgrund des jeweiligen Präkoordinationsgrades, der für GEMET an der oberen und für AGROVOC an der unteren Grenze des Bereiches für einen guten präkoordinierten Thesaurus ist, werden beide Thesauri als präkoordinierte Thesauri identifiziert.[4] Zur Begriffsintegration sind also Verfahren für die Mehrwortanalyse erforderlich.

Hinsichtlich der Flexibilität sind beide Thesauri mit sehr ähnlichen Werten unterhalb des Idealbereiches. Es wird daher eine wenig ausgeprägte Hierarchierelation vermutet.

Sehr große Unterschiede weisen die beiden Thesauri hinsichtlich einer natürlichsprachigen Definition der Begriffe auf. In AGROVOC sind solche Definitionen nur als Erläuterungen vorhanden, wenn die Thesaurusersteller die Möglichkeit einer Fehlinterpretation vorhersehen. GEMET hingegen besitzt für beinahe jeden Begriff eine Definition. Dies verstärkt den Anspruch eines

[4]Wie erwartet werden kann, bedeutet eine größerer Präkoordinationsgrad zugleich auch eine größere mittlere Länge der Deskriptoren.

Definitionsthesaurus. Jedoch ist bei einer Analyse der Definitionen festzustellen, dass in der untersuchten Version 1.5 des Thesaurus ein Teil der Definitionen nur Hinweise für den Definitionsersteller enthalten (z.B. für *freshwater biology* lautet der Definitiontext *(missing: task of CB5)*).

Schließlich kann aufgrund der Anzahl identischer Benennungen sowie der jeweiligen Äquivalenzverhältnisse 756 als minimal erwartete Anzahl der Inter-Thesaurus-Äquivalenzbeziehungen angegeben werden.

9.3.2 Analyse der Relationen

In Tabelle 9.2 sind die Analyseergebnisse, die die Relationen betreffen, aufgeführt.

Kennzahl	GEMET	AGROVOC
Konnektivität	0.98 (89 unverbundene Deskriptoren)	1.00 (2 unverbundene Deskriptoren)
Zugänglichkeit	2.31	3.51
Hierarchie-/Assoziationsverhältnis	0.19	0.86
Abstraktions-/Bestandsverhältnis	20.71	0.58
Anteil polyhierarchischer Begriffe	0.94 % (51)	3.27 % (536)

Tabelle 9.2: Kennzahlen der Analyse der Relationen

Der Konnektivitätswert von GEMET, der kleiner als 1 ist, und die somit relativ große Anzahl unverbundener Deskriptoren deutet mit den als noch fehlend markierten Definitionen darauf hin, dass mit der untersuchten Version 1.5 noch keine vollständig abgeschlossene Version vorliegt. Überarbeitungen sind zu erwarten. Durch die Integration mit AGROVOC, der eine beinahe optimale Konnektivität besitzt, können Hinweise für die Herstellung einer besseren Konnektivität sowie einer besseren Zugänglichkeit erwartet werden.

Auffallend ist der große Unterschied der Hierarchie-/Assoziationsverhältnisse. Innerhalb von GEMET existieren fünfmal soviele Hierarchiebeziehungen wie Assoziationsbeziehungen, für AGROVOC ist der Faktor kleiner als 1.2. Dies bestärkt die Annahme, dass es sich bei GEMET um einen wörterbuchartigen Thesaurus handelt, während AGROVOC für das Information Retrieval optimiert wurde und in Form von Assoziationsbeziehungen möglichst viele Querbezüge enthält. Der hohe Zugänglichkeitswert innerhalb von AGROVOC kann also durch die vergleichsweise großen Anzahl von Assoziationsbeziehungen erklärt werden. Während der Begriffsintegration sollen diese explizit berücksichtigt werden. Eine stichprobenartige Untersuchung dieser Assoziationsbeziehungen zeigt bereits, dass häufig „versteckte Hierarchien“ enthalten sind, also Hierarchiebeziehungen, die z.B. aufgrund der fehlenden Gültigkeit in allen Kontexten oder der vergleichsweise geringen Bedeutung nicht als solche ausgewiesen wurden (z.B. *agricultural chemicals* mit u.a. den assozierten Deskriptoren *algicides*, *anabolics* und *antibiotics*).

Der größte Unterschied bei der Verwendung der Relationen zeigt sich hinsichtlich des Abstraktions-/Bestandsverhältnisses. Wie in Abschnitt 8.1.2.1.2, S. 144, bereits dargestellt, existiert in AGROVOC die überwiegende Zahl der Bestandsrelationen jedoch innerhalb der lateinischen Begriffstaxonomie. Neben der Fremdsprachigkeit ist der somit offensichtlich andere Aufbau dieses Thesaurusausschnitts weiterer Grund, diesen Ausschnitt gesondert zu betrachten.

Die Polyhierarchie ist innerhalb von AGROVOC deutlich stärker ausgeprägt. Nur hier gibt es auch Begriffe (insgesamt 14) mit 3 Oberbegriffen. Dieses Ergebnis konnte bereits aufgrund der größeren Zahl der Bestandsrelationen erwartet werden. Es liefert auch einen Teilbeitrag zur Erklärung des geringeren Präkoordinationsgrades. Bei der Begriffsintegration muss die Polyhie-

rarchie berücksichtigt werden, wobei besonders bei bereits polyhierarchischen Begriffen weitere polyhierarchische Beziehungen nur etabliert werden sollen, wenn die Interpretation der Bedeutung der Deskriptoren nicht gefährdet wird.

9.3.3 Analyse der Struktur

Analysergebnisse bezüglich der Struktur der Thesauri GEMET und AGROVOC sind in Tabelle 9.3 dargestellt.

Kennzahl	GEMET	AGROVOC
Anzahl der Gruppen	35	0
Gruppengröße	154.34 (0–504)	– (441.62; 0 – 5678)
Gruppengrößenvarianz	18052.80	– (961621.69)
Größendifferenz Föderierter Gruppen	0 – 5422	
Verbindungseinheit	0.98 (100 isolierte Strukturen)	1.00 (7 isolierte Strukturen)
Mittlere Höhe	4.49	4.17
Höhenvarianz	1.42	1.34

Tabelle 9.3: Kennzahlen der Analyse der Struktur

AGROVOC enthält ursprünglich keine Gruppen. Erst innerhalb der Vorbereitungsphase (vgl. Abschnitt 8.1.3) werden die Topterme Gruppen zugewiesen. Bei einer ähnlichen thematischen Gewichtung wie GEMET wäre aufgrund der dreifachen Anzahl an Deskriptoren eine etwa dreifache Gruppengröße zu erwarten. Stattdessen kann eine Differenz von bis zum 15-fachen festgestellt werden (die Gruppe *biosphere (organisms, ecosystems)* besitzt innerhalb von AGROVOC mit 5805 Deskriptoren mehr als die 15fache Anzahl an Deskriptoren als dieselbe Gruppe in GEMET[5]). Ebenso existieren auch bei Nicht-Berücksichtigung von kleinen Gruppen in GEMET, bei denen eine automatische Zuordnung der AGROVOC-Topterme nur unzureichend möglich ist, in AGROVOC deutlich kleinere Gruppen als in GEMET (z.B. die Gruppe *risks, safety*, die in GEMET 121 in AGROVOC jedoch nur 19 Deskriptoren enthält).

Bei beiden Thesauri ist eine große Gruppengrößenvarianz festzustellen, die bei GEMET mehr als das Hundertfache der mittleren Gruppengröße beträgt, in AGROVOC mehr als das Zweihundertfache. Dies zeigt eine sehr inhomogene Gruppenstruktur, die wiederum darauf hindeutet, dass es thematische Schwerpunkte ebenso wie Randgebiete gibt, die als Gruppen zusammengefasst wurden.

Die größten Gruppen in GEMET sind *research, science* (504 Deskriptoren), *biosphere* (383 Deskriptoren), *industry, crafts; technology; equipments* (380 Deskriptoren) und *wastes, pollutants, pollution* (379 Deskriptoren). Diese unterschiedlichen Schwerpunkte belegen die allgemeine Abdeckung des übergreifenden Gebietes Umwelt, die mit GEMET erzielt werden soll. Hingegen besitzen die drei größten Gruppen AGROVOCs *biosphere (organisms, ecosystems); products, materials; chemistry, substances, processes* einen direkten Zusammenhang zur Agrarwirtschaft, ein Indiz für die größere Fokussierung des Fachthesaurus.

Die großen Größendifferenzen der gleichen Gruppen in AGROVOC und GEMET deuten – bei Übergewicht der AGROVOC-Begriffe – darauf hin, dass AGROVOC wesentlich detailliertere Begriffe enthält, die im Wesentlichen nicht mit Begriffen aus GEMET verknüpft werden können.

[5] Die 2017 Begriffe mit lateinischen Benennungen, die der Gruppe *Latin terms* zugeordnet wurden, sind beinahe ausschließlich Fachbenennungen für Pflanzen und Tiere und könnten somit die Gruppe *biosphere (organisms, ecosystems)* weiter vergrößern.

Deskriptoren in kleineren Gruppen in GEMET (z.B. *acts*) werden aufgrund fehlender Pendants in AGROVOC voraussichtlich nicht mit AGROVOC-Deskriptoren verknüpft[6]. Evtl. können Oberbegriffe zum Zusammenfassen einer Reihe solcher Deskriptoren gefunden werden[7].

Die isolierten Strukturen, die neben der großen Hauptstruktur mit über 99 % der Deskriptoren in AGROVOC existieren, konnten bis auf eine Ausnahme (der unverbundene Deskriptor *frozen storage*) auch keiner Gruppe automatisch zugeordnet werden. Da es sich aussschließlich um sehr kleine Strukturen handelt, werden Integrationsvorschläge von automatischen Begriffsintegrationsverfahren eher nicht erwartet. Zum Zwecke einer Integrationsverbesserung kann hier der menschliche Experte hinzugezogen werden.

Die geringe Anzahl von Assoziationsbeziehungen führt zu einer größeren Anzahl unverbundener Strukturen in GEMET. Aufgrund der Integration mit AGROVOC wird eine bessere Verbindungseinheit erwartet.

Hinsichtlich der mittleren Höhe und der Höhenvarianz weisen AGROVOC und GEMET sehr ähnliche Werte auf. Dies überrascht, da für einen Fachthesaurus wie AGROVOC eher eine größere Höhe als für einen allgemeineren Thesaurus wie GEMET erwartet wird. Die Erklärung ist jedoch die große Anzahl an Assoziationsbeziehungen in AGROVOC, die häufig „versteckte" Hierarchiebeziehungen sind, die bei der Berechnung der Höhe nicht berücksichtigt werden, jedoch zusätzlich von den Integrationsverfahren beachtet werden sollten. Die verhältnismäßig kleinen Werte für die Höhenvarianz geben keine Hinweise darauf, dass eine größere Anzahl herausragend tiefer Hierarchien existiert, die gesondert zu untersuchen wären.

9.3.4 Zusammenfassung der Ergebnisse

Die exemplarische Analyse der Thesauri AGROVOC und GEMET hat die große Heterogenität von Thesauri aufgezeigt, die diese besitzen können, selbst wenn die Konformität mit unserem Thesaurusmodell gegeben ist.

GEMET konnte als eher allgemeiner Thesaurus mit wörterbuchartigem Charakter (allgemeiner definitionswörterbuchartiger Thesaurus), AGROVOC hingegen als Fachthesaurus mit spezieller Retrieval-Unterstützung (Fach-Retrieval-Thesaurus) identifiziert werden.

Eine detailliertere Zusammenfassung der Ergebnisse unter der Prämisse, eher eine wörterbuchartige als eine retrieval-optimierte Föderation anzustreben, zeigt Tabelle 9.4.

Analyseergebnis	Auswirkung (Verfahren)	Erwartung (Ergebnis)
A großer, G normaler Anteil lateinischer Fachtermini	falls Integration erwünscht, besondere Algorithmen/Fachwörterbücher erforderlich	ohne besondere Algorithmen wenig Inter-Thesaurus-Beziehungen zu Fachtermini
A viele Quasy-Synonyme, G wenige, aber strenge Synonyme	Synonyme bei linguistischen Verfahren unterschiedlich bewerten	Reduktion des Begriffsumfangs von A-Begriffen bei Integration
A mit Benutze-Kombination-Beziehung	Berücksichtigung von Nicht-Deskriptoren, die in Benutze-Kombination-Relation enthalten sind, erforderlich	

Fortsetzung auf der nächsten Seite ...

[6]Eine Durchsicht der 22 GEMET-Deskriptoren der Gruppe *acts* etwa zeigt, dass es sich ausschließlich um konkrete Verordnungen und Vereinbarungen handelt (z.B. *Basel Convention*), für die es in AGROVOC tatsächlich keine Entsprechungen gibt.

[7]Alle GEMET-Begriffe der Gruppe *acts* könnten unter den AGROVOC-Begriff *international agreements* zusammengefasst werden.

... Fortsetzung

Analyseergebnis	Auswirkung (Verfahren)	Erwartung (Ergebnis)
A und G präkoordiniert	Mehrwortanalysen erforderlich	durch Mehrwortanalyse werden zusätzliche Inter-Thesaurus-Hierarchiebeziehungen gefunden
G i.d.R. natürlichsprachige Definition vorhanden, A bei missverständlichen Benennungen Erläuterungen vorhanden	zusätzliche Hinweise für menschlichen Experten vorhanden; Algorithmen zur Auswertung von Definitionstexten möglich	
1896 identische Benennungen		mehr als 756 Inter-Thesaurus-Äquivalenzbeziehungen
In A deutlich bessere Konnektivität als in G		Teil der nicht verbundenen G-Deskriptoren kann verbunden werden
in A bessere Zugänglichkeit, viele Assoziationsbeziehungen, in G kaum Assoziationsbeziehungen	Assoziationsbeziehungen in A gesondert untersuchen (versteckte Hierarchien)	Zugänglichkeit wird gegenüber dem Wert von G durch die Assoziationsrelationen in A verbessert; Assoziationsbeziehungen von A werden häufig Bestandteil von Konfliktmarkierungen (insbes. Beziehungstypdifferenzen)
in A sehr viele, in G kaum Bestandsbeziehungen; in A wesentlich mehr polyhierarchische Begriffe	Graphen nur schlecht vergleichbar (die ähnliche Höhe der Graphen ist, da Abstraktions- und Bestandsrelation gemeinsam betrachtet werden, kein Indikator für das Gegenteil); besondere Sorgfalt bei Integration polyhierarchischer Begriffe erforderlich	
A spezialisierter als G	Algorithmen zum Auffinden von Hierarchiebeziehungen erforderlich	viele A-Deskriptoren können nur über (direkte oder indirekte) Oberbegriffsbeziehungen verbunden werden
eine große und wenige kleine isolierte Strukturen in A		kleine isolierte A-Strukturen können nicht integriert werden; aufgrund der Assoziationsbeziehungen in A wird durch die Begriffsintegration die Gesamtzahl der isolierten Strukturen deutlich abnehmen
Erläuterung: A = AGROVOC G = GEMET		

Tabelle 9.4: Übersicht über die Analyseergebnisse und deren Interpretation

9.4 Resümee

Wir haben in diesem Kapitel gezeigt, dass es möglich ist, anhand der Interpretation einfacher berechenbarer Kennzahlen sowie des stichprobenartigen Untersuchens von Besonderheiten wesentliche inhaltliche Eigenschaften von Thesauri mit geringem menschlichen Aufwand zu erkennen. Anhand der erkannten Eigenschaften konnten wichtige Folgerungen sowohl über benötigte

Verfahren der Begriffsintegration und deren Einsatz als auch hinsichtlich der Erwartungen an das Ergebnis der Begriffsintegration für konkret zu integrierende Thesauri hergeleitet werden. Dieses Resultat kann in die Auswahl und Konfiguration der Verfahren ebenso einfließen wie in eine Bewertung der Thesaurusföderation.

Kapitel 10

Integrationsstrategie

Mit der in Kapitel 7 entwickelten Architektur haben wir den Grundstein gelegt, um die Problemlösungsstrategie von den Problemlösungsverfahren zu trennen. Das in Kapitel 4 vorgestellte Vorgehensmodell gibt der Problemlösungsstrategie bereits einen Rahmen. Ziel der in diesem Kapitel näher betrachteten Integrationsstrategie ist es, die Lösungsstrategie und somit das Vorgehensmodell für die komplexeste der Phasen, die Realisierungsphase, zu verfeinern (zur Einordnung vgl. Abbildung 10.1).

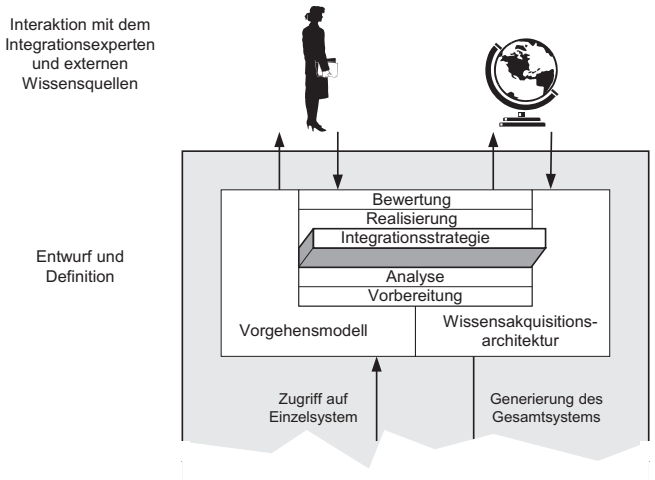

Abbildung 10.1: Die Phase der Integrationsstrategie-Festlegung detailliert das Vorgehensmodell für die komplexe Realisierungsphase

In Abschnitt 10.1 werden die Ziele dieser Integrationsstrategie detaillierter dargestellt. Anhand dieser Ziele wird in den Abschnitten 10.2 bis 10.3 top-down die Integrationsstrategie erarbeitet. Dieses Vorgehen ermöglicht eine effiziente Beherrschung der Komplexität der Strategie. Es ist angelehnt an die zu diesem Zweck in der Workflow-Forschung eingeführten vertikale Hierarchisierung (auch: Dekomposition) eines Workflow-Schemas in Form von referenzierten Subworkflows (vgl. [JBS97, S. 163]).

10.1 Ziele der Integrationsstrategie

Bei uns bekannten Ansätzen der Begriffsintegration ist die Problemlösungsstrategie häufig nur implizit gegeben[1] und immer fest vorgegeben. Die Vielzahl der in diesen Ansätzen verwendeten Problemlösungsstrategien zeigt aber, dass nicht von einer einzigen Strategie optimale Ergebnisse für verschiedene Integrationsfälle erwartet werden dürfen. Durch eine explizite und veränderteten Randbedingungen angepasste Problemlösungsstrategie ergibt sich folgender maßgeblicher Vorteil: Der menschliche Integrator aber auch maschinelle Planungsexperten können ihr während vorangegangener und der aktuellen Integrationen erworbenes Wissen zusätzlich in die Lösungsstrategie für das aktuelle Problem einbringen, ohne dass Neu- oder Weiterentwicklungen der Software-Infrastruktur erforderlich werden. Durch die Anlehnung an das Workflow-Prozessdefinitionsmodell (entsprechend der Aufgaben-Agenda, vgl. Abschnitt 7.3.2.3, S. 120) kann dies auf abstrakter Ebene geschehen, ohne dass Programmierkenntnisse erforderlich sind.

Wird das in Kapitel 4 entwickelte Vorgehensmodell als Rahmen genommen, könnte eine detailliertere Strategie für jede Phase vorgenommen werden. Wir beschränken uns aber auf eine Strategie für die komplexeste dieser Phasen, die Realisierungsphase, da die Prozesse innerhalb der übrigen Phasen als weitgehend fix angenommen werden (vgl. die entsprechenden Kapitel 8 und 9 sowie Abschnitt 12.3).

Mit der Entwicklung einer Problemlösungsstrategie für die Realisierungsphase werden folgende Ziele verfolgt:

Allgemeingültige Strategie: Eine allgemeingültige Strategie ist zu entwickeln, die als Grundlage für die Integration aller Thesauri dient. Diese Strategie zerlegt die komplexe Realisierungsphase in eine Folge von angestrebten Teilzielen. Die Entwicklung dieser allgemeingültigen Strategie kann als intellektuelle Leistung betrachtet werden, die im Wesentlichen nicht von maschinellen Experten erbracht werden kann. Stattdessen wird sie von uns im Rahmen dieser Arbeit erarbeitet und kann von menschlichen Integrationsexperten modifiziert werden.

Erstellung der Aufgaben-Agenda: Die in der allgemeingültigen Strategie spezifizierten Teilziele müssen weiter verfeinert und an eine konkret vorliegende Situation angepasst werden. Schließlich soll das Ergebnis so weit verfeinert sein, dass es eine Abfolgen von Aufgaben identifiziert, die ausgeführt werden sollen, um die Teilziele zu erreichen. Diese Folgen von Aufgaben werden der Realisierungsphase über die Aufgaben-Agenda des Blackboards bekannt gemacht (vgl. Abschnitt 7.3.2.3, S. 120).

Das Herleiten der Aufgaben-Agenda anhand der allgemeingültigen Strategie sowie der vorliegenden Situation soll semi-automatisch geschehen, d.h. von menschlichen Integrationsexperten mit maschineller Unterstützung.

Modifikation während der Ausführung: Eine Integrationsstrategie soll auch während ihrer Ausführung modifiziert werden können, um Erkenntnisse aus Zwischenergebnissen und Zwischenbewertungen einfließen zu lassen. Menschliche und maschinelle Experten sollen die Möglichkeit von Modifikationen erhalten.

In den folgenden Abschnitten werden die von uns entwickelten Teillösungen zur Erreichung der jeweiligen Ziele vorgestellt.

[1] Selbst falls eine Problemlösungsstrategie explizit aufgeführt wird, findet dies nur unvollständig und auf hohem Abstraktionsgrad statt.

10.2 Spezifikation einer allgemeingültigen Integrationsstrategie

Um zu einer allgemeingültigen Integrationsstrategie zu gelangen, müssen eine Reihe von Entscheidungen getroffen und Differenzierungen gefunden werden. Basis für diese Entscheidungen sind die Analyseergebnisse existierender Ansätze und die Verallgemeinerung der daraus gewonnen Erkenntnisse zu einem verfeinerten Vorgehensmodell für die Realisierungsphase.

Die allgemeingültige Strategie wird top-down entwickelt: Begonnen wird auf hoher Abtraktionsebene, die Schritt für Schritt verfeinert wird. Die Integrationsstrategie wird entlang dieser vertikalen Hierarchisierung in so genannte *Strategie-Ebenen* erläutert.

10.2.1 Strategie-Ebene 1: Top-Level-Integrationsstrategie

Die erste Entscheidung, die zu treffen ist, betrifft die Frage, ob alle zu einer Thesaurusföderation zu integrierenden Thesauri gleichzeitig oder nacheinander integriert werden sollen (vgl. [Con97, S. 74ff]). Diese Frage stellt sich in den bekannten Ansätzen nicht, da ausschließlich zwei Thesauri integriert werden. Wir wählen eine binäre Integrationsstrategie, da dies die Komplexität der einzelnen Integrationsschritte und somit der zu entwickelnden Problemlösungsverfahren verringert. Zudem können somit die Erkenntnisse der Ansätze für die Integration von zwei Thesauri direkt übertragen werden. Die aus dieser Entscheidung resultierende Top-Level-Integrationsstrategie wird in Abbildung 10.2 dargestellt. In der Workflow-Terminologie wird diese Top-Level-Integrationsstrategie auch Gesamtprozess genannt, der weiter in Teilprozesse oder Aktivitäten verfeinert wird.

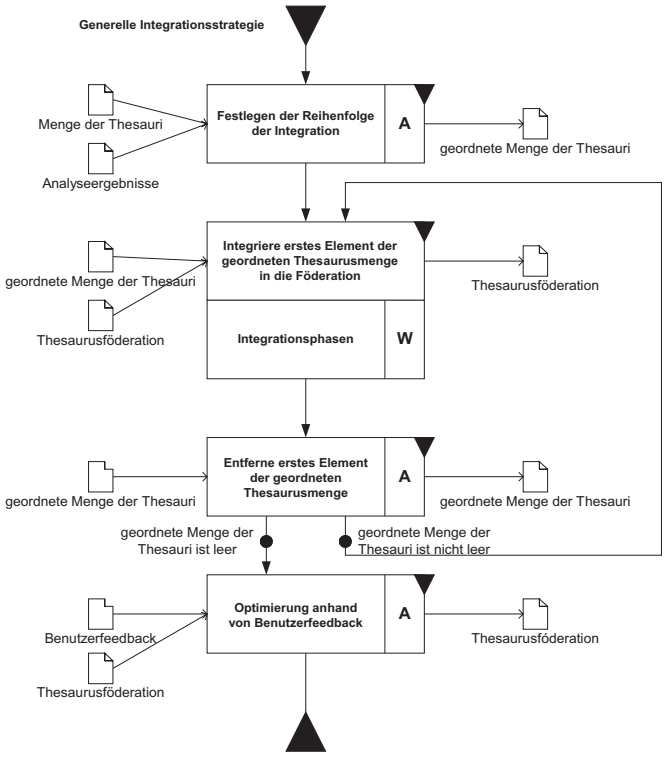

Abbildung 10.2: Strategie-Ebene 1: Top-Level-Integrationsstrategie (A bezeichnet Aktionen, W Subworkflows, die im Rahmen dieser Arbeit unter dem links neben dem W stehenden Titel näher spezifiziert werden)

Für diese binäre Top-Level-Integrationsstrategie ist die Menge der Thesauri zu ordnen. Von einer Ordnung gemäß der Thesaurusqualität wird das beste Ergebnis erwartet, da

- die Integration „guter" Thesauri bewährte Strukturen verstärken und somit als Basis für den weiteren Integrationsvorgang dienen kann,
- „schlechtere" Thesauri einfacher in bestehende „gute" Strukturen integriert werden können.

Eine einzelne Kennzahl für eine allgemeingültige Thesaurusqualität kann es aber nicht geben (vgl. Kapitel 9). Von Thesauri, die einen guten Ausgangspunkt für die Integration darstellen, erwarten wir eine eher allgemeinere als spezifischere thematische Abdeckung, gut definierte und klar abgegrenzte Begriffe und einen eher größeren Vokabularumfang, der auf eine gewisse Relevanz schließen lässt. Um Thesauri entsprechend dieser Kriterien zu bewerten, wird auf die in den Abschnitten 9.2.2 bis 9.2.4 dargestellten Kennzahlen zurückgegriffen:

Gruppengrößenvarianz (GGV): Eine geringere Varianz wird als Indikator für eine gleichmäßigere und somit allgemeinere thematische Abdeckung des Vokabulars gewertet.

Abstraktionsbeziehungen pro Deskriptor (ApD), Definitionsanteil (DA): Indikatoren für gut definierte und klar abgegrenzte Begriffe sind eine größere durchschnittliche Anzahl an Abstraktionsbeziehungen pro Deskriptor sowie ein hoher Definitionsanteil.[2]

Anzahl Deskriptoren ohne Fremdwörter (ADoF): Die Größe des Vokabulars wird durch die Anzahl der Deskriptoren definiert. Jedoch betrachten wir Fremdwörter, die im Allgemeinen Ausdruck eines speziellen Fachvokabulars sind und somit im Widerspruch zu der geforderten Allgemeingültigkeit stehen, nicht. Bezeichnet $\mathrm{FWA}(\theta.D)$ den Frendwörteranteil eines Thesaurus θ so berechnet sich ADoF aus $(1 - \mathrm{FWA}(\theta.D)) \times |\theta.D|$.

Anhand dieser Kennzahlen kann die für die Ordnung der Thesauri erforderliche Qualitätskennzahl r_θ für einen Thesaurus θ wie folgt definiert werden:

$$r_\theta = \frac{\min_{i=1}^{n}(\mathrm{GGV}(\theta_i))}{\mathrm{GGV}(\theta)} \times w_1 + \frac{\mathrm{ApD}(\theta)}{\max_{i=1}^{n}(\mathrm{ApD}(\theta_i))} \times w_2 + \mathrm{DA}(\theta) \times w_3 + \frac{ADoF}{\max_{i=1}^{n}(ADoF)} \times w_4$$

wobei n die Anzahl der zu integrierenden Thesauri ist und w_1, w_2, w_3, w_4 Gewichte sind, deren Summe 1 ergibt. Aufgrund der Normierung der einzelnen Summanden gilt $0 \leq r_\theta \leq 1$.

Je größer r_θ ist, desto früher soll ein Thesaurus integriert werden.

Beispiel 10.1 *Werden die Gewichte mit* $w_1 = w_2 = w_3 = w_4 = 0.25$ *gesetzt, womit den einzelnen Komponenten identische Bedeutung gegeben wird, ergibt sich bei einem Vergleich von GEMET und AGROVOC* $r_{AGROVOC} = 0.37$ *und* $r_{GEMET} = 0.89$. *GEMET wird also als der Thesaurus gewertet, der die Ausgangsbasis für die Integration ist. AGROVOC wird nach GEMET in die Föderation integriert.*

[2] Weiterer Indikator für die Güte der Begriffsdefinition könnte das Äquivalenzverhältnis sein. Jedoch sollten dann Quasi-Synonyme nicht berücksichtigt werden. Diese können aber nicht automatisch von strengen Synonymen unterschieden werden. Das Äquivalenzverhältnis wird daher nicht berücksichtigt.

Die Freiheitsgrade sind offensichtlich: Sollte der menschliche Experte eine von dieser berechneten Ordnung abweichende Vorstellung über die Integrationsreihenfolge haben, kann er die vorgeschlagene Ordnung überstimmen oder eine neue Formel für die Qualitätskennzahl angeben.

Sind alle Thesauri integriert, kann die Föderation aus der Entwicklungsphase in die Betriebsphase gehen. Während dieser Betriebsphase interagieren die Endnutzer mit der Föderation und haben über Rückkopplungswerkzeuge ggf. die Möglichkeit, Verknüpfungen innerhalb der Föderation zu bewerten oder neue Verknüpfungen vorzuschlagen. Diese Optimierung während des Betriebes soll aber nicht weiter betrachtet werden und sei hier nur der Vollständigkeit der Integrationsstrategie wegen aufgeführt.

Eine Bevorzugung einer n-ären vor der binären Integrationsstrategie hat Auswirkungen auf die weitere Planung sowie die Problemlösungsverfahren, die dann angepasst werden müssten (statt jeweils einen Thesaurus in die Föderation zu integrieren, müssen mehrere Thesauri oder auch Föderationen gleichzeitig integriert werden können).

10.2.2 Strategie-Ebene 2: Teilphasen der Integration

Auf der zweiten Ebene unterscheiden wir zwischen verschiedenen Teilphasen: einer initialen Integration eines Thesaurus mit einer Föderation sowie einer Optimierung dieser Integration.

Grund für diese Unterscheidung sind die unterschiedlichen Informationsquellen, auf die in diesen Phasen zugegriffen werden kann. Bei der initialen Integration gibt es noch keinerlei Verknüpfungen zwischen dem neuen Thesaurus und der Föderation. Es können aussschließlich Informationen innerhalb des Thesaurus, innerhalb der bereits vorhandenen Föderation und aus externen Quellen herangezogen werden. Hingegen stehen nach der initialen Integration zusätzliche Informationen über Verknüpfungen zur Verfügung. Hier kann wiederum unterschieden werden zwischen Verknüpfungsinformationen anhand von Zwischenergebnissen sowie – nach einer Analyse dieses Zwischenergebnisses – Bewertungsinformationen[3] (vgl. Abbildung 10.3).

Somit werden insgesamt drei Teilphasen (*initiale Integration*, *Optimierung anhand von Zwischenergebnissen* und *Optimierung anhand einer Zwischenergebnisbewertung*) identifiziert. Die beiden Optimierungsteilphasen können mehrfach hintereinander ausgeführt werden, da nach der Optimierung anhand einer Zwischenergebnisbewertung neue Zwischenergebnisse vorliegen können.

Die Partitionierung der Begriffsintegration in diese Teilphasen besitzt folgenden Vorteil: Die Teilphasen mit den dargestellten Iterationen bilden die Ausgangsbasis für ein Vorgehen der schrittweisen Verbesserung des Integrationswissens. Einfache Verfahren – im Sinne einer möglichst vollständigen Automatisierung bei kurzen Ausführungszeiten selbst bei der Berücksichtigung großer Ausschnitte der Föderation – sollen möglichst in der Anfangsphase der Integration eingesetzt werden. Aus diesen Zwischenergebnissen können andere Verfahren wiederum neues Integrationswissen herleiten. Komplexere oder auch unsicherere Verfahren – im Sinne eines schlechteren Verhältnisses von guten zu schlechten Vorschlägen und somit eines größeren Maßes an erforderlicher Kontrolle durch den menschlichen Integrationsexperten – können dediziert zur Behebung von Schwachstellen der vorläufigen Integration, also insbesondere in der dritten Teilphase, eingesetzt werden.

Wie wir in Abschnitt 11.1 zeigen werden, lassen sich alle bekannten sowie die von uns ent-

[3] Die notwendigen Methoden zur Gewinnung dieser Bewertungsinformationen werden in Kapitel 12.3 entwickelt und hier nicht weiter betrachtet.

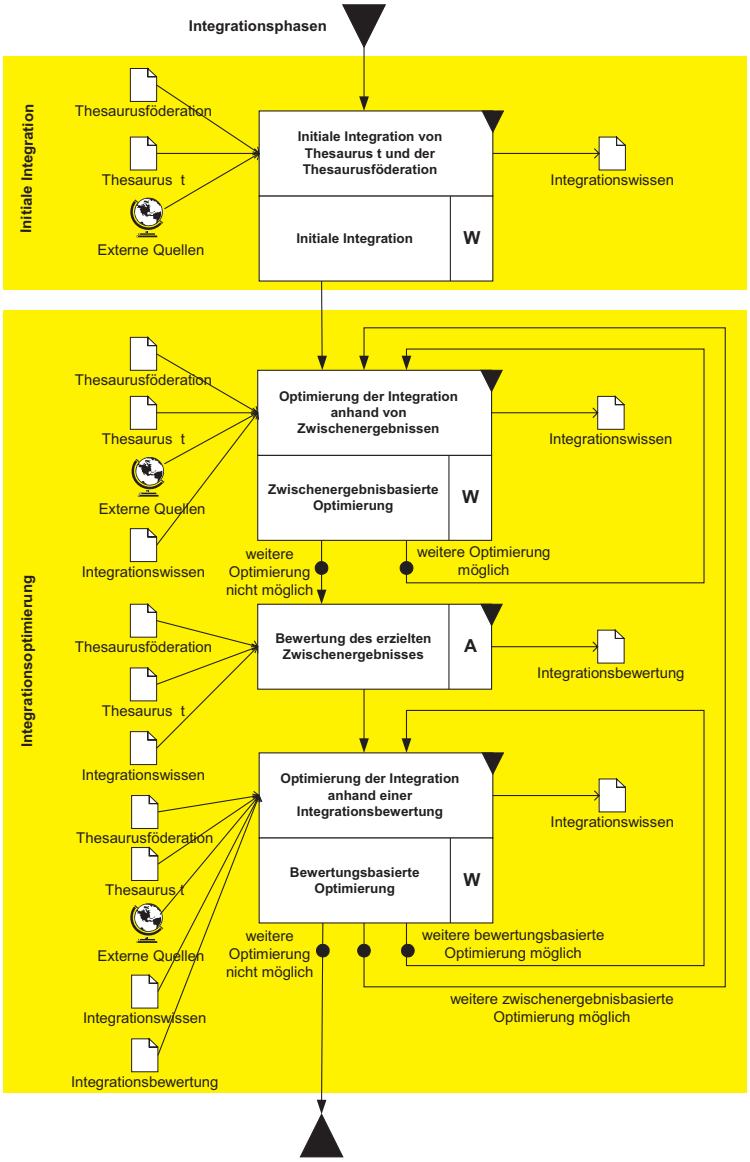

Abbildung 10.3: Strategie-Ebene 2: Teilphasen der Integration

wickelten Verfahren der Begriffsintegration entsprechend diesen Teilphasen einordnen. Somit können die Teilphasen zugleich als Basis für ein Klassifikationsschema verwendet werden, das die Eingliederung neuer Verfahren vereinfacht.

10.2.3 Strategie-Ebene 3: Ablauf innerhalb der Teilphasen

Auf der dritten Ebene wird der Ablauf innerhalb der Teilphasen spezifiziert. Eine solche Spezifikation entspricht einer allgemeinen Aufgaben-Agenda, die noch unabhängig von den tatsächlich zur Verfügung stehenden Ressourcen (wie maschinellen Integrationsexperten, aber auch den zu integrierenden Komponententhesauri) ist. Wir nennen diese Aufgaben-Agenda *vollständige Aufgaben-Agenda*, da sie alle möglichen Aufgaben enthält. Die vollständige Aufgaben-Agenda wird durch Anpassung an Zwischenergebnisse und zur Verfügung stehende Ressourcen um Aufgaben reduziert, Ergebnis ist die *adaptierte Aufgaben-Agenda*.

10.2.3.1 Initiale Integration

Erste näher zu spezifizierende Teilphase ist die initiale Integration (vgl. Abbildung 10.4). Wie bereits erwähnt, stehen bei der initialen Integration noch keinerlei Verknüpfungsinformationen zur Verfügung. Zielsetzung ist also das Herstellen erster Verbindungen zwischen der vorhandenen Thesaurusföderation (die im ersten Integrationsdurchlauf nur aus einem Thesaurus besteht) sowie dem zu integrierenden Komponententhesaurus. Diese Erstintegration kann anschließend in den weiteren Teilphasen iterativ verbessert werden.

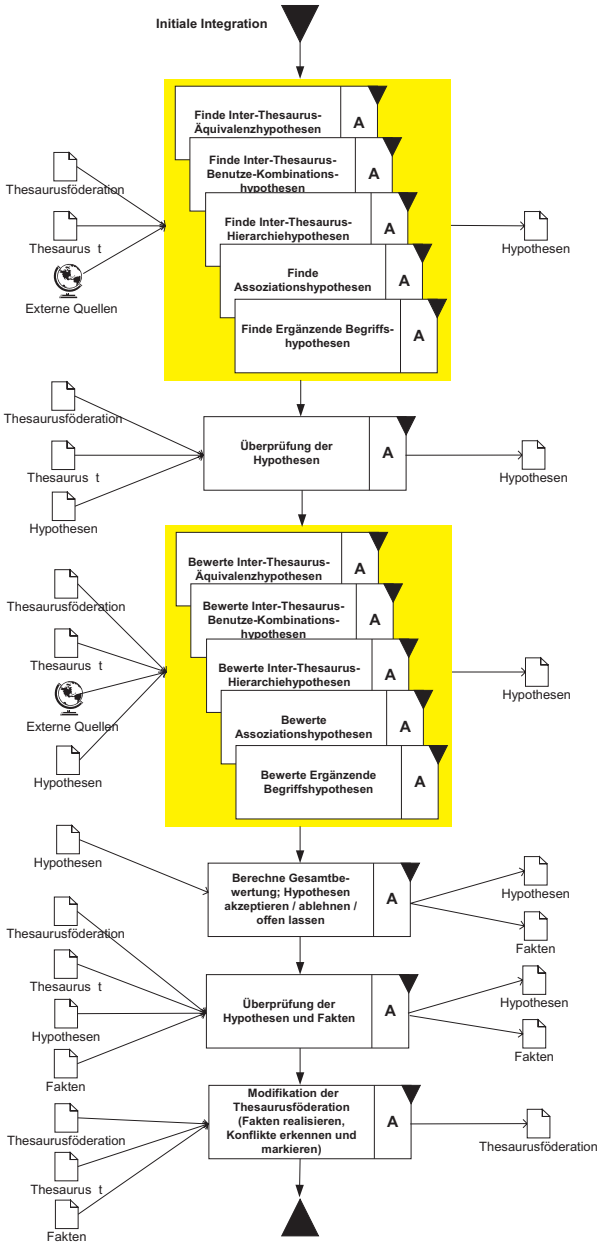

Abbildung 10.4: Strategie-Ebene 3: Initiale Integration

Um eine möglichst umfassende Basis für weitere Integrationsteilphasen bereitzustellen, soll die initiale Integration möglichst alle Arten von Verknüpfungspunkten finden. Entsprechend der von uns in Abschnitt 7.3 entwickelten Architektur bedeutet dies, Hypothesen über das Einfügen

von Inter-Thesaurus-Beziehungen (Äquivalenz-, Benutze-Kombination-, Hierarchie-, und Assoziationsbeziehungen) sowie Ergänzende Begriffe zu finden und zu bewerten. Das Finden solcher Hypothesen kann für die unterschiedlichen Hypothesentypen dabei parallel geschehen.

Nachdem die Ergebnisse aller parallel ausgeführten Arbeitsschritte zum Finden von Hypothesen vorliegen, können zu einer Bewertung dieser Ergebnisse zusätzlich die Ergebnisse selbst herangezogen werden. Wiederum können die unterschiedlichen Bewertungen parallel ausgeführt werden. Auf eine weitere Berücksichtigung der neu bewerteten Ergebnisse kann innerhalb dieser Teilphase verzichtet werden, da diese in den weiteren Teilphasen berücksichtigt werden.

Liegen die Bewertungen aller Hypothesentypen vor, kann die Gesamtbewertung für jede Hypothese berechnet werden und die Hypothese akzeptiert, abgelehnt oder offen gelassen werden. Akzeptierte Hypothesen (positive Fakten) werden schließlich zur Erstellung der initialen Integration des Komponententhesaurus verwendet. Dabei müssen ggf. entstehende Konflikte erkannt und markiert werden.

In Abbildung 10.4 sind bereits qualitative Überprüfungen der Hypothesen und Fakten vorgesehen, um Hypothesen bzw. Fakten, die den Qualitätskriterien bei der Erstellung einer Thesaurusföderation nicht entsprechen, herauszufiltern. Diese Thematik wird in Abschnitt 11.2.2 ausführlich behandelt.

10.2.3.2 Zwischenergebnisbasierte Optimierung

Sollten nach der initialen Integration keinerlei Verbindungen zwischen dem zu integrierenden Thesaurus und der Thesaurusföderation hergestellt sein, gilt eine Integration als wenig aussichtsreich. Der Vorgang kann abgebrochen werden. Falls mindestens ein weiterer Thesaurus integriert werden soll, kann ein neuer Integrationsversuch nach Integration aller weiteren Thesauri unternommen werden. Andernfalls wird der Integrationsexperte auf die erfolglose initiale Integration hingewiesen und kann über das weitere Vorgehen entscheiden.

In der Regel – und nur diesen Fall haben wir in der grafischen Darstellung des Integrationsvorgangs berücksichtigt – stehen nach der initialen Integration jedoch eine Reihe von Verbindungen als Basis für weitere Verbesserungen bereit. Die zwischenergebnisbasierte Optimierung unterscheidet sich also von der initialen Integration insofern, als dass bereits für das Finden von Hypothesen zusätzlich auf Fakten, die zu einer modifizierten Thesaurusföderation geführt haben, sowie auf Hypothesen zurückgegriffen werden kann. Somit können weitere Verfahren zum Finden von Hypothesen, die diese Informationen auswerten, eingesetzt werden. Der prinzipielle Ablauf entspricht jedoch dem der initialen Integration, wir verzichten daher auf eine erneute Abbildung.

Nach einer Iteration der zwischenergebnisbasierten Optimierung gilt es zu entscheiden, ob eine weitere Iteration durchzuführen ist oder nicht. Als Entscheidungskriterium kann z.B. die Anzahl neuer Hypothesen oder die Anzahl neuer Fakten (jeweils absolut oder relativ zur Anzahl der vorangegangenen Iteration(en)) herangezogen werden. Dieses Entscheidungskriterium kann vom menschlichen Integrationsexperten durch eine entsprechende Formel vorgegeben werden. Der menschliche Integrationsexperte sollte zusätzlich die Möglichkeit haben, eine weitere Iteration zu verlangen oder zu verhindern, auch wenn aufgrund der vorgegebenen Formel eine andere Entscheidung gefallen wäre.

10.2.3.3 Bewertungsbasierte Optimierung

Nachdem die Teilphase der zwischenergebnisbasierten Optimierung abgeschlossen ist, erfolgt, wie bereits in Abbildung 10.3 dargestellt, eine Bewertung des aktuellen Standes der Integration (vgl. ausführliche Darstellung in Kapitel 12). Auf diese Ergebnisse setzt die bewertungsbasierte Optimierung auf, die gezielt dort zu einer Verbesserung der Integration führen soll, wo die Bewertung Schwachstellen erkannt hat.

Die zur Verfügung stehenden Ergebnisse der Integrationsbewertung unterscheiden die bewertungsbasierte Teilphase grundlegend von den vorangegangenen Teilphasen der Integration. Dies verdeutlicht Abbildung 10.5, indem zum einen weitere Eingaben in Form von Bewertungen und Hypothesenzielen zur Verfügung stehen und zum anderen das Auffinden und Bewerten von Hypothesen *gezielt* stattfinden soll.

Wie solche Bewertungen aussehen wird detailliert in Kapitel 12 dargestellt. Es gilt jedoch festzuhalten, welche Auswirkungen derartige Bewertungen auf die Optimierung der Integration haben sollen. Konkret bedeutet dies, welche *Hypothesenziele* verfolgt werden sollen. Solche Hypothesenziele können etwa das Reduzieren oder Vergrößern der Anzahl der Inter-Thesaurus-Verbindungen eines bestimmten Typs innerhalb eines Teilgraphen der Föderation sein. Somit beinhaltet ein Hypothesenziel folgende Informationen:

Polarität: Die Polarität ist ein Indikator, der aussagt, ob positive oder negative Hypothesen gefunden werden sollen, also, ob die Anzahl der Inter-Thesaurus-Beziehungen bzw. Ergänzenden Begriffe vergrößert oder verkleinert werden soll.

Quantität: Ein optionaler Bestandteil ist die Aussage über die Quantität in Form einer unteren und oberen Schranke für die optimale Anzahl der Fakten, die aus den Hypothesen generiert werden soll.

Hypothesentyp: Der Hypothesentyp legt fest, nach welchem Typ von Hypothese gesucht werden soll.

Relevanter Teilgraph: Durch die Angabe eines Ausgangsknotens, der zu verfolgenden Beziehungstypen und der maximalen Pfadlänge oder optional einer Menge von Kanten kann der Teilgraph spezifiziert werden, innerhalb dessen weitere Hypothesen aufgestellt werden sollen.

Ursache: Informationen über die Ursache des Hypothesenziels (Woher kommt das Ziel?) sind insbesondere für eine Nachvollziehbarkeit erforderlich.

Menschlicher Integrationsexperte: Falls bereits festgestellt wurde, dass zum Erreichen des Hypothesenziels ein Eingriff des menschlichen Integrationsexperten erforderlich ist, kann dies entsprechend vermerkt werden.

Priorität: Die Dringlichkeit der Bearbeitung der Hypothesenziele kann durch eine Priorisierung ausgedrückt werden.

Liegen derartige Hypothesenziele vor, können gezielt entsprechende Hypothesen gesucht bzw. vorliegende Hypothesen bewertet werden.

Um aus der Integrationsbewertung die konkreten Hypothesenziele zu bestimmen, wird ein eigener Arbeitsschritt vorgesehen. Dies hat den Vorteil, dass es für die Verfahren zum Finden und Bewerten von Hypothesen ausreichend ist, die Hypothesenziele auswerten zu können. Eine Interpretation der Integrationsbewertung durch die Verfahren ist somit nicht erforderlich.

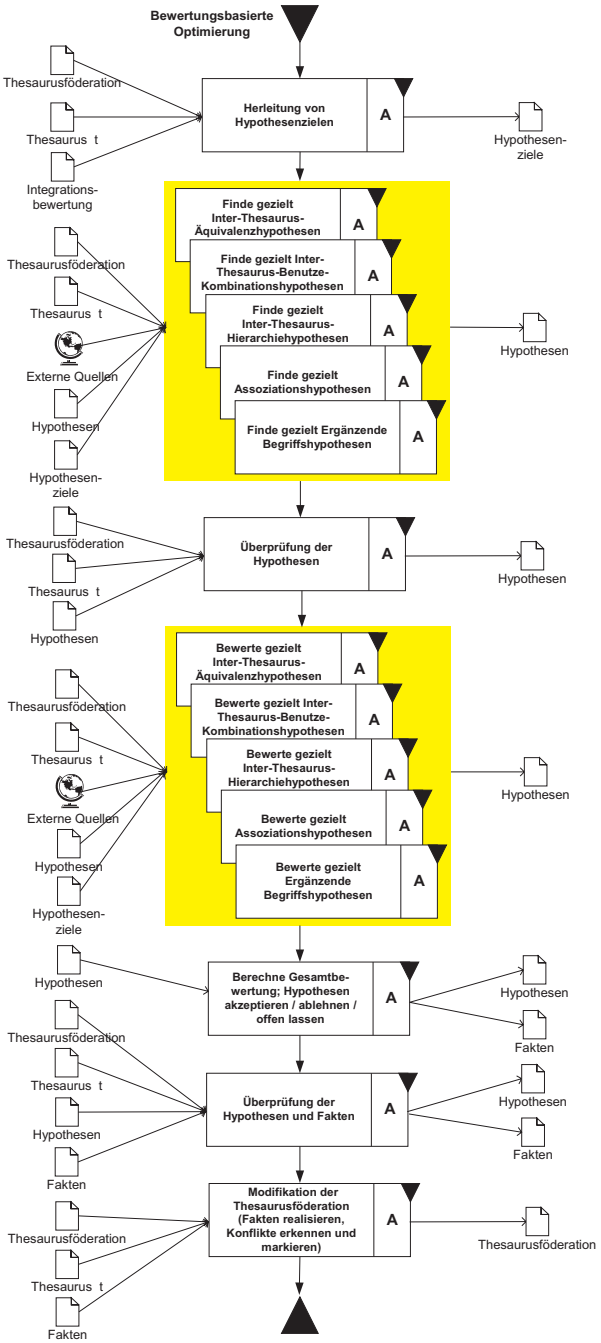

Abbildung 10.5: Integrations-Ebene 3: Bewertungsbasierte Integration

10.3 Modifikationen der Strategie und der Aufgaben-Agenda

Mit der Spezifikation aller drei Teilphasen der Ebene 3 wurde die vollständige Strategie als Aufgaben-Agenda definiert. In diesen Vorgegehensplan kann der menschliche Integrationsexperte zu jedem Zeitpunkt eingreifen, um ihn zu modifizieren. Aufgrund unserer an die Workflow-System-Architektur angelehnten Interpretation der Strategie durch den Steueragenten werden alle Änderungen sofort wirksam. Alle Modifikationen (z.B. zum Einfügen neuer Aufgaben oder Verzweigungen oder der sequenziellen statt parallelen Ausführung von Bewertungungen) müssen die Ausführung des Plans jedoch weiterhin terminieren lassen.

Die zentralen Aufgaben innerhalb der vollständigen Aufgaben-Agenda sind solche zum Finden und Bewerten von Hypothesen. Schwerpunkt der Reduktion zur adaptierten Aufgaben-Agenda bilden also diese Aufgaben. Die Reduktion kann sowohl durch den menschlichen Experten als auch durch einen Planungsagenten vorgenommen werden. Wir unterscheiden folgende Fälle:

Reduktion aufgrund verfügbarer Experten: Die Ressourcen zur Lösung von Aufgaben sind in unserer Architektur die Experten, die die Integrationsverfahren implementieren (vgl. Abschnitt 7.3, S. 115). Es gilt also festzustellen, ob entsprechende Experten zur Verfügung stehen. Die Überprüfung der verfügbaren maschinellen Experten soll automatisch geschehen, um den menschlichen Experten nicht zusätzlich zu belasten. Dazu ist zu jeder Aufgabe anzugeben, welche Ein- und Ausgaben erwartet werden und zu jedem Experten, welche Eingaben erwartet und welche Ausgaben produziert werden[4] (vgl. Anhang B). Sofern mindestens ein Agent unter Benutzung einer Teilmenge aller Eingaben alle erwarteten Ausgaben erzeugen kann, gilt die Aufgabe als prinzipiell erfüllbar. Ist dies nicht der Fall, kann die Aufgabe noch immer durch den menschlichen Integrator gelöst werden. Ob dieser sich dazu bereit erklärt, soll jedoch von ihm erfragt werden. Bei ablehnender Antwort muss die Aufgabe aus der Agenda entfernt werden.

Reduktion aufgrund bisheriger Ergebnisse: Es kann auch bei prinzipieller Erfüllbarkeit einer Aufgabe vorkommen, dass von einem oder mehreren Experten in einer spezifischen Situation keine Ausgaben produziert werden. Stehen nach Ausführung aller parallelen Aufgaben keinerlei entsprechende Aufgaben in Form von Hypothesen auf dem Blackboard oder ist das Blackboard ingesamt nicht modifiziert, wird der menschliche Integrationsexperte explizit aufgefordert entsprechende Ausgaben zu erzeugen[5]. Lehnt dieser ab, muss eine Überarbeitung des weiteren Planes erfolgen. Das bedeutet, dass alle Aufgaben, die unbedingt entsprechende Eingaben erwarten, entfernt werden, bis wiederum eine Aufgabe solche Ausgaben produzieren kann. Als Beispiel sei die Aufgabe *Bewerte Assoziationshypothesen* genannt, die in einer Iteration der zwischenergebnisbasierten Optimierung entfernt wird, wenn zuvor keine Assoziationshypothesen auf dem Blackboard sind oder das Blackboard seit der letzten Durchführung dieser Aufgabe nicht modifiziert wurde.

Reduktion aufgrund der Hypothesenziele: Sind Hypothesenziele Eingaben für Aufgaben, können alle Aufgaben, die nichts zu den Hypothesenzielen beitragen, ebenfalls solange entfernt werden, bis neue Hypothesenziele erzeugt werden, zu deren Erfüllung die Aufgaben

[4] Es sei darauf hingewiesen, dass die Abbildungen 10.4 und 10.5 hinsichtlich der Ausgaben der Finde- und Bewerte-Aufgaben vereinfacht sind. Etwa erzeugt die Bearbeitung einer Aufgabe *Finde Inter-Thesaurus-Äquivalenzhypothesen* als Ausgabe eine Menge von Äquivalenhypothesen, die in der Abbildung zu Hypothesen verallgemeinert sind.

[5] Diese Aufforderung des menschlichen Experten kann während der initialen Integration und der zwischenergebnisbasierten Optimierung unterbleiben, sollte aber bei der bewertungsbasierten Optimierung erfolgen, um die Hypothesenziele möglichst zu erreichen.

beitragen können. Ob eine Aufgabe zu einem Hypothesenziel beiträgt, kann festgestellt werden, indem überprüft wird, ob der Hypothesentyp des Hypothesenziels von der Aufgabe als Ausgabe erzeugt wird.

Der Zeitpunkt des Eingriffs des Planungsagenten ist, wie oben ersichtlich wird, der Moment des Übergangs von einer Aufgabe (oder einer Menge parallel auszuführender Aufgaben) zu einer neuen Aufgabe (oder einer Menge parallel auszuführender Aufgaben).

Das tatsächliche Zuordnen und Starten von Agenten zur Erfüllung von Aufgaben ist nicht mehr Teil der Integrationsstrategie, sondern wird vom Steueragenten innerhalb der Realisierungsphase durchgeführt. Die Verzahnungen zwischen Integrationsstrategie und Realisierungsphase werden durch den Eingriff des Planungsagenten sowie durch mögliche Eingriffe des menschlichen Experten notwendig.

Weitere Strategiemodifikationen durch den Planungsagenten sind in unserer Arbeit nicht vorgesehen. Insbesondere können also solche Modifikationen, die zu einer Erweiterung der Strategie führen, nur durch den menschlichen Integrationsexperten durchgeführt werden. Die Architektur und Modularität unseres Systems erlaubt jedoch durchaus zukünftig auch die Verwendung mächtigerer Planungsagenten.

10.4 Resümee

Das Ergebnis dieses Kapitels ist eine Lösungsstrategie in Form einer detailierten Prozessdefinition. Somit wird die Begriffsintegration erstmals explizit als Prozess dargestellt, der unabhängig von den eingesetzten Lösungsverfahren ist. Die Prozessdefinition in der vollständigen Form erfolgt als intellektuelle Leistung durch den menschlichen Experten, die Reduktion hinsichtlich einer konkreten Situation kann jedoch maschinell erfolgen. Eine solche Reduktion ist als Optimierung zu verstehen, mit der insbesondere das Laufzeitverhalten verbessert werden kann.

Aufgrund der Architektur können Eingriffe in den laufenden Prozess – unabhängig davon, ob sie maschinell oder durch den Menschen erfolgen – sofort berücksichtigt werden. Dies unterstützt eine rollierende Planung.

Die in diesem Kapitel von uns für den Planungsagenten vorgesehenen Aufgaben sind noch recht einfach. Wir haben damit die Umsetzbarkeit der Architektur hinsichtlich einer teilautomatisierten Planung gezeigt. Komplexere Planungsagenten z.B. auf Basis der Künstlichen Intelligenz sind möglich, ohne dass Änderungen an der Architektur erforderlich werden.

Kapitel 11

Realisierungsphase

Als Ergebnis der vorangegangenen Phase liegt eine Lösungsstrategie vor, die im Rahmen der Realisierungsphase zur Gewinnung und Festschreibung des Integrationswissens umgesetzt werden soll. Die Realisierungsphase als zentrale Phase setzt dazu auf den Ergebnissen sämtlicher vorangegangener Phasen (vgl. Kapitel 8 bis 10) auf (zur Einordnung vgl. Abbildung 11.1).

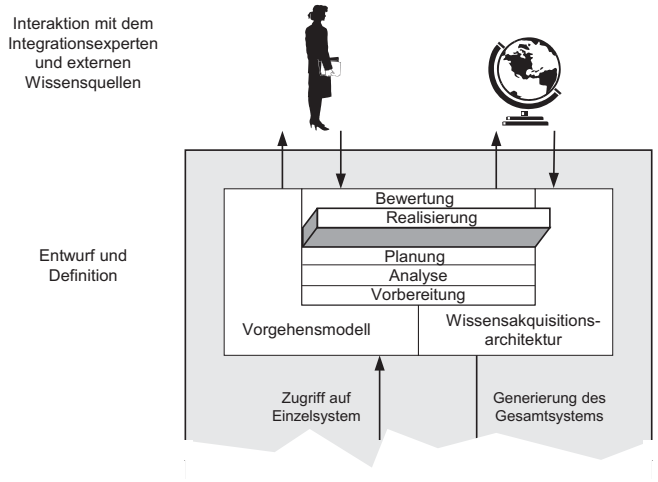

Abbildung 11.1: Die Realisierungsphase nutzt zur Gewinnung des Integrationswissens die Ergebnisse aller vorangegangener Phasen

Bei näherer Betrachtung der Realisierungsphase können folgende Teilaufgaben und Fragestellungen identifiziert werden, die in den folgenden Abschnitten detaillierter betrachtet werden:

Steuerung der Abarbeitung der Aufgaben-Agenda: Die Umsetzung der Problemlösungsstrategie erfolgt – wie in Abschnitt 7.3.6, S. 126 dargestellt – durch das Abarbeiten der Aufgaben-Agenda. Innerhalb der Realisierungsphase muss also die entsprechende Prozesssteuerung übernommen werden. Die Ausführungssteuerung im Sinne der Prozessablaufsteuerung übernimmt in unserer Architektur der Steueragent (vgl. Abschnitt 7.3.6, S. 126). Ein solcher Steueragent ist vergleichbar dem Workflow-Ausführungsdienst innerhalb von Workflowsystemen und von dort übernommen worden. Im Rahmen dieser Arbeit kann daher auf eine weitergehende allgemeine Betrachtung verzichtet werden.

Einbringen der Problemlösungsverfahren: Es ist weiterhin aufzuzeigen, wie wann welche Problemlösungsverfahren, die in unserer Architektur als Experten implementiert werden, möglichst optimal in die Strategie eingebracht werden können (vgl. Abschnitt 11.1).

Finden von Gesamtbeiträgen zur Lösungsverbesserung: Wird ein Problemlösungsverfahren separat betrachtet, liefert es einen Teilbeitrag zur Lösungsverbesserung. Aus den Teilbeiträgen ist ein Gesamtbeitrag herzuleiten (vgl. Abschnitt 11.2). Dies ist in unserer Architektur die Aufgabe des Moderators (vgl. Abschnitt 7.3.7). Der Moderator benötigt dazu sowohl Zugriff auf die Hypthesen und Fakten als auch auf die Föderation inklusive der abgelehnten Fakten. Ist eine Konfliktbereinigung bzw. -markierung erforderlich, kann der Moderator über entsprechende Aufgaben den menschlichen Experten mit Unterstützung des Benutzeragenten in den Prozess einbeziehen.

Einbringen von Verbesserungsbeiträgen: Schließlich gilt es, gefundene Gesamtbeiträge in die bisherige Lösung einzubringen und festzuhalten (vgl. Abschnitt 11.3). Dafür ist in unserer Architektur der Ausführungsagent verantwortlich.

11.1 Einbringen der Problemlösungsverfahren

Die eigentliche Gewinnung des Integrationswissens geschieht durch den Einsatz der Problemlösungsverfahren. In unserer Architektur ist die Aufgabe solcher Problemlösungsverfahren das Finden und Bewerten von Integrationsvorschlägen (vgl. Abschnitt 7.3, S. 115). Die Verfahren werden in Form von (maschinellen) Experten eingebracht.

Bereits in den Abschnitten 3.2.2 und 3.2.3 wurden die aus der Literatur bekannten Verfahren vorgestellt und hinsichtlich der zugrunde liegenden Methodik klassifiziert. An dieser Stelle entwickeln wir darüber hinaus Kriterien zum Zuordnen von Verfahren zu Teilphasen und Aufgaben, stellen exemplarisch die Einordnung einer Reihe von Verfahren dar und betrachten schließlich den Aspekt der Konfiguration der Verfahren.

11.1.1 Ordnungskriterien zur Einbringung der Verfahren

Im vorangegangenen Kapitel 10 wurde die Realisierungsphase in die drei Teilphasen Initiale Integration, Zwischenergebnisbasierte Optimierung und Bewertungsbasierte Optimierung zerlegt. In jeder dieser Teilphasen wiederum gibt es Aufgaben zum Finden und Aufgaben zum Bewerten von Hypothesen. Für alle Lösungsverfahren muss nun fesgelegt werden, in welcher dieser Teilphasen das jeweilige Verfahren zum Finden oder Bewerten herangezogen werden soll. Um diese Entscheidung treffen zu können, sind Informationen darüber erforderlich, was das entsprechende Verfahren an Voraussetzungen benötigt (Eingaben) und welche Ergebnisse erzeugt werden (Ausgaben). Diese Ein-/Ausgaben-Beschreibung ist in Anhang B dargestellt und wurde schon für die Registrierung der Experten beim Blackboard verwendet (vgl. Abschnitt 7.3.2.3).

Anhand dieser Informationen ist es bereits möglich zu beurteilen, *wozu* und *unter welchen Voraussetzungen* die Experten Lösungsbeiträge liefern. Auch auf die Frage, *wann*, d.h. in welcher Teilphase, der Einsatz eines Experten optimal ist, ermöglichen die Informationen über die Voraussetzungen erste Einschränkungen. Zu einer festen Zuordnung sind jedoch weitere Kriterien erforderlich. Wir betrachten dazu die entsprechenden Teilphasen:

Initiale Integration: Bei der Initialen Integration sind noch keine Inter-Thesaurus-Beziehungen etabliert und auch Hypothesen liegen noch nicht vor. Alle Verfahren, die in

dieser Teilphase eingesetzt werden können, dürfen daher *weder etablierte Inter-Thesaurus-Beziehungen noch bereits vorhandene Hypothesen*[1] *als Voraussetzung erfordern.*

Die Ergebnisse der Initialen Integration sind zudem Basis jeder weiteren Optimierung. Um eine Vielzahl von Korrekturen in späteren Schritten zu vermeiden, wird daher angestrebt, möglichst gute Fakten zu erzeugen (*zuverlässige Verfahren*) und bei Unsicherheiten die Hypothesen entsprechend offen zu lassen. Eine entsprechend *differenzierte Bewertung der Hypothesen* sollte z.B. über die Konfiguration der Verfahren möglich sein.

Die initiale Integration betrachtet die gesamten Begriffsmengen der bereits vorhandenen Föderation und des zu integrierenden Thesaurus. Aufgrund dieses umfangreichen Datenbestandes sowie einer angestrebten Minimierung des menschlichen Einsatzes sollen alle Verfahren innerhalb dieser Phase möglichst *keinerlei Interaktionen mit dem menschlichen Experten* – z.B. zum Auflösen von Unsicherheiten oder während der Verfahrensvorbereitung – erfordern. Sollen die Ressourcen *Zeit* und *Rechenleistung* möglichst wenig in Anspruch genommen werden, kann es zudem sinnvoll sein, besonders aufwendige Verfahren erst in späteren Teilphasen, innerhalb derer nur Ausschnitte der Begriffsnetze betrachtet werden, einzusetzen.

Zwischenergebnisbasierte Optimierung: Für die Zwischenergebnisbasierte Optimierung kann davon ausgegangen werden, dass bereits Hypothesen zuvor generiert und Fakten in Form von Inter-Thesaurus-Beziehungen, Ergänzenden Begriffen und Konfliktmarkierungen etabliert wurden. Verfahren, *die diese Informationen (Hypothesen, etablierte Fakten) nutzen*, um neue Hypothesen zu erzeugen oder existierende Hypothesen zu bewerten, sind für den Einsatz innerhalb dieser Teilphase prädestiniert. Es gilt wiederum, dass *möglichst selten auf den menschlichen Experten zurückgegriffen werden* sollte. Falls eine Einschränkung auf die Betrachtung von Teilbereichen des Begriffsnetzes gegeben ist – z.B. durch Betrachtung des näheren Graphen um eine Hypothese herum – dürfen diese Verfahren durchaus aufwendiger sein, da die betrachtete Datenmenge reduziert ist.

Bewertungsbasierte Optimierung: Verfahren, die innerhalb der Bewertungsbasierten Optimierung eingesetzt werden, sollen *gezielt auf bestimmten Teilgraphen operieren* können, da die Optimierung für diese Teilgraphen angestrebt wird. Da eine Optimierung als notwendig eingeschätzt wurde und üblicherweise kleinere Ausschnitte betrachtet werden, sind sowohl *unsichere* Verfahren als auch Verfahren, die einen *größeren menschlichen Aufwand* erfordern, möglich.

Eine Bewertung eines einzuordnenden Verfahrens erfordert genaue Kenntnisse des Verfahrens. Der Verfahrensersteller hat die entsprechenden Informationen daher mit dem Verfahren selbst bereitzustellen. Die Erkenntnisgewinne während einer Integration ermöglichen es dem menschlichen Experten zudem, diese Metainformationen zu ergänzen oder zu korrigieren. Die Zuordnung der Verfahren zu den verschiedenen Teilphasen kann bei vorhandenen Angaben zu den Teilphasen sowie dem Vorliegen aller genannten Informationen automatisch (durch den Steueragent) geschehen. Bei unzureichenden Informationen trifft der menschliche Experte die entsprechenden Entscheidungen.

Es soll zusätzlich erwähnt werden, dass bei der Bewertung von neu erzeugten Hypothesen alle Verfahren zur Bewertung von Hypothesen aus vorangegangenen Teilphasen auch in späteren Teilphasen eingesetzt werden können. Um jedoch neue Fakten zu erzeugen, ist den Bewertungen

[1] Beim Bewerten von Hypothesen ist das Vorhandensein der Hypothesen selbstverständliche Voraussetzung, die Einschränkung hinsichtlich vorhandener Hypothesen gilt daher nur für das Finden von Hypothesen.

ggf. ein anderes Gewicht zu geben. Insbesondere während der Bewertungsbasierten Optimierung kann eine größere Unsicherheit zugelassen werden.

11.1.2 Spezielle Verfahren und deren Einordnung

Es ist nicht das Ziel dieser Arbeit, vollkommen neue Integrationsverfahren zu entwicklen. Vielmehr wird die Möglichkeit geschaffen, eine Vielzahl von Verfahren in die Lösungsfindung einzubringen. In diesem Abschnitt belegen wir diese Möglichkeit, indem wir bereits aus der Literatur bekannte Verfahren aufgreifen, diese zum Teil optimieren, zugleich aber auch insbesondere für Bereiche, die in der Literatur bisher kaum beachtet wurden, Verfahrensvorschläge entwickeln. Diese werden anhand der im vorangegangenen Abschnitt vorgestellten Kriterien den Teilphasen zugeordnet. Schließlich zeigen wir das Zusammenspiel der Verfahren.

Nachdem wir in Kapitel 9 bereits eine Reihe von Maßen für die Analyse von Thesauri eingeführt haben, wird es in diesem Abschnitt notwendig, weitere Maße für die Bewertung von Hypothesen zu definieren. Wir führen diese Maße jeweils bei der Erläuterung der Verfahren ein, die für die Hypothesenbewertung auf die Maße zurückgreifen.

Lexikalische Gleichheit (Deskriptoren, Nicht-Deskriptoren, externe Quellen)

Eingabe: Thesaurusföderation, Thesaurus, externe Quellen (optional)

Ausgabe: positive Äquivalenzhypothesen

Beschreibung: Anhand lexikalischer Vergleiche werden identische Benennungen gesucht. Dieser Vergleich kann direkt zwischen den Benennungen (Deskriptoren/Nicht-Deskriptoren) der Thesaurusföderation bzw. des Thesaurus stattfinden. Des Weiteren ist es möglich, zu der Föderations-/Thesaurusbenennung auch in externen Quellen (z.B. WordNet) nach Synonymen zu suchen und diese in die lexikalischen Vergleiche miteinzubeziehen.

Bei lexikalischer Gleichheit wird eine Äquivalenzhypothese aufgrund übereinstimmender Buchstabenfolgen von Benennungen erzeugt. Die Bewertung der Hypothese erfolgt anhand der Typen der identischen Benennungen. Unterschieden wird zwischen lexikalischer Identität zwischen Deskriptoren, zwischen Deskriptoren und Nicht-Deskriptoren oder, falls ebenfalls externe Quellen berücksichtigt werden, zwischen Deskriptoren und Deskriptor-entsprechenden Benennungen in externer Quelle, zwischen Nicht-Deskriptor und Deskriptor-entsprechenden Benennungen in externer Quelle bzw. zwischen Nicht-Deskriptor und Nicht-Deskriptor-entsprechenden Benennung in externer Quelle.[2] Die Bewertungen sollen abhängig von der Zuverlässigkeit der Nicht-Deskriptoren in den Thesauri sowie der Benennungen in externen Quellen geschehen.

Wir haben somit das in der Literatur am Häufigsten verwendete Verfahren der Begriffsintegration aufgegriffen und mit den unterschiedlichen Bewertungen die Möglichkeit geschaffen, differenziertes Vertrauen in aufgrund unterschiedlicher Vergleichbasis gefundene Hypothesen auszudrücken.

Einordnung: Initiale Integration

Lexikalische Gleichheit zwischen Deskriptoren und BK-Nicht-Deskriptoren

Eingabe: Thesaurusföderation, Thesaurus (wobei in der Föderation und/oder dem Thesaurus BK-Nicht-Deskriptoren vorhanden sind), externe Quellen (optional)

[2] Eine detailliertere Unterscheidung in Abhängigkeit der externen Quelle oder der eindeutigen/nicht-eindeutigen Zuordnung in der externen Quelle ist möglich. Diese weitere Detaillierung ist verfahrensabhängig festzulegen.

Ausgabe: positive Benutze-Kombination-Hypothesen

Beschreibung: Wie Lexikalische Gleichheit (Deskriptoren, Nicht-Deskriptoren, externe Quellen), jedoch wird bei Gleichheit zwischen Deskriptor und BK-Nicht-Deskriptor oder deren Synonymen in externen Quellen eine Benutze-Kombination-Hypothese zwischen dem Deskriptor und den in der BK-Beziehung stehenden Deskriptoren erzeugt.

Diese gesonderte Berücksichtigung der BK-Nicht-Deskriptoren ermöglicht das direkte Auffinden von Benutze-Kombination-Hypothesen.

Einordnung: Initiale Integration

Lexikalische Gleichheit zwischen BK-Nicht-Deskriptoren

Eingabe: Thesaurusföderation, Thesaurus (wobei in der Föderation und dem Thesaurus BK-Nicht-Deskriptoren vorhanden sind), externe Quellen (optional)

Ausgabe: positive Ergänzender-Begriff-Hypothesen

Beschreibung: Wie Lexikalische Gleichheit (Deskriptoren, Nicht-Deskriptoren, externe Quellen), jedoch wird bei Gleichheit zwischen BK-Nicht-Deskriptor und BK-Nicht-Deskriptor oder deren Synonymen in externen Quellen eine Ergänzende-Begriff-Hypothese erzeugt. Da offensichtlich dem BK-Nicht-Deskriptor in verschiedenen Thesauri eine gewisse Bedeutung zukommt, wird mit der Ergänzender-Begriff-Hypothese vorgeschlagen, den entsprechenden Begriff als eigenständigen Deskriptor in die Thesaurusföderation aufzunehmen.

Einordnung: Initiale Integration

Lexikalische Ähnlichkeit

Eingabe: Thesaurusföderation, Thesaurus, Teilgraph, externe Quellen (optional)

Ausgabe: positive Äquivalenzhypothesen

Beschreibung: Wie Lexikalische Gleichheit (Deskriptoren, Nicht-Deskriptoren, externe Quellen), jedoch wird ein zu definierender lexikalischer Abstand zugelassen (z.B. das Vorhanden-/Nichtvorhandensein von Bindestrichen oder Leerzeichen, typische Unterschiede in Schreibvarianten (etwa hinsichtlich der Varianten in neuer und alter deutscher Rechtschreibung) bis hin zur Bestimmung etwa des Edit- oder Hamming-Abstandes). Die Bewertung der Hypothese ist abhängig vom aufgefundenen Abstand.

Kann bei kleinem lexikalischem Abstand (unterschiedliche Bindestriche, Leerzeichen) noch von häufig zutreffenden Hypothesen ausgegangen werden, wird die Zuverlässigkeit solcher Hypothesen mit größerem Abstand geringer. Es bietet sich daher der gezielte Einsatz für die Untersuchung bestimmter Teilmengen der Benennungen an, innerhalb derer zuvor mit anderen Verfahren keine Äquivalenzbeziehungen gefunden werden konnten.

Einordnung: Bewertungsbasierte Optimierung

Analysen von Mehrwortbenennungen

Eingabe: Thesaurusföderation, Thesaurus

Ausgabe: positive Abstraktionshypothesen, positive Bestandshypothesen

Beschreibung: Neben der zeichenweisen Untersuchung von Benennungen kann eine wortweise Untersuchung vorgenommen werden. Dies bietet sich insbesondere für die englische Sprache an, da dort komplexere Begriffe häufig durch eine aus mehreren Wörtern

bestehende Benennung repräsentiert werden. Im Deutschen hingegen werden mehrere Wörter zu *einem* neuen Wort zusammengesetzt (Beispiel: *Abfallverbrennungsanlage*), eine Zerlegung ist zusätzlich erforderlich[3]. Dies zeigt bereits die große Abhängigkeit von Mehrwortanalysen von der zugrunde liegenden Sprache.

Das Durchführen einer Mehrwortanalyse bedeutet das Aufstellen von sprachspezifischen Regeln, deren Implikationsteil die aufzustellende Hypothese beinhaltet. Bereits in der Vorbereitungsphase wurden solche Regeln zur Unterscheidung von Hierarchiebeziehungen in Abstraktions- und Bestandsbeziehungen aufgestellt, vgl. Tabelle 8.3, S. 141, Regeln 5 und 6. Diese Regeln können zum Auffinden von Hierarchiehypothesen aufgegriffen (trifft Regel 5 zu, wird eine Abstraktionshypothese vorgeschlagen, bei Regel 6 eine Bestandshypothese) und um weitere Regeln ergänzt werden. Als Beispiel für eine weitere Regel sei der Fall genannt, dass eine Mehrwortbenennung b_1 aus weniger Wörtern bestehe als eine andere Mehrwortbenennung b_2 und die Benennung b_1 identisch mit dem Beginn der Benennung von b_2 ist (mögliche Zusatzbedingung: falls in b_2 eine Präposition folgt). Trifft diese Regel zu, wird eine Abstraktionshypothese generiert (z.B. für *environmental impact* und *environmental impact of agriculture*).

Eine Bewertung der Hypothese findet aufgrund der Bewertung der die Hypothese erzeugenden Regel statt. Es soll wiederum unterschieden werden können, ob die Regel anhand einer Betrachtung der Deskriptoren oder Nicht-Deskriptoren zutraf.

Einordnung: Initiale Integration

Hierarchien in externen Quellen

Eingabe: Thesaurusföderation, Thesaurus, externe Quellen mit Hierarchiebeziehungen, Teilgraph (optional)

Ausgabe: positive Abstraktionshypothesen, positive Bestandshypothesen

Beschreibung: Wurde die lexikalische Übereinstimmungen von Benennungen mit Benennungen in externen Quellen, die Informationen über Hierarchiebeziehungen enthalten (z.B. WordNet mit Abstraktions- und Bestandsbeziehungen), festgestellt, kann anhand einer Hierarchiebeziehung in der externen Quelle auf eine herzustellende Hierarchiebeziehung in der Thesaurusföderation (Hierarchiehypothese) geschlossen werden (ähnlich Regeln 1 und 2 in Tabelle 8.3, S. 141). Eine Bewertung kann anhand der Unterscheidung zwischen lexikalischer Gleichheit mit der Benennung in der externen Quelle und Deskriptor bzw. Nicht-Deskriptor und der eindeutigen bzw. nichteindeutigen Zuordnung zu einem Begriff in der externen Quelle bestimmt werden.

Für Abstraktionsbeziehungen kann die Transitivität dieser Beziehungen verwendet werden, um mehrere Ebenen (Pfadlänge ≥ 1) innerhalb der externen Quelle zu berücksichtigen. Als Parameter sollte die maximale Pfadlänge angegeben werden können, um das Verfahren so entsprechend der Zielsetzung konfigurieren zu können. Bei der Bewertung sollte die Pfadlänge berücksichtigt werden, zumindest aber unterschieden werden, ob es sich um direkte Unterbegriffe oder um indirekte Unterbegriffe handelt.

Einordnung: Initiale Integration, mit größerer maximaler Pfadlänge auch innerhalb der Bewertungsbasierten Optimierung

Assoziationen in externen Quellen

[3] Für eine Zerlegung deutscher Komposita und die Herleitung von hierarchischen Beziehungen anhand heuristischer Verfahren sei auf [HL77] verwiesen.

Eingabe: Thesaurusföderation, Thesaurus, externe Quelle mit Assoziationsbeziehungen

Ausgabe: positive Assoziationshypothesen

Beschreibung: Entsprechend Hierarchiebeziehungen in externen Quellen. Aufgrund der fehlenden Transitivität der Assoziationsbeziehungen sollte die Pfadlänge jedoch ausschließlich 1 sein. In Wörterbüchern können auch die Verweise ausgewertet werden, um Assoziationshypothesen aufzustellen.

Assoziationsbeziehungen sind zu Beginn der Integration kein bedeutendes Integrationsziel. Stattdessen werden sie zur Auflösung von Konflikten benötigt. Konnten für Teilgraphen aber keine semantisch stärkeren Beziehungen etabliert werden, kann mit diesem Verfahren versucht werden, die Konnektivität zu verbessern.

Einordnung: Bewertungsbasierte Optimierung

Analyse von Mehrwortbenennungen und Hierarchien in externen Quellen

Eingabe: Thesaurusföderation, Thesaurus, externe Quelle mit Hierarchiebeziehungen

Ausgabe: positive Abstraktionshypothesen, positive Bestandshypothesen

Beschreibung: Wie bei der Mehrwortanalyse können Regeln aufgestellt werden, die für die Bestandteile von Mehrwortbenennungen beim Auffinden dieser Bestandteile in einer externen Quelle mit Hierarchiebeziehungen Hypothesen mit Beziehungen zu den Ober- bzw. Unterbegriffen als Implikation besitzen. Als Beispiel seien zweiwortige Benennungen genannt, bei denen beide Wörter Substantive sind. Als Hypothesen können dann Abstraktionbeziehungen zu allen Abstraktionsoberbegriffen des zweiten Wortes vorgeschlagen werden, sofern das zweite Wort in der externen Quelle Oberbegriffe hat, deren Benennungen wiederum in der Thesaurusföderation oder dem Thesaurus vorhanden sind (vgl. auch Regeln 3 und 4 in Tabelle 8.3, S. 141).

Es sollte möglich sein, pro Regel eine Bewertung anzugeben, damit entsprechend differenziert bewertet werden kann.

Einordnung: Initiale Integration

Übereinstimmungsgrad der Gruppen

Eingabe: Hypothese, Thesaurusföderation und Thesaurus mit identischen Gruppen

Ausgabe: Bewertung für Äquivalenz-, Benutze-Kombination-, Hierarchiehypothese

Beschreibung: Falls innerhalb der Thesaurusföderation $\Im$ sowie des zu integrierenden Thesaurus die identische Menge von Gruppen verwendet wird, kann der Übereinstimmungsgrad der Gruppen UeG für zwei Begriffe a, b wie folgt definiert werden:

$$UeG = \frac{|groups(a) \cap groups(b)|}{|groups(a) \cup groups(b)|}$$

wobei $groups(x)$ alle Gruppen eines Deskriptors x liefert.

Je näher UeG an 1 ist, desto besser wird eine Hypothese bewertet. Die Bedeutung dieser Bewertung kann sich für die unterschiedlichen Hypothesentypen unterscheiden, so dass zusätzlich eine hypothesentypspezifische Gewichtung angegeben werden kann.

Ebenso kann die Bewertung in verschiedenen Teilphasen abhängig von der Zielsetzung verschieden ausfallen. Innerhalb der Bewertungsbasierten Optimierung kann mit dem Ziel, weitere Hierarchiehypothesen zu finden, etwa die Bewertung optimistischer ausfallen.

Einordnung: Initiale Integration

Abstand von Äquivalenzhypothesen

Eingabe: Äquivalenzhypothesen, Thesaurusföderation, Thesaurus

Ausgabe: Bewertung für Äquivalenzhypothesen

Beschreibung: Als Äquivalenzhypothesenabstand AeD bezeichnen wir die mittlere Länge der kürzesten Hierarchiepfade von den Benennungen a, b der einen Äquivalenzhypothese zu den Benennungen c, d der anderen Äquivalenzhypothese innerhalb der Thesaurusföderation bzw. des Thesaurus:

$$AeD = \frac{len(a,c) + len(b,d)}{2}$$

wobei $len(a,c)$ die Anzahl der Knoten auf dem kürzesten Pfad von a nach c bezeichnet und 0 ist, falls kein Pfad existiert.

Wenn $len(a,c) \neq 0$ und $len(b,d) \neq 0$ ist, gilt, je näher AeD an 1 ist, desto besser ist die Bewertung beider Hypothesen. Es ist offensichtlich, dass die optimale Bewertung erreicht wird, wenn zwei Äquivalenzhypothesen Deskriptoren beinhalten, die jeweils direkter Unterbegriff sind. Unterschieden werden kann, ob der Pfad ein Abstraktionspfad oder aber ein Bestands- oder gemischter Hierarchiepfad ist.

Einordnung: Initiale Integration

Abstand von Äquivalenzhypothesen zu Äquivalenbeziehungen

Eingabe: Äquivalenzhypothesen, etablierte Äquivalenzbeziehungen, Thesaurusföderation, Thesaurus

Ausgabe: Bewertung für Äquivalenzhypothesen

Beschreibung: Wie Abstand von Äquivalenzhypothesen, jedoch wird der Abstand einer Äquivalenzhypothese zu einem Tupel von bereits per etablierter Äquivalenzbeziehung verbundenen Knoten gemessen.

Einordnung: Zwischenergebnisbasierte Optimierung

Äquivalenzhypothesen aufgrund gemeinsamer Unterbegriffe

Eingabe: Thesaurusföderation, Thesaurus, bereits etablierte Inter-Thesaurus-Äquivalenzbeziehungen

Ausgabe: positive Äquivalenzhypothesen

Beschreibung: Selbst wenn in einem ersten Schritt zwei Begriffe nicht als äquivalent erkannt wurden, kann es gerechtfertigt sein, aufgrund der engen Beziehungen der Unterbegriffe eine Äquivalenz anzunehmen. Konkret wird eine Äquivalenzhypothese zwischen zuvor nicht in Beziehung stehenden Begriffen aus der Föderation und dem Thesaurus vorgeschlagen, wenn die Begriffe mehr als einen gemeinsamen Föderierten Unterbegriff besitzen.

Die Bewertung kann dabei die Anzahl der gemeinsamen Unterbegriffe und die Art der Hierarchiebeziehungen berücksichtigen.

Einordnung: Zwischenergebnisbasierte Optimierung

Analyse der Erläuterungen

Eingabe: Äquivalenzhypothese, Thesaurusföderation, Thesaurus

Ausgabe: Bewertung für Äquivalenzhypothesen

Beschreibung: Falls zu einem der Begriffe einer Äquivalenzhypothese eine Erläuterung angegeben ist, kann bei lexikalischer Gleichheit der Erläuterung (oder einem oder mehreren Wörtern der Erläuterung) mit Benennungen von Oberbegriffen im anderen Thesaurus bzw. der Thesaurusföderation eine positive Bewertung abgegeben werden. Hier wird die Tatsache ausgenutzt, dass Erläuterungen häufig der Homonymauflösung dienen und dazu Angaben zum fachlichen Kontext gemacht werden.

Einordnung: Initiale Integration

Reverse Suche in Definitionen (Bewertung)

Eingabe: Hierarchiehypothese oder Assoziationshypothese, Thesaurusföderation, Thesaurus

Ausgabe: Bewertung für Hierarchiehypothesen bzw. Assoziationshypothesen

Beschreibung: In Definitionstexten wird häufig Bezug auf Oberbegriffe genommen oder verwandte Begriffe werden erwähnt. Jedoch ist eine umfassende Auswertung der natürlichsprachigen Definitionen schwierig. Wiederum können jedoch Regeln aufgestellt werden, die anhand des Auffindens von Benennungen des Oberbegriffs bzw. assoziierten Begriffs in den Definitionen des Unterbegriffs bzw. des anderen assoziierten Begriffs ein verstärktes Vertrauen in die entsprechenden Hypothese ausdrücken.

Die entsprechenden Regeln sollen jeweils wieder unterschiedlich bewertet werden können.

Weitergehend kann auch versucht werden, anhand der Definitionen in den Thesauri Hierarchiehypothesen herzuleiten. In [Hea92] werden dazu ebenfalls sprachabhängige Regeln, die auf Verb- und Nominalgruppen operieren, aufgestellt.

Einordnung: Initiale Integration

Reverse Suche in Definitionen (Erzeugung)

Eingabe: Thesaurusföderation, Thesaurus, Teilgraph

Ausgabe: positive Hierarchie- bzw. Assoziationshypothesen

Beschreibung: Wie Reverse Suche in Definitionen (Bewertung), jedoch wird versucht, anhand der Definitionen in den Thesauri Hierarchiehypothesen herzuleiten. Dazu können etwa die in [Hea92] aufgestellten (sprachabhängigen) Regeln, die auf Verb- und Nominalgruppen operieren, verwendet werden.

Einordnung: Bewertungsbasierte Optimierung

Schwesternanalysen

Eingabe: etablierte Äquivalenzbeziehungen, Thesaurusföderation, Thesaurus

Ausgabe: positive und negative Hierarchiebeziehungen

Beschreibung: Mit den Mitteln der Mehrwortbenennungsanalysen und dem Betrachten von Hierarchien in externen Quellen werden Schwestern noch einmal auf eine mögliche Einordnung untereinander untersucht. Dabei können die Bewertungen aufgrund des eingeschränkten Untersuchungsraumes deutlicher ausfallen als zuvor. Wird eine solche Einstufung durch eine Hierarchiehypothese vorgeschlagen, wird zugleich eine negative Hypothese über die aufzulösende Beziehung aufgestellt.

Einordnung: Zwischenergebnisbasierte Optimierung

Abstraktionsdistanz

Eingabe: etablierte Äquivalenzbeziehungen, Thesaurusföderation, Thesaurus

Ausgabe: bewertete Äquivalenzbeziehungen

Beschreibung: Die bei der Darstellung des Standes der Forschung in Abschnitt 3.2.3.2.2, S. 41, vorgestellten semantischen Abstandsmaße können zu einer Bewertung von Äquivalenzhypothesen herangezogen werden. An [SC96] und [SC97] angelehnt, kann etwa die Abstraktionsdistanz zweier Begriffe als Summe der Wichtigkeit ihrer nichtgemeinsamen Oberbegriffe definiert werden. Die Wichtigkeit wiederum ist der Kehrwert der Tiefe innerhalb der Hierarchierelation. Je geringer die Abstraktionsdistanz ist, desto besser fällt die Bewertung der Äquivalenzhypothesen aus.

Für eine zuverlässige Berechnung der Abstraktionsdistanz ist es erforderlich, dass die übergeordneten Begriffshierarchien bereits integriert sind. Daher kann dieses Verfahren erst in einem fortgeschrittenen Stadium der Integration eingesetzt werden. Das Vertrauen in die Bewertung kann also mit dem Fortschreiten der Integration vergrößert werden.

Einordnung: Zwischenergebnisbasierte Optimierung

Homonymanalysen

Eingabe: etablierte Äquivalenzbeziehungen, Thesaurusföderation, Thesaurus, Teilgraph

Ausgabe: negative Äquivalenzhypothesen

Beschreibung: Einfache Tests auf lexikalische Gleichheit zum Finden von Äquivalenzhypothesen produzieren aufgrund möglicher Homonymie auch nicht zutreffende Vorschläge. Um solche unzutreffenden Vorschläge zu finden, wurden bereits das semantische Abstandsmaß der Abstraktionsdistanz, der Übereinstimmungsgrad von Gruppen und der Abstand von Äquivalenzhypothesen eingeführt. Zusätzlich kann die Analyse der Erläuterungen dazu dienen, solche Äquivalenzhypothesen für homonyme Begriffe herauszufiltern. Eine weitere Möglichkeit ist, mit Hilfe externer Quellen ein Homonymie-Potential zu berechnen. Wird dazu auf WordNet zugegriffen, wird das *Homonymie-Potential* für eine Benennung als Anzahl des Auftretens der Benennung in verschiedenen Synsets bestimmt. Je größer das Homonymie-Potential ist, desto geringer wird die Hypothese bewertet. Die Qualität der externen Quelle sowie die Häufigkeit des Auffindens von Benennungen des Thesaurus in der externen Quelle bestimmen das Vertrauen in die Ergebnisse dieses Verfahrens.

Der Einsatz solcher Homonymanalysen kann zur Reduktion von Äquivalenzbeziehungen in gegebenen Teilgraphen ebenso wie zur Bewertung von Äquivalenzhypothesen dienen.

Einordnung: Bewertungsbasierte Optimierung

Analyse von Assoziationsbeziehungen

Eingabe: Thesaurusföderation, Thesaurus

Ausgabe: positive Hierarchiehypothesen, negative Assoziationshypothesen

Beschreibung: Wie in Abschnitt 9.3.2, S. 163, gezeigt wurde, kann die Analyse der Thesauri ergeben, dass es sich bei Assoziationsbeziehungen innerhalb eines Thesaurus um versteckte Hierarchiebeziehungen handelt. Ist dies der Fall, sollen während der Integration entsprechende Assoziationsbeziehungen wie Hierarchiebeziehungen behandelt werden. Ein erster Schritt dazu ist es, solche versteckten Hierarchiebeziehungen zu

erkennen. Dies kann durch lexikalische Vergleiche der Benennungen mit den Benennungen in anderen Thesauri oder externen Quellen geschehen. Werden Übereinstimmungen zwischen Benennungen assoziierter Begriffe und Benenungen mit Hierarchiebeziehungen in anderen Komponenten-Thesauri oder externen Quellen oder externen Thesauri gefunden, wird eine positive Hierarchiehypothese und zugleich eine negative Assoziationshypothese aufgestellt.

Einordnung: Initiale Integration

Gezielte Analyse schwacher Hypothesen

Eingabe: Hypothesen, Teilgraph

Ausgabe: positive Hypothesen, negative Hypothesen

Beschreibung: Sollen innerhalb von Graphenausschnitten weitere Inter-Thesaurus-Beziehungen gefunden werden, die verschiedenen Verfahren jedoch keine weiteren Hypothesen liefern, deren Bewertung ausreichend für eine Faktenerzeugung ist, kann die Bewertung vorhandener Hypothesen, die Knoten in den Teilgraphen beinhalten, verbessert werden. Dazu können z.B. die bisher am besten bewerteten Hypothesen zusätzlich so gut bewertet werden, dass sie zu Fakten werden. Oder diese Hypothesen werden gezielt dem menschlichen Experten vorgelegt, damit dieser darüber entscheidet, ob sie zu Fakten werden sollen oder nicht.

Dieses Verfahren kann entsprechend auch auf negative Hypothesen angewandt werden.

Einordnung: Bewertungsbasierte Optimierung

Die Auflistung von Verfahren erhebt keinerlei Anspruch auf Vollständigkeit, weitere Verfahren existieren, neue Verfahren können entwickelt werden. Die aufgeführten Kriterien und Beispiele ermöglichen eine entsprechende Einordnung solch neuer Verfahren.

Es sei zusätzlich darauf hingewiesen, dass auch während der Faktenerzeugung, Konfliktmarkierung und -auflösung Hypothesen aufgestellt und weitere Fakten erzeugt werden können, vgl. Abschnitt 11.2. Insbesondere wird dort gezeigt, dass Ergänzende Begriffe einen wichtigen Beitrag zur Auflösung von Konflikten beitragen können.

Da wir im Rahmen dieser Arbeit die Entwicklung und Bewertung der Verfahren nicht in den Vordergrund stellen, beschränken wir uns auf eine Darstellung von Ergebnissen ausgewählter Verfahren, um deren Wirkung zu verdeutlichen (vgl. auch die Studienarbeit [Sun01] in der Verfahren exemplarisch implementiert und die Ergebnisse ausgewertet wurden).

Beispiel 11.1 *Das Verfahren Lexikalische Gleichheit (Deskriptoren, Nicht-Deskriptoren, externe Quellen) ist das häufigst verwendete Verfahren zum Finden von Äquivalenzbeziehungen. Bei der Anwendung dieses Verfahrens auf die Thesauri GEMET und AGROVOC unter Berücksichtigung von Deskriptoren, Nicht-Deskriptoren und Synonymen innerhalb der externen Quelle WordNet wurden die in Tabelle 11.1 gezeigten Anzahlen von Äquivalenzhypothesen erzeugt.*

Es ist offensichtlich, dass ein ausschließlicher Vergleich der Deskriptoren bei weitem nicht alle Äquivalenzhypothesen gefunden hätte. Basis zur Festlegung der für die Bewertung erforderlichen Konfidenzfaktoren sind die Ergebnisse der Thesaurusevaluierung. Aufgrund der festgestellten großen Anzahl von Quasy-Synonymen in AGROVOC ist es sinnvoll, den Äquivalenzhypothesen, die auf Basis der AGROVOC-Nicht-Deskriptoren gefunden werden, eine deutlich geringere Bewertung zu geben, als denen, die auf Basis der GEMET-Nicht-Deskriptoren gefunden wurden.

GEMET	AGROVOC	Anzahl Hypothesen	Konfidenzfaktor
Deskriptor	Deskriptor	1883	0.8
Deskriptor	Nicht-Deskriptor	584	0.6
Nicht-Deskriptor	Deskriptor	235	0.8
Nicht-Deskriptor	Nicht-Deskriptor	54	0.4
Deskriptor	WordNet-Synonym	1113	0.2
WordNet-Synonym	Deskriptor	15	0.3
Nicht-Deskriptor	WordNet-Synonym	124	0.2
WordNet-Synonym	Nicht-Deskriptor	261	0.2
WordNet-Synonym	WordNet-Synonym	10730	0.1

Tabelle 11.1: Anzahl der anhand lexikalischer Gleichheit erzeugten Äquivalenzhypothesen

Durch Betrachtung der WordNet-Synonyme vergrößert sich die Menge der gefundenen Äquivalenzhypothesen erheblich, jedoch ist aufgrund der häufig nicht eindeutigen Zuordnung einer Thesaurusbenennung zu einem WordNet-Synset hier mit vielen unzutreffenden Hypothesen zu rechnen. Daher dürfen diese Hypothesen nur niedrig bewertet werden. Sie können jedoch als Basis für weitere Verfahren (z.B. Homonymanalysen oder gezielte Analyse schwacher Hypothesen) dienen.

Falls mehrfach lexikalische Gleichheit festgestellt werden kann, wird anhand der verschiedenen Konfidenzfaktoren mittels der in Abschnitt 7.4.1, S. 130, dargestellten Aggregationsformel die kombinierte Bewertung berechnet.

Beispiel 11.2 *Werden die im obigen Beispiel mittels lexikalischem Vergleich erzeugten Äquivalenzhypothesen auf ihren Abstand hin untersucht, kann dieser für 186 Hypothesen, also ca. 2 % aller Hypothesen, ermittelt werden. Werden die anhand von WordNet-Synonymen gefundenen, häufig nicht zutreffenden Hypothesen jedoch nicht berücksichtigt, steigt der Anteil auf ca. 10 %. Davon entfällt der weitaus überwiegende Anteil auf Hypothesen mit einem Abstand von 1 (vgl. Tabelle 11.2). Eine manuelle Überprüfung dieser Hypothesen ergibt in allen Fällen, dass die zu etablierende Beziehung korrekt ist. Somit können hohe Konfidenzfaktoren konfiguriert werden, die zu einer entsprechend guten Bewertung führen.*

Hypothesenabstand	Anzahl Hypothesen	Konfidenzfaktor
1	138	0.9
1.5	40	0.8
≥ 2	8	0.7

Tabelle 11.2: Anzahl der Äquivalenzhypothesen, für die ein Hypothesenabstand berechnet werden kann.

Beispiel 11.3 *Sind bereits erste Äquivalenzbeziehungen etabliert, können Äquivalenzhypothesen aufgrund mehrerer gemeinsamer Untebegriffe vorgeschlagen werden. Auf der Basis von ca. 1200 etablierten Äquivalenzbeziehungen konnten so insgesamt 387 Hypothesen erzeugt werden, davon waren 330 neue Vorschläge (Beispiele:* animal product *und* processed animal products *haben die gemeinsamen Unterbegriffe* leather *und* fur, farming technique *und* cultural method *mit den gemeinsamen Unterbegriffen* irregation farming / irregated farming *und* contour farming*). Ca. 75 % dieser Hypothesen werden dabei aufgrund gemeinsamer Abstraktionsunterbegriffe, 20 % aufgrund gemeinsamer Bestandsunterbegriffe und 5 % aufgrund gemeinsamer*

Abstraktions- und Bestandsuntebegriffe gefunden. Die Vorschläge aufgrund gemeinsamer Abstraktionsunterbegriffe sind am zuverlässigsten, so dass sie die beste Bewertung erhalten.

Eine Reihe der erzeugten Vorschläge betreffen Äquivalenzbeziehungen, die im Zusammenhang nur zweier Thesauri Begriffe föderieren, im strengeren Sinne jedoch Hierarchiebeziehungen erfordern (Beispiel: animal housing *und* agricultural building*). Eine entsprechende zusätzliche Untersuchung dieser Hypothesen ist sinnvoll.*

Bereits diese Beispiele zeigen, wie die Verfahren zusammenspielen, um schrittweise die Lösung zu verbessern. Lexikalische Verfahren sind die Grundlage der Integration, Struktur- und zusätzliche Informationen auswertende Verfahren sichern die Ergebnisse ab und schaffen neue Vorschläge.

In Tabelle 11.3 werden die vorgestellten Verfahren und deren Einordnung noch einmal in der Übersicht dargestellt. Alle Bewertungsverfahren können in späteren Teilphasen ebenfalls angewandt werden.

11.1.3 Konfigurieren von Verfahren

Konfigurieren der Verfahren ist notwendig, damit die Verfahren situationsabhängig die Güte ihrer Hypothesen, die aufgrund verschiedener Algorithmen, Regeln und zugrundeliegender Informationen aufgestellt werden, bewerten können. Bereits im vorangegangen Abschnitt 11.1.2 haben wir bei der Beschreibung der Verfahren auf mögliche differenzierte Bewertungen hingewiesen. Unter Konfigurieren verstehen wir das Einstellen von Konfidenzfaktoren abhängig vom jeweiligen Algorithmus, der entsprechenden Regel und Information, die zu einer Hypothese bzw. deren Bewertung führen. Als *Konfiguration* bezeichnen wir die Menge von Konfidenzfaktoren für ein Verfahren.

Die Möglichkeiten, zu den für die Bewertungen erforderlichen Konfidenzfaktoren zu gelangen, sind vielfältig und nicht zuletzt Erfahrungswerte, die die menschlichen Entwickler in die Verfahren mit einbringen. Generell kann jedoch, um den unterschiedlichen Situationen gerecht zu werden, Folgendes gefordert werden:

- Für ähnliche Thesauri sollen die Verfahren ähnlich konfiguriert werden. Das Ähnlichkeitsmaß hängt dabei von der Art des Verfahrens ab: Für lexikalische Verfahren kann die Ähnlichkeit anhand ähnlicher (Teil-) Ergebnisse der quantitativen Analyse der Benennungen, für strukturbasierte Verfahren anhand ähnlicher (Teil-) Ergebnisse der quantitativen Analyse der Relationen bestimmt werden. Ist die Ähnlichkeit groß genug, kann eine Konfiguration beibehalten werden, ansonsten ist das Verfahren neu zu konfigurieren.

- Nur konfigurierte Verfahren dürfen zum Finden oder Bewerten von Hypothesen verwandt werden. Soll ein nicht-konfiguriertes Verfahren angewandt werden, ist dies zuvor zu konfigurieren.[4]

- Zur Konfiguration der Verfahren liefern die Ergebnisse der Komponententhesaurusanalyse wichtige Hinweise, beispielhaft seien die Zuverlässigkeit der Nicht-Deskriptoren (strenge Synonyme vs. Quasy-Synonyme) und das Abstraktions-/Bestandsverhältnis zur Beurteilung der Wichtigkeit dieser Hierarchierelationstypen genannt. Zur Konfiguration bietet sich daher eine verfahrensspezifische Auswertung der Analyseergebnisse an.

[4] Dies festzustellen und die Konfiguration zu beauftragen, ist eine weitere Aufgabe des Steueragenten.

	Initiale Integration		Zwischenergebnisbasierte Optimierung		Bewertungsbasierte Optimierung	
	Finden	Bewerten	Finden	Bewerten	Finden	Bewerten
ÄH	Lexikalische Gleichheit (D, ND, ext. Quellen)	Übereinstimmungsgrad der Gruppen	ÄH aufgrund gemeinsamer Unterbegriffe	Abstand von ÄH zu Äquivalenzbeziehungen	Lexikalische Ähnlichkeit	Gezielte Analyse schwacher Hypothesen
		Abstand von ÄH	Abstraktionsdistanz		Homonymanalyse	
		Analyse der Erläuterungen				
BKH	Lexikalische Gleichheit zwischen D und BK-ND	Übereinstimmungsgrad der Gruppen				Gezielte Analyse schwacher Hypothesen
HH	Analyse von Mehrwortbenennungen	Übereinstimmungsgrad der Gruppen	Schwesternanalyse			Gezielte Analyse schwacher Hypothesen
	Hierarchien in externen Quellen	Reverse Suche in Definitionen (Bewertung)			Reverse Suche in Definitionen (Erzeugung)	
	Analyse von Assoziationsbez.					
AH					Assoziationen in externen Quellen	Gezielte Analyse schwacher Hypothesen
EBH	Lexikalische Gleichheit zwischen BK-ND					Gezielte Analyse schwacher Hypothesen

Erläuterung:
ÄH = Äquivalenzhypothese(n)
BKH = Benutze-Kombination-Hypothese(n)
HH = Hierarchiehypothese(n)
AH = Assoziationshypothese(n)
EBH = Ergänzende-Begriff-Hypothese(n)
D = Deskriptor(en)
ND = Nicht-Deskriptor(en)

Tabelle 11.3: Geordnete Übersicht über die vorgestellten Integrationsverfahren

- Werden die vorliegenden Analyseergebnisse der Komponententhesauri sowie die Erfahrung der Verfahrensersteller und menschlichen Experten für eine Bewertung einzelner Verfahrensbereiche (z.B. einer Regel innerhalb eines Verfahrens) als nicht ausreichend eingeschätzt, kann die Konfiguration anhand von Testdaten durchgeführt werden (vgl. auch Abschnitt 8.1.2.1.2, S. 142).

11.2 Faktenerzeugung und Konfliktmarkierung

Aufgabe der Faktenerzeugung ist es, die vorhandenen Hypothesen qualitativ zu überprüfen und zu bewerten, so dass Fakten entstehen, die Änderungen am Integrationswissen nach sich ziehen. Diese Aufgabe beinhaltet die Berechnung eines aggregierten Konfidenzfaktors aus den Einzelbewertungen einer Hypothese (vgl. Abschnitt 11.2.1) sowie die qualitative Überprüfung von Hypothesen und Fakten einschließlich einer Konfliktmarkierung (vgl. Abschnitt 11.2.2).

11.2.1 Berechnung einer aggregierten Hypothesenbewertung

Für die Bewertung der Hypothesen haben wir bereits in Abschnitt 7.4 ein Bewertungsmodell erarbeitet, das aus den Bewertungen der Agenten sowie den von den Agenten abgegebenen Hypothesenbewertungen eine Gesamtbewertung berechnet. Dieses Bewertungsmodell kann vom Moderator verwendet werden.

In Abschnitt 7.3.7 haben wir ebenfalls verschiedene Alternativen eingeführt, wie mit Hypothesen verfahren werden kann, die aufgrund einer Gesamtbewertung zwischen unterem und oberem Grenzwert nicht zu Fakten werden. In der Realisierungsphase haben wir – im Gegensatz zur Vorbereitungsphase – nicht den Bedarf, dass über eine unsichere Hypothese auf jeden Fall entschieden werden muss. Bleibt sie unsicher, kann sie solange Hypothese bleiben, bis vom menschlichen Experten entschieden wird, alle aufgrund einer zu unsicheren Bewertung für die weitere Betrachtung nicht mehr relevanten Hypothesen zu löschen. Wiederum kann ein Bewertungsintervall angegeben werden, innerhalb dessen unsichere Hypothesen gelöscht werden. Der Zeitpunkt einer solchen Aktion wird üblicherweise die Feststellung sein, dass innerhalb einer Teilphase der Integration (vgl. Abschnitt 10.2.2) keine weitere Optimierung möglich ist.

11.2.2 Qualitative Überprüfung von Hypothesen und Fakten

Das Ziel einer qualitativen Überprüfung von Hypothesen bzw. Fakten ist es, solche Hypothesen bzw. Fakten zu finden, die von Experten erzeugt wurden, den Qualitätskriterien bei der Erstellung einer Thesaurusföderation aber nicht entsprechen und daher nicht weiter betrachtet werden sollen. Diese Überprüfung ist erforderlich, da für die Experten nicht vorausgesetzt werden kann, dass diese alle Qualitätskriterien kennen und bei der Aufstellung und Bewertung von Hypothesen berücksichtigen.

Zur qualitativen Überprüfung von Blackboard-Einträgen gilt es folgende Fragen zu beantworten:

- Welches sind die Qualitätskriterien, anhand derer die Überprüfung stattfinden sollen?
- Welche Einträge sind zu prüfen? Sollen offene Hypothesen, Fakten (mit Hilfe des Bewertungsmodells akzeptierte oder abgelehnte Hypothesen) oder alle Hypothesen (offene Hypothesen und Fakten) überprüft werden? Welche Rolle spielen implizierte Beziehungen?

- Welche Aktionen erfolgen bei Nichterfüllung der Qualitätskriterien?

Zur Beantwortung dieser Fragen ist es erforderlich, das Ziel der qualitativen Überprüfung von Hypothesen und Fakten in Teilziele zu zerlegen:

Hypothesen-Reduktion: Hypothesen, die den Qualitätskriterien nicht entsprechen, sollen frühzeitig erkannt und entfernt werden, um die Menge der zu verarbeitenden Daten möglichst klein zu halten.

Auflösung von Widersprüchen zwischen Fakten: Alle aufgrund der Faktenmenge resultierenden Widersprüche (Verstöße gegen Invarianten, Akzeptanz einer zuvor abgelehnten Beziehung) sollen *erkannt und aufgelöst* werden.

Auflösen oder Markieren weiterer Widersprüche: Alle weiteren Verstöße gegen Invarianten, also solche, die *nicht ausschließlich durch direkte Fakten* erzeugt wurden, sollen *erkannt* und *aufgelöst oder markiert* werden.

Alle Qualitätskriterien für offene Hypothesen und Fakten lassen sich aus der Tatsache ableiten, dass möglichst weitgehende Widerspruchsfreiheit gegen die in Abschnitt 6.4.3, S. 104ff, aufgestellten Invarianten angestrebt wird. Die Bedeutung der Widerspruchsfreiheit wird in den Abschnitten 11.2.2.1 bis 11.2.2.4 daher für unterschiedliche Datenausschnitte betrachtet.

Um eine hohe Flexibilität zu erreichen, kann die Spezifikation der zu überprüfenden Qualitätskriterien und der Reaktionen bei Verletzung durch Bedingungs-/Aktionsregeln geschehen. Das die Überprüfungsregeln auslösende Ereignis ist das Schreiben einer Hypothese auf das Blackboard (nur bei Betrachtung einzelner Hypothesen) bzw. das Erzeugen einer neuen Gesamtbewertung der Hypothesen durch den Moderator.

11.2.2.1 Überprüfung einzelner Hypothesen

Jede einzelne Hypothese über einer Inter-Thesaurus-Beziehung hat der Invariante des richtigen Einsatzes der Thesaurus-verbindenden Kanten bzgl. der unterschiedlichen Thesauruszugehörigkeit der Knoten zu genügen (vgl. Abschnitt 6.4.3.3, 105). Es resultiert folgendes Kriterium (Bedingung):

Q1: Zwei Knoten, die durch eine $\mathcal{BS}$-, $\mathcal{BK}$-, $\mathcal{UA}$-, $\mathcal{UP}$- oder $\mathcal{VB}$-Kante verbunden werden, müssen aus unterschiedlichen Thesauri stammen.

Verletzt eine Hypothese dieses Kriterium Q1, wird sie unabhängig davon, ob es sich um eine offene Hypothese oder ein Faktum handelt, verworfen (Aktion, falls Bedingung nicht erfüllt).

11.2.2.2 Überprüfung der Menge der Hypothesen

Wird die gesamte Menge der Hypothesen betrachtet, gelten eine Reihe weiterer Kriterien. Prinzipiell könnten alle in Abschnitt 6.4.3, S. 104, vorgestellten Invarianten auf ihre Gültigkeit innerhalb der Menge der offenen Hypothesen bzw. Faktenmenge überprüft werden. Aus Performanz-Gründen gilt es abzuwägen, welcher Aufwand zur Überprüfung dieser Mengen getrieben werden soll, da man eine vollständige Überprüfung früherer Fakten, deren Auswirkungen bereits in das Integrationswissen übernommen wurden, zusätzlich berücksichtigen muss.

Wir beschränken uns daher hier auf Invarianten, bei denen Verstöße mit großem Wirkungsgrad zu erkennen sind. Konkret bedeutet dies, bei geringem Überprüfungsaufwand (maximal logarithmischer Aufwand) Verstöße gegen für explizite Beziehungen obligatorische Invarianten zu erkennen, deren Auswirkung auf die Hypothesenmenge durch das Entfernen von Hypothesen zu einer Verringerung der Datenmenge für Folgeoperationen führt. Des Weiteren kann auch die Erfolgsquote beim Auffinden von Verstößen darüber entscheiden, ob die Überprüfung bereits an dieser Stelle oder aber erst bei zusätzlicher Betrachtung des gesamten Integrationswissens (vgl. Abschnitt 11.2.2.3, S. 198ff) durchgeführt werden soll. Das bedeutet, dass Überprüfungen, die in einem Durchlauf bei einer bestimmten Mindestanzahl von zu überprüfenden Hypothesen keine oder sehr wenige Konflikte festgestellt haben, in der Zukunft bei der gleichen Menge von Agenten bzw. einer Teilmenge dieser Agenten, die die Hypothesen aufgestellt und bewertet haben, nicht mehr durchgeführt werden.

Aus der Invariante des richtigen Einsatzes der Thesaurus-verbindenden Kanten (vgl. wiederum Abschnitt 6.4.3.3, S. 105) können für die Überprüfung der Menge der Kanten folgende Kriterien hergeleitet werden:

Q2: Wird für einen Knoten $n \in N$ eine $\mathcal{BK}$-Kante $(n, \mathcal{BK}, m_1)$ vorgeschlagen, muss für diesen Knoten eine weitere $\mathcal{BK}$-Kante $(n, \mathcal{BK}, m_2)$ existieren, so dass m_1 und m_2 aus dem selben Thesaurus stammen.

Q3: Werden für einen Knoten $n \in N$ mehrere $\mathcal{BS}$-Kanten vorgeschlagen, müssen die Thesauri der jeweiligen Knoten paarweise disjunkt sein.

Q4: Wird für einen Knoten $n \in N$ eine $\mathcal{BS}$-Kante zu einem Knoten $m \in N$ vorgeschlagen, darf nicht gleichzeitig eine $\mathcal{BK}$-Kante für n zu einem Knoten innerhalb des Thesaurus von m vorgeschlagen werden.

Da wir von allen Experten erwarten, bei $\mathcal{BK}$-Kantenvorschlägen alle relevanten Knoten innerhalb eines Thesaurus zu liefern, können Hypothesen, die Q2 nicht erfüllen, entfernt werden.

Bei offenen Hypothesen ist es nicht erforderlich, solche zu entfernen, die das Kriterium Q3 verletzen, solange nicht mindestens eine dieser Hypothesen Faktum geworden ist. Schließlich soll eine Bewertung der jeweiligen offenen Hypothesen durch die verschiedenen Agenten möglich sein, um über eine Gesamtbewertung zwischen relevanten und irrelevanten Hypothesen entscheiden zu können. Werden jedoch mehrere dieser offenen Hypothesen zu akzeptierten Fakten, gilt es das am besten bewertete Faktum auszuwählen. Bei wenig sicheren Entscheidungen (Bewertungsdifferenz unterschreitet vorgegebenes Minimum) kann der menschliche Experte die Auswahl aus den Möglichkeiten treffen. Die nicht-selektierten Fakten werden als abgelehnte Fakten mit der Begründung „Mehrfachäquivalenz“ gespeichert. Wird genau eine der offenen Hypothesen Faktum, werden die anderen Hypothesen entfernt.

Beispiel 11.4 *In AGROVOC existiert ein Deskriptor* pulp and paper industry *mit den dort synonym gesetzten Nicht-Deskriptoren* pulp industry, paper industry *und* cellulose industry. *Die Agenten schlagen $\mathcal{BS}$-Kante von* pulp and paper industry *zu den GEMET-Deskriptoren* pulp industry, paper industry *und* cellulose industry *vor, mit jeweils einer ausreichenden Gesamtbewertung, um als Faktum anerkannt zu werden (vgl. Abbildung 11.2). Durch die Überprüfung von Q3 wird die Kante aber schließlich nur zwischen* pulp and paper industry *und* paper industry *etabliert, die weiteren Kanten werden als abgelehnte Kanten gespeichert. Diese auf den ersten Blick überraschende Entscheidung wird durch die strukturelle Ähnlichkeit begründet, die schließlich den Ausschlag für die beste Bewertung gibt: Sowohl* pulp and paper industry *als auch* paper

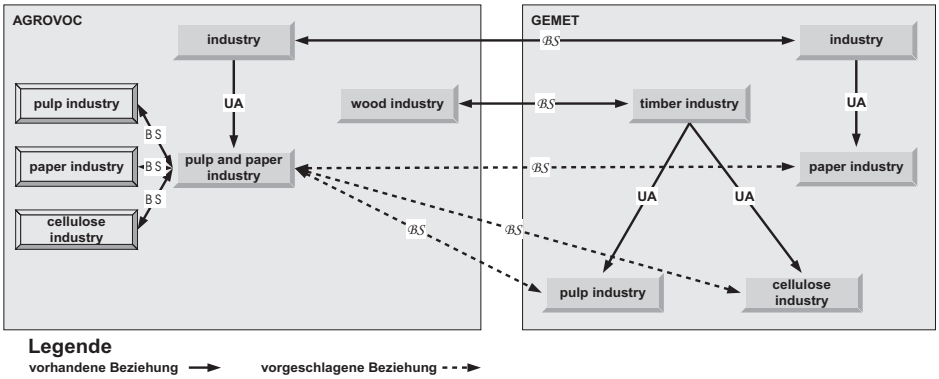

Abbildung 11.2: Mehrfachvorschläge für $\mathcal{BS}$-Kanten mit identischem Deskriptor

industry *sind direkte Unterbegriffe von* industry. *Die GEMET-Deskriptoren* pulp industry *und* cellulose industry *hingegen sind Abstraktionsunterbegriffe von* timber industry, *das zuvor bereits mit* wood industry *aus AGROVOC über eine $\mathcal{BS}$-Kante verbunden wurde.*

Bei Verstoß gegen Q4 kann ähnlich verfahren werden wie bei Verstoß gegen Q3. Allerdings wird der $\mathcal{BS}$-Kante mehr Bedeutung eingeräumt als der $\mathcal{BK}$-Kante und somit werden bei Fakten für beide Kantentypen ausgehend von einem identischen Knoten die $\mathcal{BK}$-Kanten-Fakten entfernt und als abgelehnte Fakten („$\mathcal{BS}$-Kante vorhanden“) gespeichert.

Die Invariante der Einzigartigkeit einer Kante darf nur aufgrund implizierter Beziehungen verletzt werden (vgl. Abschnitt 6.4.3.5, S. 105). Hypothesen und Fakten jedoch enthalten explizite Beziehungen. Daher gilt als Kriterium:

Q5: Zwischen zwei Knoten $m, n \in N, m \neq n$ darf maximal eine Inter-Thesaurus-Kante vorgeschlagen werden.

Kriterium Q5 kann entsprechend Q3 behandelt werden. Die Ablehnungsbegründung der entfernten Fakten lautet „Verletzung Einzigartigkeit der Kanten“.

Die Invariante der Schwesternassoziationen (vgl. Abschnitt 6.4.3.7, S. 107 und Abschnitt 6.3.7.2, S. 97) darf ebenfalls nur durch implizierte bzw. Intra-Thesaurus-Beziehungen verletzt werden. Daher gilt:

Q6: Werden zwei Knoten $m_1, m_2 \in N$ als Abstraktions- oder Bestandsunterbegriffe zu einem Begriff $n \in N$ vorgeschlagen, darf nicht gleichzeitig eine Assoziationsbeziehung zwischen m_1 und m_2 vorgeschlagen werden.

Bei Kriterium Q6 gilt, dass die Assoziationsbeziehung aus den Hypothesen entfernt werden kann, sobald die Hierarchiebeziehungen Fakten werden und umgekehrt. Werden Assoziations- und Hierarchieziehungen Fakten, wird die semantisch schwächere Assoziationsbeziehung entfernt und als abgelehntes Faktum gespeichert („Verletzung Schwesternassoziation“).

11.2.2.3 Überprüfung von Faktenmenge und vorhandenem Integrationswissen

Die zeitliche Differenz bei der Generierung der Fakten darf für die erforderliche Auflösung von Widersprüchen zwischen den Fakten keine Rolle spielen. Daher erfordert das Erkennen aller

widersprüchlichen Fakten die Betrachtung der *aktuellen Faktenmenge* (Fakten, die zu einem Zeitpunkt t_1 auf dem Blackboard vorhanden sind, aber noch nicht in das persistente Integrationswissen übernommen wurden) zusammen mit allen *persistenten Fakten* (Fakten, die vor t_1 erzeugt wurden und zu dem zum Zeitpunkt t_1 persistenten Integrationswissen beigetragen haben). Widersprüche können unterschieden werden in Widersprüche gegen die Invarianten und in Widersprüche bzgl. der Akzeptanz bzw. Ablehnung.

11.2.2.3.1 Widersprüche gegen die Invarianten Die Kriterien Q1 und Q2 gelten für jede einzelne Hypothese bzw. für eine gleichzeitig vorliegende aktuelle Menge von Fakten und sind daher für die gemeinsame Prüfung aktueller und persistenter Fakten nicht mehr relevant. Die Kriterien Q3 bis Q6 hingegen gelten übertragend auch für diese gemeinsame Prüfung. Wir bezeichnen sie im Folgenden Q3' bis Q6'.

Bei Verstoß gegen Q3' (das bedeutet es gibt zwei Kanten $(n, \mathcal{BK}, m_1)$, $(n, \mathcal{BK}, m_2)$, $m_1 \neq m_2$) wird der Begriffsinhalt zweier durch die Knoten m_1 und m_2 repräsentierten Begriffe als Bestandteil des Begriffsinhaltes des durch n repräsentierten Begriffes angesehen. Zur Auflösung des resultierenden Konfliktes werden daher als Alternativen vorgeschlagen, auf eine der $\mathcal{BS}$-Kanten zu verzichten oder sie durch eine $\mathcal{VB}$-Kante zu ersetzen (und somit diese Sichtweise nicht beizubehalten) oder die Kanten durch $(n, \mathcal{UA}, m_1)$ und $(n, \mathcal{UA}, m_2)$ bzw. $(n, \mathcal{BK}, m_1)$ und $(n, \mathcal{BK}, m_2)$ zu ersetzen (und damit die beschriebene Sichtweise auf formal korrekte Weise dem jeweiligen Standpunkt angemessen auszudrücken). Die Entscheidung wird durch den menschlichen Experten getroffen.

Zur Behandlung bei Widersprüchen gegen weitere Invarianten unterteilen wir die Kanten in vier verschiedene Klassen, die wir aufsteigend nach ihrer semantischen Stärke ordnen: $\{\mathcal{VB}\}$, $\{\mathcal{BK}\}$, $\{\mathcal{UA}, \mathcal{UP}, \mathcal{OA}, \mathcal{OP}\}$ und $\{\mathcal{BS}\}$. Bei Verstößen gegen Q4' bis Q6' gilt es, das die semantisch schwächere Kante belegende Faktum zu entfernen, wenn die semantisch stärkere Kante nicht zugleich abgelehnte Kante ist. Stammen die Kanten aus einer Klasse, wird dem menschlichen Experten die Entscheidung zwischen den Alternativen überlassen.

Als weiteres Kriterium wird die Redundanzfreiheit der Inter-Thesaurus-Abstraktionspfade gefordert (vgl. Abschnitt 6.4.3.7, S. 107 und Abschnitt 6.3.6.3, S. 89f). Das bedeutet:

Q7': Falls es einen Pfad von $m \in N$ über $\mathcal{UA}-$Kanten zu $n \in N$ mit einer Länge größer 2 gibt, darf es nicht gleichzeitig eine Kante $(m, \mathcal{UA}, n) \in E$ geben.

Bei Verstoß gegen Q7' wird das die Kante $(m, \mathcal{UA}, n)$ ausdrückende Faktum entfernt und als abgelehntes Faktum mit der Begründung „Redundante Abstraktionskante" gespeichert.

Ebenso wird die Zyklenfreiheit der Inter-Thesaurus-Hierarchiepfade – die zugleich die Zyklenfreiheit der Inter-Thesaurus-Abstraktions- und Bestandspfade bedeutet – gefordert (vgl. Abschnitt 6.4.3.7, S. 107 und Abschnitt 6.3.6.5, S. 94f). Da bei Betrachtung aktueller und persistenter *Fakten* nur explizite Beziehungen betrachtet werden, bedeutet dies:

Q8': Es darf keinen zyklischen Pfad über $\mathcal{UA}$-, $\mathcal{UP}$- und $\mathcal{BS}$-Kanten geben.

Bei Verstoß gegen Q8' wird dem menschlichen Experten aufgetragen, eine der am Zyklus beteiligten Kanten zu entfernen oder als $\mathcal{VB}$-Kante zu klassifizieren. Das entsprechende Faktum wird als abgelehntes Faktum mit der Begründung „Zyklenbestandteil" gespeichert.

11.2.2.3.2 Widersprüche bzgl. Akzeptanz und Ablehnung In Fällen, in denen ein persistentes Faktum eine Aussage über die Ablehnung einer Intra-Thesaurus-Beziehung bzw. eines Ergänzenden Begriffes macht (negatives Faktum), ein aktuelles Faktum diese Beziehung bzw. diesen Begriff aber akzeptiert (positives Faktum), sowie in umgekehrten Fällen muss entschieden werden, welches Faktum Gültigkeit behält.

Kriterien für eine Entscheidung sind die Begründung des persistenten Faktums, die Zustimmung des menschlichen Experten zum persistenten bzw. aktuellen Faktum (dem menschlichen Experten wird als letzte Entscheidungsinstanz ein größeres Gewicht eingeräumt) und wie nahe die Gesamtbewertung des aktuellen Faktums der optimalen Gesamtbewertung von -1 bzw. 1 ist (große Gesamtkonfidenzfaktoren schließen Unsicherheiten weitestgehend aus). Anhand dieser Kriterien kann eine Entscheidungstabelle aufgestellt werden, die an die Präferenzen der Integrationsexperten angepasst werden kann.

Für den Fall eines negativen persistenten Faktums und eines positiven aktuellen Faktums schlagen wir folgende Standardbelegung vor:

Begründung																
Persistentes negatives Faktum (Ablehnung)																
begründet durch Verstoß gegen ein Qualitätskriterium	j	j	j	j	j	j	j	j	n	n	n	n	n	n	n	n
abgelehnt von menschl. Experten	j	j	j	j	n	n	n	n	j	j	j	j	n	n	n	n
Aktuelles positives Faktum (Akzeptanz)																
akzeptiert von menschl. Experten	j	j	n	n	j	j	n	n	j	j	n	n	j	j	n	n
hohe Gesamtbewertung	j	n	j	n	j	n	j	n	j	n	j	n	j	n	j	n
Aktionen																
entferne persistentes Faktum					×	×			×	×			×	×		
entferne aktuelles Faktum			×	×			×	×			×	×				×
menschl. Experte entscheidet	×	×													×	

Tabelle 11.4: Entscheidungstabelle für die Behandlung von Widersprüchen zwischen Akzeptanz und Ablehnung

Diese Belegung der Entscheidungstabelle wurde gewählt, um folgendes Verhalten zu erreichen: Bei Ablehnungen durch Verstoß gegen ein Qualitätskriterium überstimmt die Ablehnung die Akzeptanz, solange der menschliche Experte nicht auch seine Akzeptanz ausgedrückt hat. Das Ersetzen einer solchen Ablehnung durch eine Akzeptanz bedeutet dann zugleich, dass die Einhaltung des Qualitätskriteriums weiter zu gewährleisten ist. Ggf. müssen dazu unter Mitwirkung des menschlichen Experten bisher akzeptierte Fakten abgelehnt werden. Kann diese Bedingung nicht erfüllt werden, ist ein Entfernen der Ablehnung nicht möglich, stattdessen muss die Akzeptanz verworfen werden. Wurde das negative Faktum nicht durch Verstoß gegen ein Qualitätskriterium begründet, ist die Meinung des menschlichen Experten ausschlaggebend. Die zu dem spätere Zeitpunkt getroffene Entscheidung überwiegt dabei die frühere Entscheidung. Liegt keine Expertenentscheidung vor, wird entweder der menschliche Experte um eine Entscheidung gebeten oder das aktuelle Faktum entfernt.

Nicht-Standardbegründungen – z.B. durch den menschlichen Experten angegebene natürlichsprachige Begründungen – können nicht automatisch ausgewertet werden. Zur Vereinfachung werden diese Begründungen weiterhin als gültig angenommen. Bei ausreichenden Indikatoren gegen das derart begründete Faktum kann mit der Entscheidung durch den menschlichen Experten auch eine Überprüfung der Begründung einhergehen.

11.2.2.4 Überprüfung bei zusätzlicher Betrachtung der durch die Hypothesen/Fakten implizierten Beziehungen

Eine Hypothese wird mit einer entsprechenden Gesamtbewertung sowie positiven Ergebnissen der qualitativen Überprüfung für einzelne Hypothesen, der Menge der Hypothesen sowie der zusätzlichen Überprüfung von Faktenmenge und vorhandenem Integrationswissen Faktum. Trotz der dargestellten Überprüfungen kann es Widersprüche mit den Modelleigenschaften der Thesaurusföderation geben. Denn alle bisherigen Überprüfungen haben nur solche Beziehungen berücksichtigt, die durch Hypothesen und Fakten *explizit* ausgedrückt werden. Es gilt also zusätzlich diese implizierten Beziehungen zu berücksichtigen.

11.2.2.4.1 Widersprüche gegen die Invarianten Für Widersprüche gegen die Invarianten durch implizierte Beziehungen gibt es zwei mögliche Ursachen, die eine unterschiedliche Konfliktbehandlung erfordern:

Widersprüchliche Inter-Thesaurus-Beziehungen: Die Widersprüche werden ausschließlich durch (explizite und implizite) *Inter*-Thesaurus-Beziehungen erzeugt, vgl. etwa Abbildung 11.3.

Die Inter-Thesaurus-Beziehungen können so modifiziert werden, dass der Konflikt aufgelöst werden kann. Dabei muss zusätzlich zu dem in den vorangegangenen Abschnitten beschriebenen Vorgehen berücksichtigt werden, dass implizierte Beziehungen nicht alleine entfernt werden können, sondern nur gemeinsam mit den expliziten Beziehungen, die diese implizieren.

Widersprüchliche Inter- und Intra-Thesaurus-Beziehungen: Die Widersprüche entstehen durch eine Beteiligung von (expliziten und impliziten) *Inter*-Thesaurus-Beziehungen sowie von *Intra*-Thesaurus-Beziehungen.

Im diesem Fall können zwar die Inter-Thesaurus-Beziehungen modifiziert werden, aufgrund der Autonomie der Komponententhesauri nicht aber die Intra-Thesaurus-Beziehungen. Wird im Kontext der Föderation jedoch eine andere Sichtweise als die im Komponententhesaurus ausgedrückte bevorzugt, ist eine entsprechende Konfliktmarkierung, vgl. Abschnitt 6.3.4, S. 88f, erforderlich.

Die Kriterien Q3' bis Q8' gelten entsprechend auch für den Fall der zusätzlichen Berücksichtigung von implizierten Kanten und werden als Q3" bis Q8" notiert.

Prinzipiell kann zur Auflösung eines Verstoßes gegen Q3" wie bei der Auflösung bei Verstoß gegen Q3' vorgegangen werden. Es ist aber zusätzlich zu berücksichtigen, dass eine Entscheidung, die implizierte Kante zu entfernen, bedeutet, explizite Kanten zu entfernen (bzw. durch $\mathcal{VB}$-Kanten zu ersetzen), so dass anschließend diese Kante nicht mehr impliziert wird.

Beispiel 11.5 *Ein Beispiel für einen Verstoß gegen Q3" ist in Abbildung 11.3 dargestellt. Um die implizierte Kante* (G.land conservation, $\mathcal{BS}$, A.land management) *zu beseitigen, könnte die*

*explizite Kante (*V.land management, $\mathcal{BS}$, G.land conservation) *wie auch die explizite Kante (*V.land management, $\mathcal{BS}$, A.land management) *entfernt werden. Die Auswahl muss vom menschlichen Experten getroffen werden. Die ausgewählte zu entfernende explizite Kante sowie die implizierte Kante werden dann als abgelehnte Beziehungen mit der Ablehnungsbegründung „Verstoß gegen Q3"" gespeichert.*

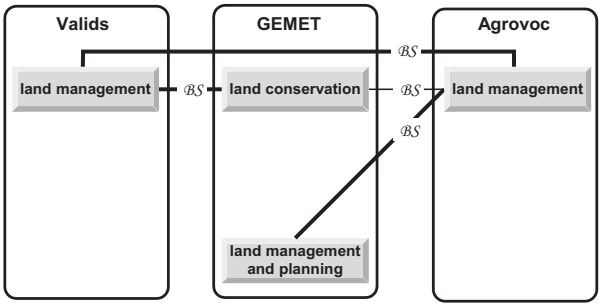

Abbildung 11.3: Verstoß gegen Q3", da A.land management explizit mit G.land management and planning und implizit mit G.land conservation äquivalent gesetzt wird.

Bei einer Verletzung von Q3" und Q4" sind aufgrund der Definition dieser Kriterien anhand von $\mathcal{BS}$- und $\mathcal{BK}$-Kanten, die nicht durch Intra-Thesaurus-Beziehungen impliziert werden können[5], ausschließlich Inter-Thesaurus-Beziehungen beteiligt. Hingegen können bei einer Verletzung von Q5" bis Q8" auch Intra-Thesaurus-Beziehungen Mitverursacher der Verletzung sein. Ist dies nicht der Fall (so ist z.B. in Abbildung 6.11 ein Verstoß gegen Q5" mit ausschließlich Inter-Thesaurus-Beziehungen als Verursacher dargestellt), kann entsprechend Q5' bis Q8' vorgegangen werden, wobei ggf. implizierte Beziehungen entsprechend der bei Verstoß gegen Q3' beschriebenen Vorgehensweise entfernt werden. Der Verstoß wird also beseitigt, indem auf Inter-Thesaurus-Beziehungen verzichtet wird.

Sind hingegen auch Intra-Thesaurus-Beziehungen Mitverursacher des Verstoßes gilt es, zusätzlich zu entscheiden, ob im Rahmen der Thesaurusföderation den Intra-Thesaurus-Beziehungen oder den Inter-Thesaurus-Beziehungen ein stärkeres Gewicht gegeben werden soll. Diese Entscheidung muss ebenfalls durch den menschlichen Integrationsexperten getroffen werden. Im ersten Fall sind die Inter-Thesaurus-Beziehungen zu entfernen, die im Widerspruch zu den Intra-Thesaurus-Beziehungen stehen. Im zweiten Fall kann eine Entfernung der Intra-Thesaurus-Beziehungen aufgrund der Autonomie der Komponententhesauri nicht stattfinden. Stattdessen erfolgt eine Markierung der Intra-Thesaurus-Beziehungen. Abstrakte Beispiele für solche Markierungen wurden bereits für Q5" in Abschnitt 6.3.7.1, für Q6" in Abschnitt 6.3.7.2, für Q7" in Abschnitt 6.3.6.3 und für Q8" in Abschnitt 6.3.6.5, aufgeführt. Dort wird auch ersichtlich, dass die Menge der Konfliktverursacher automatisch festgestellt werden kann, aus dieser im Allgemeinen aber nur mit Unterstützung des menschlichen Experten die Mengen r_1, r_2 und s definiert werden können. Das Expertenwissen wurde jedoch bereits durch die Entscheidung eingebracht, dass die Inter-Thesaurus-Beziehungen den Intra-Thesaurus-Beziehungen gegenüber stärker bewertet werden. Somit kann sowohl in r_1 als auch in s die Intra-Thesaurus-Beziehung aufgenommen werden. Sind jedoch mehrere Intra-Thesaurus-Beziehungen beteiligt, muss wiederum der menschliche Integrationsexperte entscheiden, welche Intra-Thesaurus-Beziehung(en) in r_1 und s aufgenommen werden sollen. Wurden r_1 und s spezifiziert, kann r_2 automatisch

[5]BS- und BK-Kanten innerhalb eines Thesaurus sind zwischen Nicht-Deskriptoren und Deskriptoren definiert und können daher keine Inter-Thesaurus-Beziehungen zwischen Deskriptoren implizieren

abgeleitet werden, wie dies bereits in den entsprechenden Abschnitten des Kapitels 6 anhand der Konfliktmarkierungen ersichtlich wird.

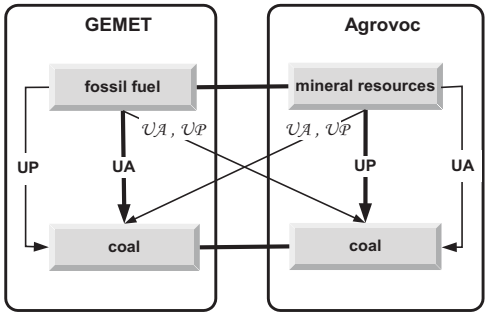

Abbildung 11.4: Beispiel eines Verstoßes gegen Q5" unter Beteiligung von Intra-Thesaurus-Beziehungen

Beispiel 11.6 *Abbildung 11.4 zeigt ein Beispiel für einen Verstoß gegen Q5". Die durch die Deskriptoren* fossil fuel *in GEMET und* mineral resources *in AGROVOC repräsentierten Begriffe wurden aufgrund mehrerer identischer Unterbegriffe durch eine Inter-Thesaurus-Äquivalenzbeziehung verbunden. Ebenso wurde* coal *mit dem gleichlautenden Pendant äquivalent gesetzt. Jedoch ist* coal *in GEMET Abstraktionsunterbegriff von* fossil fuel *und Bestandsunterbegriff von* mineral resources *in AGROVOC. Zwischen den Deskriptoren bestehen somit verursacht durch die implizierten Beziehungen sowohl Abstraktions- als auch Bestandsbeziehungen. Nach Betrachtung der weiteren Unterbegriffe* (lignite, peat, petroleum)*, die verschiedene fossile Energieträger bezeichnen aber keine weiteren Mineral-Ressourcen, wird entschieden, die GEMET-Sichtweise als die für die Föderation relevante beizubehalten. Das bedeutet, die Konfliktmarkierungsmengen sehen wie folgt aus:*

$$
\begin{aligned}
v &= \{(\textit{fossil fuel}, \textit{mineral resources}), (\textit{coal}, \textit{coal}), (\textit{fossil fuel}, \textit{coal}),\\
&\quad (\textit{mineral resources}, \textit{coal})\}\\
r_1 &= \{(\textit{fossil fuel}, \textit{coal})\}\\
r_2 &= \{(\textit{mineral resources}, \textit{coal})\}\\
s &= \{(\textit{mineral resources}, \textit{coal})\}
\end{aligned}
$$

Verstöße gegen Q6", Q7" und Q8" können entsprechend Verstößen gegen Q5" behandelt werden.

Beispiel 11.7 *Ein Beispiel für einen Verstoß gegen Q7" ist in Abbildung 11.5 dargestellt. Verursacher sind sowohl Beziehungen zu einem Ergänzenden Begriff* (conventional fuel) *als auch die Intra-Thesaurus-Beziehung* (fuel, *UA,* diesel fuel)*. Im Zusammenhang mit der Föderation kann die Intra-Thesaurus-Beziehung als redundant markiert werden, d.h.*

$$
\begin{aligned}
v &= \{(\textit{fuel}, \textit{fuel}), (\textit{fuel}, \textit{conventional fuel}), (\textit{conventional fuel}, \textit{diesel fuel}),\\
&\quad (\textit{fuel}, \textit{diesel fuel})\}\\
r_1 &= \{(\textit{fuel}, \textit{conventional fuel}), (\textit{conventional fuel}, \textit{diesel fuel})\}\\
r_2 &= \{(\textit{fuel}, \textit{diesel fuel})\}\\
s &= \{(\textit{fuel}, \textit{diesel fuel})\}
\end{aligned}
$$

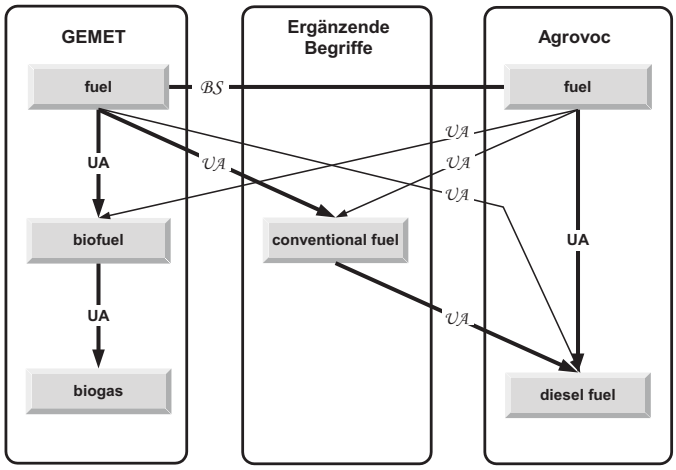

Abbildung 11.5: Beispiel eines Verstoßes gegen Q7" unter Beteiligung von Intra-Thesaurus-Beziehungen und einem Ergänzenden Begriff

Bei der zusätzlichen Betrachtung der implizierten Fakten wird, da es sich hier um die am weitest gehende Prüfung handelt, auch die Einhaltung aller bisher nicht betrachteten Invarianten überprüft. Somit werden zwei weitere Kriterien geprüft (vgl. Abschnitte 6.4.3.7, S. 107 und 6.3.6.3, S. 92f sowie 6.4.3.7, S. 107 und 6.3.7.3, S. 98):

Q9": Verschiedene Abstraktionspfade von einem Begriff x zu einem Begriff y sollen gleich lang sein.

Q10": Wenn innerhalb eines Komponententhesaurus alle Unterbegriffe über denselben Hierarchierelationstypen verbunden sind, soll durch Inter-Thesaurus-Beziehungen diese Einheitlichkeit nicht gestört werden.

Beispiele für Verstöße gegen diese Kriterien wurden bereits in den Abschnitten 6.3.6.3 und 6.3.7.3 aufgeführt. Sowohl bei einem Verstoß gegen Q9" als auch gegen Q10" kann versucht werden, durch das Einführen eines Ergänzenden Begriffs den Verstoß aufzulösen. Die möglichen Positionen innerhalb des Thesaurusgraphen können automatisch vorgeschlagen werden, die Festlegung der Bezeichner kann jedoch nur durch den menschlichen Experten getroffen werden. Die Positionsbestimmung ist trivial: Bei Verstoß gegen Q9" kommen alle Positionen zwischen Knoten des kürzeren Pfades in Frage. Die Anzahl der einzufügenden Knoten entspricht der Differenz der Pfadlängen. Bei Verstoß gegen Q10" werden Ergänzende Begriffe so vorgeschlagen, dass nach Einführen dieser die Einheitlichkeit der Inter-Thesaurus-Hierarchiebeziehungen mit den Intra-Thesaurus-Hierarchiebeziehungen wieder hergestellt ist (vgl. Abbildung 11.6). Das bedeutet, die Ergänzenden Begriffe werden als Mittelglied in die abweichenden Inter-Thesaurus-Beziehungen eingeführt. Die ursprünglichen Intra-Thesaurus-Beziehungen werden als redundant markiert bzw. die Inter-Thesaurus-Beziehungen entsprechend mit dem Ergänzenden Begriff etabliert.

Der menschliche Experte entscheidet in beiden Fällen, ob Ergänzende Begriffe sinnvoll sind.

Beispiel 11.8 *Abbildung 11.6 zeigt ein Beispiel für einheitliche Hierarchierelationstypen (Abstraktionsbeziehungen) ausgehend von dem Deskriptor* water management *innerhalb GEMET, die urprünglich mit andersartigen Inter-Thesaurus-Relationen (Bestandsrelationen) zu den Begriffen* water storage *und* transfer of water *kombiniert wurden. Durch Einführen des Begriffes*

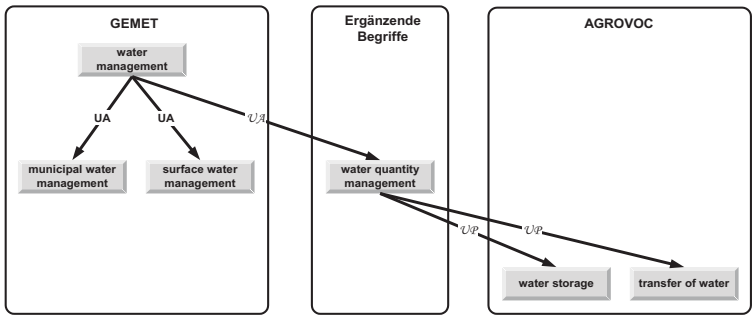

Abbildung 11.6: Unterschiedliche Typen von Hierarchierelationen bei Schwesterknoten werden durch Einführen eines Ergänzenden Begriffs vereinheitlicht.

water quantity management *kann die Einheitlichkeit der Hierarchierelationstypen wieder hergestellt werden.*

11.2.2.4.2 Widersprüche bzgl. Akzeptanz und Ablehnung Wie bereits in Abschnitt 11.2.2.3.2 für Fakten dargelegt, kann es auch zwischen implizierten Beziehungen und Fakten Widersprüche hinichtlich der Akzeptanz bzw. Ablehnung geben. Da die Menge der durch akzeptierte Fakten implizierten (akzeptierten) Beziehungen eine Obermenge der durch abgelehnte Fakten implizierten (abgelehnten) Beziehungen ist[6], genügt es, diese zu berechnen. Bei so aufgedeckten Widersprüchen kann prinzipiell wie in Abschnitt 11.2.2.3.2 dargestellt vorgegangen werden. Es ist aber zusätzlich zu berücksichtigen, dass implizierte Beziehungen nur durch Entfernen eines Faktums, das zur Implikation dieser Beziehung beiträgt, entfernt werden können. Die durch dieses Faktum ausgedrückte Beziehung sowie die entfernte implizierte Beziehung werden als abgelehnte Fakten mit der Begründung „Widerspruch der implizierten Kante zu abgelehntem Faktum“ gespeichert.

11.2.2.4.3 Übersicht der Kriterien und ihrer Herleitung Wie wir gezeigt haben, werden alle Qualitätskriterien aus den Modelleigenschaften der Thesaurusföderation hergeleitet. Um die Zusammenhänge zwischen den Modelleigenschaften und den Kriterien noch einmal zu verdeutlichen, zeigt Tabelle 11.5 zusammenfassend, aus welcher Modelleigenschaft welches Kriterium abgeleitet wurde.

11.3 Erweiterungen und Veränderungen am Integrationswissen

Wurden vom Moderator Fakten erzeugt, ist es Aufgabe des Ausführungsagenten, die Implikationen dieser Fakten durch Erweiterungen und Veränderungen am Integrationswissen festzuschreiben. Bereits in Abschnitt 7.3.8, S. 127, haben wir festgestellt, dass der Status eines Faktums als akzeptiert bzw. abgelehnt sowie das Vorhandensein der durch das Faktum ausgedrückten Beziehung unterschiedliche Modifikationen des Integrationswissens erfordern. In den Abschnitten 11.3.1 bis 11.3.4 betrachten wir die erforderlichen Aktionen genauer.

[6] Akzeptierte und abgelehnte Beziehungen implizieren gleichartige Beziehungen zwischen allen Deskriptoren der Föderierten Begriffe (vgl. z.B. Abbildung 6.6 mitte rechts, S. 85), akzeptierte Beziehungen können zusätzliche Abstraktions-/Bestands-/Assoziationsbeziehungen implizieren, die von den anderen Deskriptoren des Föderierten Begriffs übertragen werden (vgl. z.B. Abbildung 6.6 mitte links, S. 85).

Modell-eigenschaft	abgeleitetes Kriterium		
	Qi	Qi'	Qi"
Richtiger Einsatz der thesaurus-verbindenden Kante	Q1, Q2, Q3, Q4	Q3', Q4'	Q3", Q4"
Einzigartigkeit einer Kante	Q5	Q5'	Q5"
Schwesternassoziationen	Q6	Q6'	Q6"
Redundanzfreiheit der Inter-Thesaurus-Abstraktionspfade		Q7'	Q7"
Zyklenfreiheit der Inter-Thesaurus-Hierarchiepfade		Q8'	Q8"
Freiheit von Abstraktionsniveaudifferenzen			Q9"
Beibehaltung des Hierarchierelationstyps (keine AP oder PA-Schwestern)			Q10"

Tabelle 11.5: Aus den Modelleigenschaften der Thesaurusföderation hergeleitete Kriterien für die qualitative Überprüfung von Hypothesen und Fakten

Vorweg sei erwähnt, dass jedes Faktum, dessen Implikationen in das Integrationswissen übernommen wurden, vom Blackboard gelöscht und in das Faktenarchiv transferiert wird. Innerhalb dieses Faktenarchivs ersetzt es dort ggf. vorhandene ältere Fakten des gleichen Behauptungstyps über die gleichen Begriffe und Beziehungen unabhängig davon, ob das ältere Faktum und das neue Faktum den gleichen Bewertungsstatus (akzeptiert oder abgelehnt) besitzen oder nicht. Somit wird sichergestellt, dass innerhalb des Faktenarchivs ausschließlich solche Fakten enthalten sind, deren Implikationen im aktuellen Integrationswissen widergespiegelt werden.

Für eine akzeptierte Kante e im Faktenarchiv A notieren wir $e \in A$, für eine abgelehnte Kante $\overline{e} \in A$.

11.3.1 Einfügen von Inter-Thesaurus-Beziehungen

Akzeptierte Fakten über Inter-Thesaurus-Beziehungen führen zur Aufnahme dieser Beziehungen in das Integrationswissen. Wurden diese Beziehungen im Thesaurusföderationsgraphen bisher noch nicht ausgedrückt, ist die Kante dort zu ergänzen. Wurde die Beziehung bereits durch eine implizierte Kante ausgedrückt, ist also im Faktenarchiv noch kein Faktum mit der entsprechenden Kante vorhanden, im Föderationsgraphen aber bereits die Kante enthalten, wird der Status der Kante von einer bisher implizierten zu einer nun expliziten Kante modifiziert. Nur falls die Kante bisher nicht bereits impliziert war, gilt es zusätzlich, alle implizierten Kanten zu bestimmen und diese als solche ebenfalls im Föderationsgraphen zu ergänzen.

Der komplexeste Fall ist das Einfügen einer Inter-Thesaurus-Äquivalenzbeziehung ($\mathcal{BS}$-Kante). Wir betrachten diesen Fall daher exemplarisch detaillierter.

Abbildung 11.7 zeigt an einem Beispiel, dass eine neue Inter-Thesaurus-Äquivalenzbeziehung eine ganze Reihe von implizierten Kanten bedeuten kann. Auf implizierte $\mathcal{BS}$-Kanten wurde der besseren Übersicht wegen in dieser Abbildung verzichtet.

Der Algorithmus *FügeBSKanteEin* zeigt das Einfügen einer $\mathcal{BS}$-Kante sowie aller implizierten Kanten: Ist die Kante bereits im Faktenarchiv vorhanden, sind keine weiteren Aktionen erforderlich. Ist dies nicht der Fall, wird das Faktenarchiv ergänzt[7]. Der Sub-Algorithmus *FügeBSKanteInFöderationsgraphEin* überprüft, ob die $\mathcal{BS}$-Kante bereits als (implizite) Kante im Föderationsgraphen vorhanden ist und ergänzt die Kante erforderlichenfalls[8]. Sollte die Kante bereits vorhanden sein, wurde durch das Eintragen in das Faktenarchiv bereits der Status von impli-

[7] Auf eine formale Darstellung dieses trivialen Algorithmus *FügeKanteInFaktenarchivEin* wird hier verzichtet.
[8] Aufgrund Definition 6.22, S. 103, wird die Kante nur in eine Richtung eingefügt.

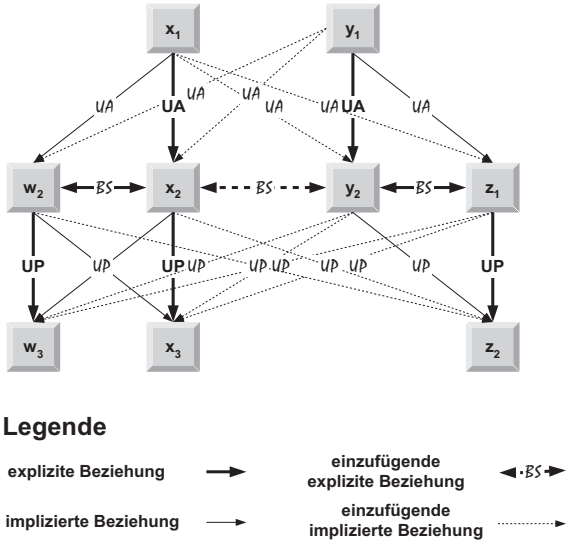

Abbildung 11.7: Implizierte Hierarchiebeziehungen

zit auf explizit geändert. Weitere Aktionen sind nicht erforderlich. Existierte die Kante zuvor jedoch weder als explizite noch als implizite Kante, sind nun zusätzlich alle impliziten Kanten zu bestimmen und zu ergänzen. Dazu werden alle zu den Knoten der neuen $\mathcal{BS}$-Kante mit einer $\mathcal{BS}$-Kante verbundenen Knoten, also die Äquivalenzklasse S der Knoten der einzufügenden Kanten, bestimmt. Anschließend wird von jedem Knoten der Äquivalenzklasse S zu jedem anderen Knoten eine $\mathcal{BS}$-Kante etabliert[9], um die implizierten $\mathcal{BS}$-Kanten zu erfassen. Implizierte Kanten zu Ober- und Unterbegriffen sowie zu Verwandten Begriffen werden bestimmt, indem zu allen Knoten der Äquivalenzklasse S die entsprechenden Unter-/Oberbegriffe bzw. Verwandten Begriffe ermittelt werden und bisher nicht vorhandene Kanten ergänzt werden.

Algorithmus *FügeBSKanteEin*(F, e, A)
Eingabe: Föderationsgraph $F = (N, E)$, eine BS-Kante $e = (x, \mathcal{BS}, y)$, Faktenarchiv A
Ausgabe: modifizierter Föderationsgraph F

1. (∗ Prüfen, ob Kante bereits im Faktenarchiv ∗)
2. **if** $e \in A$
3. **then return** F (∗ Keine Aktion erforderlich, fertig ∗)
4. **else** FügeKanteInFaktenarchivEin$(A, (x, \mathcal{BS}, y)$
5. (∗ Prüfen, ob Kante im Föderationsgraph vorhanden und ggf. ergänzen ∗)
6. **if** (FügeBSKanteInFöderationsgraphEin(F, x, y) = „Kante vorhanden“))
7. **then** (∗ zuvor implizite Kante ist nun explizite Kante ∗)
8. (∗ somit brauchen weitere implizite Kanten nicht berechnet werden ∗)
9. **return** F (∗ Fertig ∗)
10. **else** (∗ Implizierte Kanten bestimmen und ergänzen ∗)
11.
12. (∗ Bestimme alle Knoten, die mit x, y in einer Äquivalenzklasse sind ∗)
13. $S \leftarrow \{x, y\} \cup \{a : (a, \mathcal{BS}, x) \in E \vee (x, \mathcal{BS}, a) \in E \vee (a, \mathcal{BS}, y) \in E \vee (y, \mathcal{BS}, a) \in E\}$
14. (∗ Implizierte $\mathcal{BS}$-Kanten einfügen ∗)
15. (∗ $\star a$ bezeichnet eindeutigen Identifikator von a, s. Definition 6.8, S. 83 ∗)
16. **for all** $(a, b) \in S \times S, \star a < \star b$

[9] wiederum unter Berücksichtigung der aus Definition 6.22 resultierenden Einschränkung

```
17.                 if  (a, ℬ𝒮, b) ∉ E
18.                   then FügeBSKanteInFöderationsgraphEin(F, a, b)
19.
20.         (∗ Bestimme alle 𝒰𝒜-, UA- und UA-Unterknoten ∗)
21.         U ← {b : ((a, 𝒰𝒜, b) ∈ E, a ∈ S) ∨ ((a, UA, b) ∈ E, a ∈ S) ∨ ((a, UA, b) ∈ E, a ∈ S)}
22.         (∗ 𝒰𝒜-, UA- und UA-Kanten ergänzen ∗)
23.         for all (a, u) ∈ S × U
24.                 if (thesaurus(a) = thesaurus(u) ∧ (a, UA, u) ∉ E)
25.                   then (∗ implizierte Intra-Thesaurus-Kanten einfügen ∗)
26.                         E ← E ∪ (a, UA, u)
27.                   else  if (thesaurus(a) ≠ thesaurus(u) ∧ (a, 𝒰𝒜, u) ∉ E)
28.                           then (∗ implizierte Inter-Thesaurus-Kanten einfügen ∗)
29.                                 E ← E ∪ (a, 𝒰𝒜, u)
30.
31.         ... (∗ entsprechend für 𝒪𝒜-, OA-, 𝒰𝒫-, UP-, 𝒪𝒫-, OP-, 𝒱ℬ-, VB-Kanten ∗)
32.         return F
```

Algorithmus *FügeBSKanteInFöderationsgraphEin*(F, a, b)
Eingabe: Föderationsgraph $F = (N, E)$, Deskriptorknoten a, b
Ausgabe: modifizierter Föderationsgraph F, Status neue / vorhandene Kante

```
1.  (∗ ungerichtete ℬ𝒮-Kante einfügen ∗)
2.  if (⋆a ≥ ⋆b)
3.    then e ← (b, ℬ𝒮, a)
4.    else  e ← (a, ℬ𝒮, b)
5.  if (e ∈ E)
6.    then return „Kante vorhanden“
7.    else  E ← E ∪ e
8.          return „Kante neu“
```

11.3.2 Einfügen von Ergänzenden Begriffen

Der Unterschied beim Einfügen von Ergänzenden Begriffen im Vergleich zu allen Operationen, die sich auf Komponententhesauri beziehen, besteht darin, dass nicht nur Kanten und Konfliktmarkierungen in die Föderation eingebracht werden, sondern zusätzlich Knoten. Diese Knoten werden in den Thesaurus der Ergänzenden Begriffe eingefügt. Des weiteren ist das Einfügen eines Ergänzenden Begriffs aufgrund Definition 6.5, S. 82f, immer auch mit dem Einfügen von mindestens zwei Kanten verbunden. Dies kann wie im vorangegangenen Abschnitt 11.3.1 beschrieben geschehen. Falls der Ergänzende Begriff zum Auflösen eines AP- oder PA-Schwesternkonfliktes (unterschiedliche Typen von Hierarchierelationen bei Schwesterknoten) eingefügt wurde, werden die so redundant gewordenen Kanten entfernt (im Beispiel in Abbildung 11.6, S. 205, wurde die zuvor vorhandene $\mathcal{UP}$-Kante zwischen den Knoten *G.water management* und *A.water storage* entfernt (vgl. dazu Abschnitt 11.3.3)). Ebenso werden für den Fall des „Verlängerns“ von Pfaden durch Einfügen von Ergänzenden Begriffen in einen Pfad zum Auflösen von Abstraktionsniveaudifferenzen redundant werdende Kanten entfernt oder, falls dies nicht möglich ist, als redundant markiert.

11.3.3 Entfernen von Inter-Thesaurus-Beziehungen

Abgelehnte Hypothesen, also negative Fakten, über Inter-Thesaurus-Beziehungen bedeuten, dass die entsprechenden Inter-Thesaurus-Beziehungen, falls vorhanden, aus dem Integrationswissen entfernt werden sollen. Zusätzlich zu der explizit aufgeführten Beziehung, deren Identifikation und Entfernung trivial ist, gilt es auch die implizierten Beziehungen zu entfernen sowie die Konfliktmenge zu aktualisieren.

Implizierte Kanten können nicht unabhängig von der oder den Kanten enfernt werden, die diese Kante implizieren. Daher sind zum Entfernen einer impliziten Kante alle expliziten Kanten aufzufinden, die diese implizieren, und zu entfernen.

11.3.3.1 Entfernen von Kanten

Der komplexeste Fall des Entfernens einer Kante betrifft wiederum Inter-Thesaurus-Äquivalenzbeziehungen. Wir zeigen für diesen Fall detailliert, wie neben der expliziten Kante auch alle implizierten Kanten entfernt werden können.

Grundsätzliche Idee des Algorithmus *EntferneBSKante* ist, alle implizierten Beziehungen, die von einem Knoten, der mit einer $\mathcal{BS}$-Kante mit einem Knoten der zu entfernenden Kante verbunden ist, zu entfernen. Ebenfalls werden alle $\mathcal{BS}$-Kanten zwischen einem Knoten der zu entfernenden Kante und allen anderen Knoten entfernt. Durch das erneute Einfügen der $\mathcal{BS}$-Kanten im Faktenarchiv und der erneuten Berechnung aller implizierten Beziehungen wird der neue Gesamtzustand bestimmt. Ein zentraler Teil des Algorithmus kann somit zurückgeführt werden auf den bereits vorgestellten Algorithmus *FügeBSKanteInFöderationsgraphEin*. Dem Nachteil des nicht geringst möglichen Aufwandes steht der Vorteil der Einfachheit durch Zurückführung auf bekannte Algorithmen entgegen. Aufgrund der üblicherweise geringen Anzahl von zu entfernenden Kanten ist der Aufwand vernachlässigbar, so dass der Einfachheit größeres Gewicht gegeben wird.

Algorithmus *EntferneBSKante*(F, e, A)
Eingabe: Föderationsgraph $F = (N, E)$, eine BS-Kante $e = (x, \mathcal{BS}, y)$, Faktenarchiv A
Ausgabe: modifizierter Föderationsgraph F, Ergebnismeldung

1. (∗ Prüfen, ob Kante bisher implizierte Kante war (kein Faktum im Faktenarchiv) ∗)
2. **if** $(e \notin A)$
3. **then return** „implizierte Kante kann nicht entfernt werden“
4. **else** (∗ explizite Kante entfernen ∗)
5.
6. (∗ Bestimme alle Knoten, die bisher mit x, y in einer Äquivalenzklasse sind ∗)
7. $S \leftarrow \{x, y\} \cup \{a : (a, \mathcal{BS}, x) \in E \vee (x, \mathcal{BS}, a) \in E \vee (a, \mathcal{BS}, y) \in E \vee (y, \mathcal{BS}, a) \in E\}$
8. (∗ Kante im Föderationsgraph entfernen ∗)
9. $E \leftarrow E - (x, \mathcal{BS}, y)$
10. (∗ Faktenarchiv aktualisieren (positives Faktum durch negatives ersetzen) ∗)
11. FügeAbgelehnteKanteInFaktenarchivEin$(A, (x, \mathcal{BS}, y))$
12.
13. (∗ Alle BS-Kanten innerhalb der Äquivalenzklasse entfernen ∗)
14. **for all** $(a, b) \in S \times S, \star a < \star b$
15. E ← E − $(a, \mathcal{BS}, b)$
16.
17. (∗ Alle nicht expliziten Kanten zu Knoten der Äquivalenzklasse entfernen ∗)

```
18.     (∗ Bestimme alle UA-Unterknoten ∗)
19.     U ← {b : ((a, UA, b) ∈ E, a ∈ S)}
20.     for all (a, u) ∈ S × U
21.         if (a, UA, u) ∉ A
22.           then E ← E − (a, UA, u)
23.
24.     (∗ entsprechend für UA, OA-, OA-, UP-, UP-, OP-, OP-, VB-, VB-Kanten ∗)
25.      ...
26.
27.     (∗ BS-Kanten aus Faktenarchiv (mit allen implizierten Kanten) wieder einfügen ∗)
28.     for all (a, b) ∈ S × S, ⋆a < ⋆b
29.         if ((a, BS, b) ∈ A
30.           then FügeBSKanteEin(F, (a, BS, b), A)
31.     if ((x, BS, y) ∈ E) ∨ ((y, BS, x) ∈ E)
32.       then return „Kante weiterhin impliziert“
33.       else return „Kante erfolgreich entfernt“
```

Es kann nicht ausgeschlossen werden, dass eine Kante, die zuvor explizit war, nach Entfernen des Faktums weiterhin durch andere Kanten impliziert wird. In Abbildung 11.7 könnte z.B. eine explizite $\mathcal{BS}$-Kante zwischen x_2 und z_1 weiter die enfernte explizite Kante $(x_2, \mathcal{BS}, y_2)$ implizieren. Um auch diese implizierte Kante zu entfernen gilt es, zusätzlich eine oder mehrere explizite Kanten zu entfernen, die diese Kante implizieren. Im Falle einer zu entfernenden implizierten $\mathcal{BS}$-Kante zwischen zwei Knoten x_2, y_2 bedeutet dies, solche Kanten zu entfernen, die dazu beitragen, dass x_2, y_2 Elemente einer Äquivalenzklasse sind. Falls es verschiedene Möglichkeiten gibt, disjunkte Äquivalenklassen herzustellen, wird die Auswahl der zu entfernenden Kanten dem menschlichen Experten überlassen.

11.3.3.2 Aktualisieren der Konfliktmenge

Das Ergänzen oder Entfernen von Kanten muss auch bei den Konfliktmarkierungen berücksichtigt werden. Die Konsequenzen beim Hinzufügen von Kanten wurden bereits in Abschnitt 11.2.2.4 ausführlich behandelt. Wir betrachten daher an dieser Stelle nur den Fall des Entfernens einer Kante näher.

Beim Entfernen einer Kante e zwischen den Knoten x, y müssen alle Konflikte k, deren Konfliktverursachermenge $k.v$ die Knoten x, y enthalten, untersucht werden. Da sowohl explizite als auch implizite Beziehungen Konflikte verursachen können, ist eine solche Untersuchung in beiden Fällen erforderlich.[10] Enthält ein Konflikt k die Knoten x, y der entfernten Kante e, sind die erforderlichen Aktionen abhängig vom Typ des Konflikts. Anhand zweier Konflikttypen zeigen wir dies exemplarisch. Aufgrund der kompakteren Erfassung der Konflikte mittels des in Abschnitt 6.3 eingeführten formalen Modells nutzen die folgenden Algorithmen dieses anstelle des Graphenmodells. Eine Transformation ist mittels der Herleitung des Graphenmodells aus dem formalen Modell in Abschnitt 6.4, S. 100ff, möglich.

In Algorithmus *BeziehungstypdifferenzTest* werden die Aktionen beschrieben, die erforderlich

[10] Beim Entfernen von expliziten Beziehungen durch Algorithmus *EntferneBSKante* werden zwar implizite Beziehungen entfernt, die Menge der entfernten implizierten Beziehungen liegt anschließend jedoch nicht vor. Um die Konfliktmenge zu aktualisieren, können die entfernten implizierten Beziehungen jedoch als Differenz der vor und nach dem Entfernen vorhandenen implizierten Beziehungen (also solcher Beziehungen, die zwar als Kanten existieren, aber nicht im Faktenarchiv sind) berechnet werden.

sind, falls die Knoten x, y der entfernten Kante e als Tupel in der Konfliktverursachermenge eines Konflikts des Typs Beziehungstypdifferenz enthalten sind. Dieser Konflikttyp bedeutet, dass zwischen zwei Knoten mindestens zwei Beziehungen unterschiedlichen Typs etabliert sind, also mindestens zwei verschiedene Kanten die Knoten verbinden. Das Knotentupel wird aus der Konfliktverursachermenge entfernt. Anschließend wird die Anzahl der unterschiedlichen Kanten zwischen den Knoten festgestellt. Ist diese nun kleiner zwei, existiert kein Konflikt mehr und die Konfliktmarkierung kann entfernt werden.

Algorithmus *BeziehungstypdifferenzTest*(F, K, k, x, y)
Eingabe: Föderationsgraph $F = (N, E)$, Konfliktmenge K, Konflikt $k \in K$, Kanten x, y
Ausgabe: modifizierte Konfliktmenge

1. $(*$ Kanten aus Konfliktmarkierungsmengen entfernen $*)$
2. $k.v \leftarrow k.v - (x, y)$
3. $k.r_1 \leftarrow k.r_1 - (x, y)$
4. $k.r_2 \leftarrow k.r_2 - (x, y)$
5. $(*$ Prüfen, ob durch Entfernen der Kante Konflikt entfällt $*)$
6. **if** $|\{beztyp : (beztyp \in \{\mathcal{BS}, \mathcal{UA}, \mathcal{UP}, \mathcal{VB}, BS, UA, UP, VB, \mathsf{UA}, \mathsf{UP}, \mathsf{VB}\}) \wedge ((x, beztyp, y) \in E \vee (y, beztyp, x) \in E)\}| < 2$
7. **then** $K \leftarrow K - k$
8. **return** K

Ist eine entfernte Kante Element der Konfliktverursachermenge $k.v$ eines Abstraktionsredundanzkonfliktes k, wird durch Algorithmus *AbstraktionsredundanzTest* geprüft, ob die Konfliktmarkierung entfernt werden kann. Zu Beginn werden die Knoten x, y der entfernten Kante aus der Konfliktverursachermenge entfernt. Anschließend wird überprüft, ob die entfernte Kante verantwortlich für die direkte Abstraktionsbeziehung war (festgehalten in $k.r_1$, die im Widerspruch zu der durch $k.r_2$ beschriebenen indirekten Abstraktionsbeziehung steht). Ist dies der Fall, existiert der Konflikt nicht weiter und die Konfliktmarkierung kann entfernt werden. Ist dies nicht der Fall, ist das entfernte Knotentupel Teil des indirekten Pfades. Falls durch Entfernen aus diesem indirekten Pfad kein indirekter Abstraktionspfad mehr zwischen den in $k.r_1$ auch direkt durch eine Abstraktionskante verbundenen Knoten mehr besteht, existiert der Konflikt ebenfalls nicht länger und die Konfliktmarkierung wird entfernt.

Algorithmus *AbstraktionsredundanzTest*(F, K, k, x, y)
Eingabe: Föderationsgraph $F = (N, E)$, Konfliktmenge K, Konflikt $k \in K$, Kanten x, y
Ausgabe: modifizierte Konfliktmenge

1. $(*$ Kanten aus Konfliktmarkierungsmenge entfernen (vorerst nur Verursachermenge) $*)$
2. $k.v \leftarrow k.v - (x, y)$
3. $(*$ Prüfen, ob durch Entfernen der Kante Konflikt entfällt $*)$
4. **if** $(x, y) \in k.r_1 \vee (y, x) \in k.r_1$
5. **then** $(*$ zwischen x und y bestand $\mathcal{UA}$-Kante, die entfernt wurde $*)$
6. $K \leftarrow K - k$ $(*$ Konflikt aus Konfliktmenge entfernen $*)$
7. **else** Seien m, n die Knoten des Tupels $k.r_1$ mit $(m, \mathcal{UA}, n) \in E$
8. $k.r_2 \leftarrow k.r_2 - (x, y)$
9. **if** $(m \overset{A^\star_{k.r_2}}{\hookrightarrow} n) \wedge (|k.r_2| \geq 2)$
10. **then** $(*$ Es gibt weiter direkte Kante $*)$
11. $(*$ und Abstraktionspfad über Knoten in $k.r_2$ Pfad mit Länge ≥ 2 $*)$
12. $(*$ Konflikt kann nicht entfernt werden $*)$

13. **else** $K \leftarrow K - k$ ($*$ sonst entferne Konflikt $*$)
14. **return** K

11.3.4 Entfernen von Ergänzenden Begriffen

Wie bereits in Abschnitt 11.3.2 erläutert, besteht der wesentliche Unterschied zu anderen Operationen beim Einfügen bzw. Entfernen von Ergänzenden Begriffen darin, dass zusätzlich zu den Auswirkungen auf Kanten und Konfliktmarkierungen auch Knoten eingefügt oder entfernt werden. Neben dieser einfachen Operation kann – entsprechend wie in Abschnitt 11.3.2 dargestellt – vollständig auf bereits bekannte Algorithmen zurückgegriffen werden. Beim Entfernen eines Ergänzenden Begriffs sind dies das Entfernen aller Beziehungen bei gleichzeitiger Aktualisierung der Konfliktmengen.

11.4 Resümee

Wir haben in diesem Kapitel gezeigt, wie im Rahmen der Umsetzung der Lösungsstrategie das für die Erstellung der Thesaurusföderation notwendige Integrationswissen gewonnen wird. Dazu wurde gezeigt, wie unterschiedlichste Problemlösungsverfahren eingebracht werden können. Als Basis wurden Klassifizierungskriterien so entwickelt, dass die Lösungsverfahren den verschiedenen Teilphasen zugeordnet werden können. Exemplarisch wurden eine Reihe von Problemlösungsverfahren, die zum Teil aus der Literatur stammen und zum Teil in diesem Kapitel (weiter-) entwickelt wurden, entsprechend klassifiziert. Mit den vorgestellten Verfahren wurde bereits ein umfangreicher Grundstock zur Gewinnung des Integrationswissens bereitgestellt. Weitere Verfahren können aufgrund der generellen Offenheit der Blackboard-Architektur sowie mittels der Klassifizierungskriterien einfach eingebracht werden.

Von den Lösungsverfahren werden Hypothesen erzeugt, die Basis der Gewinnung des Integrationswissens sind. Dazu gilt es, aus den Hypothesen Fakten herzuleiten. Wir haben gezeigt, dass hierzu neben der Berechnung einer Gesamtbewertung insbesondere eine qualitative Überprüfung der Hypothesen erforderlich ist. Diese qualitative Überprüfungen konnten anhand der angestrebten Widerspruchsfreiheit gegen die Modelleigenschaften der Thesaurusföderation hergeleitet werden. Ebenso wurde unter Berücksichtigung der Ziele der qualitativen Überprüfung gezeigt, dass diese Überprüfungen unterschiedliche Ausschnitte der Hypothesenmenge und des bereits vorhandenen Integrationswissens betrachten müssen. Als Bestandteil der qualitativen Überprüfung wurde auch das Erkennen und ggf. Beseitigen oder Markieren von Konflikten vorgestellt.

Neben der Gewinnung des Integrationswissens ist weiterer wesentlicher Bestandteil der Realisierungsphase, dieses Integrationswissen festzuschreiben. Sowohl für positive als auch für negative Fakten haben wir Verfahren entwickelt, diese mit sämtlichen Implikationen festzuschreiben.

Aufbauend auf den vorangegangenen Phasen ist die hier vorgestellte Realsierungsphase somit als entscheidende Phase bei der Erstellung einer Thesaurusföderation verdeutlicht worden.

Kapitel 12

Analyse und Bewertung von Thesaurusföderationen

Für die iterative Ergebnisoptimierung ist die Bewertung des Zwischenstandes erforderlich (vgl. Abbildung 10.3, S. 174, dort insbesondere den Schritt *Bewertung des erzielten Zwischenergebnisses*). Eine solche Bewertung, deren Ergebnisse Grundlage der weiteren Optimierung sind, ist in unserem Phasenmodell die letzte Phase der Begriffsintegration (zur Einordnung vgl. Abbildung 12.1). Zielsetzung der Bewertung ist es, Schwachstellen in der aktuellen Integration zu identifizieren, um diese in der bewertungsbasierten Optimierungsphase beseitigen zu können.

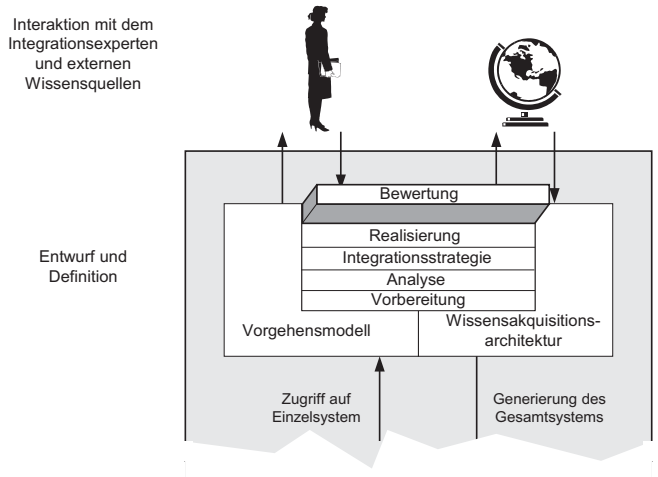

Abbildung 12.1: Die Bewertung als letzte Phase der Begriffsintegration

Die in diesem Kapitel entwickelten Kriterien für die Zwischenergebnisbewertung können zugleich in eine Bewertung des erreichten Endergebnisses einfließen. Weitere Bewertungskriterien, die sich aus der Übertragung der Kriterien zur Bewertung von Komponententhesauri ergeben, jedoch von uns im Rahmen dieser Arbeit nicht für die Zwischenergebnisbewertung zur Integrationsoptimierung vorgesehen sind, können diese ergänzen. Wir merken dies an den entsprechenden Stellen in diesem Kapitel an, um mittels einer solchen Endergebnisbewertung in Zukunft einen Vergleich von Thesaurusföderationen ebenso wie eine Bewertung verschiedener Ansätze und Verfahren zu ermöglichen.

In den Abschnitten 12.2 und 12.3 werden Kriterien für die Bewertung des Zwischenergebnisses

erarbeitet. Hier steht das Überprüfen der Erfüllung der in der Analysephase aufgestellten Erwartungen (im Folgenden durch (**EE**) angezeigt) im Vordergrund. Können diese Erwartungen (vgl. Abschnitte 9.2.2 bis 9.2.5) nicht erfüllt werden, müssen daraus die in Abschnitt 10.2.3.3, S. 177, eingeführten Hypothesenziele hergeleitet werden (im Folgenden durch (**HZ**) hervorgehoben). Diese Hypothesenziele steuern – wie wir in Abschnitt 12.1 noch einmal veranschaulichen – die weitere Optimierung. Zusätzliche Aspekte, die zu einer abschließenden oder vergleichenden Bewertung herangezogen werden können, sind durch (**B**) markiert. Insofern der menschlichen Experte diese Bewertungskriterien interpretiert, können Sie auch zur Optimierung der Integration herangezogen werden.

12.1 Steuerung der Begriffsintegration durch Hypothesenziele

Bereits in Kapitel 10 wurde bei der Entwicklung der Teilphasen der Integrationsstrategie dargestellt, wie die Bewertung eines erzielten Zwischenstandes der Integration als Grundlage der bewertungsbasierten Optimierung verwendet wird. Abbildung 12.2 veranschaulich den Zusammenhang mit dem Fokus auf Erwartungen, deren Nicht-Erfüllungen und daraus hergeleiteten Hypothesenzielen.

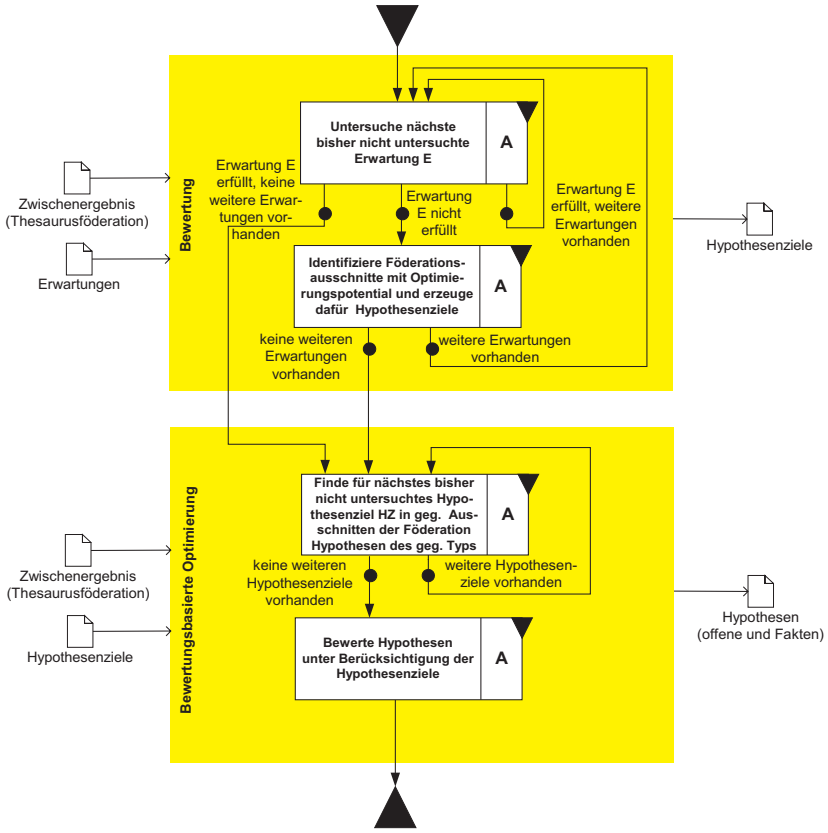

Abbildung 12.2: Herleitung von Hypothesenzielen aus nicht erfüllten Erwartungen als Basis der bewertungsbasierten Optimierung

Wurde eine Erwartung nicht erfüllt, kann analysiert werden, welche Weiterentwicklung der Föderation notwendig wäre, um die Erwartung besser zu erfüllen (z.B. bei weniger als der erwarteten Anzahl von Äquivalenzbeziehungen weitere Äquivalenzbeziehungen). Evtl. können auch Ausschnitte der Föderation bestimmt werden, die besonders intensiv betrachtet werden sollen

(z.B. bisher nicht verbundene Teilgraphen). Dieser Weiterentwicklungsbedarf wird in Form von Hypothesenzielen für jede nicht erfüllte Erwartung formuliert.

Innerhalb der bewertungsbasierten Optimierung wird nun versucht, Hypothesen entsprechend der Hypothesenziele zu finden. Dabei können sowohl andere Verfahren als zuvor eingebracht werden (vgl. Kapitel 11 und hier insbesondere Tabelle 11.3, S. 194) als auch die Bewertungen für bereits vorhandene oder neu generierte Hypothesen anders als zuvor ausfallen (z.B. können Hypothesen, deren Bewertungen bisher geringfügig unterhalb des Akzeptanzmaßstabes lagen, akzeptiert werden, falls diese Hypothesen zur Erfüllung mindestens eines Hypothesenziels beitragen).

12.2 Quantitative Analyse einer Thesaurusföderation

Grundlage der quantitativen Analyse sowohl eines Zwischen- als auch eines Endergebnisses der Integration sind wie bei der Analyse der Komponententhesauri Kennzahlen. Wir übertragen daher die in den Abschnitten 9.2.2 bis 9.2.4 vorgestellten Kennzahlenkriterien auf Thesaurusföderationen. Unter Berücksichtigung der zusätzlichen Eigenschaften von Thesaurusföderationen leiten wir weitere Kriterien her.

12.2.1 Quantitative Analyse der Benennungen

Bereits in Abschnitt 9.2.2 wurden die (Intra-Thesaurus-) Äquivalenzbeziehungen bei der Analyse der Benennungen berücksichtigt. Wir führen dies hier für Inter-Thesaurus-Äquivalenzbeziehungen fort. Denn relevant für eine Analyse der Thesaurusföderation ist nicht nur die Anzahl der Deskriptoren (die sich als Summe der Deskriptoren in den Komponententhesauri ergibt), sondern insbesondere die Anzahl der Inter-Thesaurus-Äquivalenzbeziehungen und die Anzahl der Föderierten Begriffe:

Anzahl der Inter-Thesaurus-Äquivalenzbeziehungen: Die Anzahl der Inter-Thesaurus-Äquivalenzbeziehungen berechnet sich aus der Anzahl expliziter und implizierter Inter-Thesaurus-Äquivalenzbeziehungen. (**B**)

In Abschnitt 9.2.2 wurde anhand der identischen Benennungen in den verschiedenen Komponententhesauri eine minimale Anzahl der voraussichtlich aufzufindenden Inter-Thesaurus-Äquivalenzbeziehungen ($IT\ddot{A}_{min}$) bestimmt. Nun kann überprüft werden, inwiefern diese Zahl bereits erreicht oder noch unterschritten ist (**EE**). Darüber hinaus können Teilgraphen identifiziert werden, in denen die mittlere Anzahl von Inter-Thesaurus-Äquivalenzbeziehungen deutlich geringer als im Durchschnitt ist. Diese Teilgraphen sollen bei einer weiteren Suche nach Inter-Thesaurus-Äquivalenzbeziehungen besonders berücksichtigt werden. (**HZ**)

Anzahl der Föderierten Begriffe: Bei der Föderation von zwei Komponententhesauri kann anhand der Anzahl der Inter-Thesaurus-Äquivalenzbeziehungen die Anzahl der Föderierten Begriffe abgeleitet werden. Bei einer Föderation von mehr als zwei Thesauri gilt dies nicht mehr, da z.B. drei Inter-Thesaurus-Äquivalenzbeziehungen aus drei Deskriptoren einen Föderierten Begriff oder aus sechs Deskriptoren drei Föderierte Begriffe bilden können (vgl. Abbildung 12.3). Daher wird die Anzahl der Föderierten Begriffe zusätzlich bestimmt.

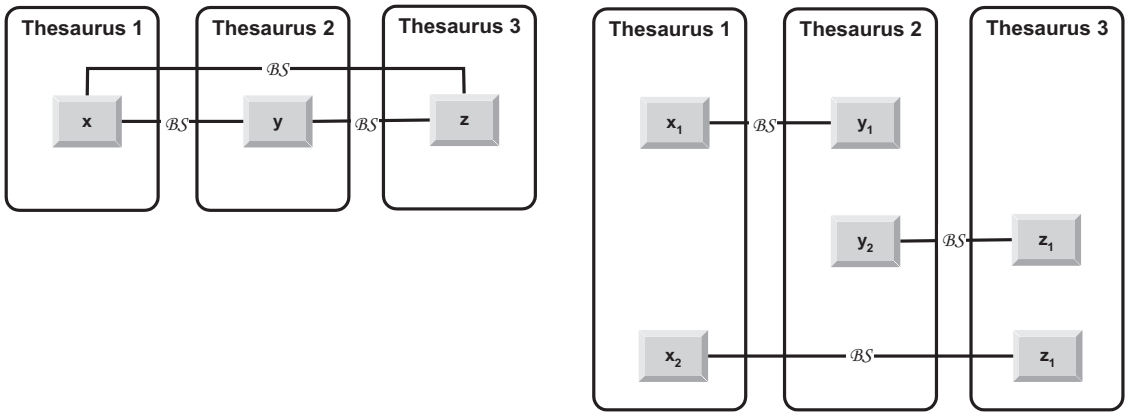

Abbildung 12.3: Die gleiche Abzahl von Inter-Thesaurus-Äquivalenzbeziehungen kann zu verschiedenen Anzahlen von Föderierten Begriffen führen

Die Anzahl der Föderierten Begriffe vermittelt einen Eindruck von der begrifflichen Vielfalt der Thesaurusföderation. (**B**)

Benutze-Kombination-Beziehungs-Anteil: Wie für Komponententhesauri kann auch für die Thesaurusföderation bestimmt werden, wie groß das Verhältnis der Anzahl der (Inter-Thesaurus-) Benutze-Kombination-Beziehungen zur Anzahl der (Inter-Thesaurus-) Äquivalenzbeziehungen ist. (**B**)

Wurden während der Analyse der Komponententhesauri sehr unterschiedliche Äquivalenzverhältnisse festgestellt, sind die Erwartungen an die Föderation abhängig von der Zielsetzung (vgl. S. 155): Soll die Thesaurusföderation eher als Information-Retrieval-unterstützende Föderation entstehen (vgl. Abschnitt 9.2.5, S. 160), wird ein größerer Benutze-Kombination-Beziehungs-Anteil erwartet (**EE**). Wird hingegen eine eher definitionswörterbuchartige Föderation angestrebt, wird ein kleinerer Benutze-Kombination-Beziehungs-Anteil erwartet (**EE**). Stattdessen wird für Begriffe aus dem Thesaurus mit großem Äquivalenzverhältnis, die durch einen Deskriptor und eine Reihe von Nicht-Deskriptoren repräsentiert sind, eine Einschränkung des Begriffsumfangs erwartet (**EE**). Eine solche Einschränkung kann implizit geschehen (Nicht-Deskriptoren dieses Thesaurus werden nachrangig behandelt) oder aber explizit durch „Ersetzen“ des Deskriptors mit seinen Nicht-Deskriptoren durch spezifischere Deskriptoren / Nicht-Deskriptoren als Ergänzende Begriffe[1].

Ist entgegen der Erwartungen der Anteil der Benutze-Kombination-Beziehungen bzw. die Anzahl der expliziten Begriffseinschränkungen gering, sollen insbesondere Deskriptoren mit vielen Nicht-Deskriptoren, für die noch keine Benutze-Kombination-Beziehungen etabliert bzw. für die noch nicht der Begriffsumfang eingeschränkt wurde, untersucht werden. (**HZ**)

Anzahl Ergänzender Begriffe: Die Anzahl der Ergänzenden Begriffe vermittelt einen Eindruck vom Umfang des Thesaurus der Ergänzenden Begriffe. (**B**)

Wie im vorherigen Punkt ersichtlich wurde, wird für den Fall einer expliziten Begriffsumfangreduktion bei Thesauri mit unterschiedlichen Äquivalenzverhältnissen und der Zielsetzung einer definitionswörterbuchartigen Föderation eine größere Anzahl Ergänzender Begriffe erwartet. (**EE**)

[1] „Ersetzen“ bedeutet, dass der Ergänzende Begriff alle erforderlichen Beziehungen des ersetzten Begriffs eingeht und alle Beziehungen zu diesem ersetzten Begriff als redundant markiert werden.

Alle weiteren Kriterien für die Analyse von Benennungen in Komponententhesauri sind für die Analyse von Thesaurusföderationen nicht relevant und werden nicht übertragen.

12.2.2 Quantitative Analyse der Relationen

Nachdem die Inter-Thesaurus-Äquivalenzbeziehungen bereits bei der quantitativen Analyse der Benennungen einbezogen wurde, konzentriert sich die quantitative Analyse der Relationen auf die weiteren Relelationstypen.

Konnektivität: Wie für Komponententhesauri kann auch für eine Thesaurusföderation die Konnektivität als Verhältnis der Anzahl der Deskriptoren, die mindestens eine Beziehung zu einem anderen Deskriptor eingehen, zu der Gesamtzahl der Deskriptoren berechnet werden (**B**). Inter-Thesaurus-Äquivalenzbeziehungen dürfen dabei jedoch nicht berücksichtigt werden, da die Konnektivität durch diese nicht verbessert wird.[2]

Als idealer Wert für die Konnektivität auch der Thesaurusföderation wird 1 angesehen.

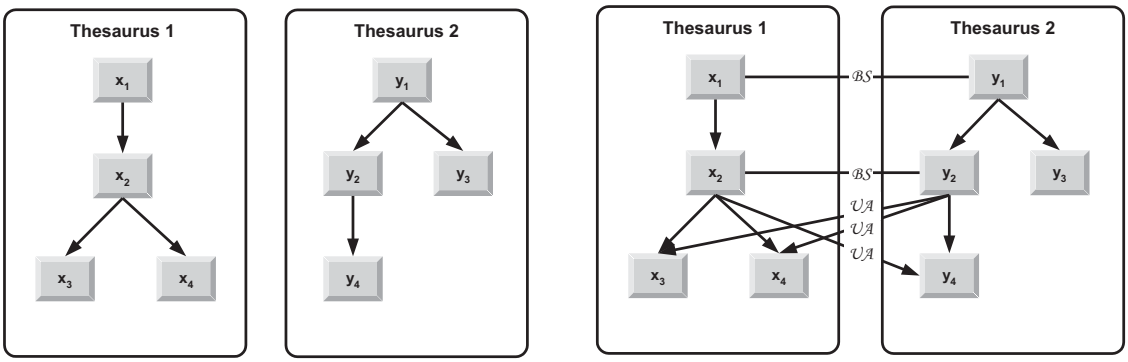

Abbildung 12.4: Bei einer *Konnektivität* von 1 in beiden Föderationen ist die *Inter-Thesaurus-Konnektivität* in der linken Föderation 0, in der rechten $\frac{7}{8} = 0.875$

Inter-Thesaurus-Konnektivität: Während die Konnektivität ein Maß für alle Verbindungen zwischen den Knoten der Föderation ist, wird zusätzlich ein Maß für die reinen Inter-Thesaurus-Verbindungen benötigt. Wir führen daher zusätzlich die Inter-Thesaurus-Konnektivität ein, die als Verhältnis der Anzahl der Deskriptoren, die mindestens eine Inter-Thesaurus-Beziehung zu einem anderen Deskriptor eingehen, zu der Gesamtzahl der Deskriptoren berechnet wird (**B**). Um einen Eindruck über die gesamten Inter-Thesaurus-Beziehungen zu erhalten, werden Inter-Thesaurus-Äquivalenzbeziehungen ebenfalls berücksichtigt.

So ist z.B. möglich, dass die Konnektivität aufgrund größtmöglicher Konnektivität innerhalb der Komponententhesauri bereits 1 ist, ohne dass eine einzige Inter-Thesaurus-Beziehung existiert. Die Inter-Thesaurus-Konnektivität wäre in diesem Fall jedoch 0 (vgl. auch die Beispiele in Abbildung 12.4).

Erfahrungswerte für die Inter-Thesaurus-Konnektivität existieren noch nicht. Die Inter-Thesaurus-Konnektivität kann nicht nur für den gesamten Föderationsgraphen berechnet

[2]Zwei isolierte Deskriptoren x und y, die durch eine Inter-Thesaurus-Äquivalenzbeziehung verbunden werden, bleiben isoliert, da der gesamte erreichbare Teilgraph ausschließlich aus Knoten der Äquivalenklasse besteht, der Begriff also mit keinem anderen Begriff verbunden ist. Sobald jedoch einer der Deskriptoren x oder y zusätzlich mit z.B. einer Hierarchiebeziehung mit einem weiteren Deskriptor z verbunden wird, ist aufgrund der implizierten Hierarchiebeziehung des anderen Deskriptors die Isolation des Begriffs beendet.

werden, sondern auch für Ausschnitte (z.B. innerhalb von Gruppen) und somit einen Vergleich dieser Ausschnitte ermöglichen (**B**). In Gruppen, die in verschiedenen Komponententhesauri jeweils eine größere Anzahl von Begriffen beinhalten, wird eine gute Inter-Thesaurus-Konnektivität erwartet (**EE**). Falls dies nicht der Fall ist, sollen diese Gruppen gezielt nach weiteren Inter-Thesaurus-Beziehungen untersucht werden. (**HZ**)

Zugänglichkeit: Das von der Komponententhesaurusanalyse bekannte Kriterium der Zugänglichkeit kann direkt auf Thesaurusföderationen übertragen werden (**B**). Statt auf Deskriptoren bezieht sich der Wert jedoch auf föderierte Begriffe, die durch die Äquivalenzklasse der Deskriptoren repräsentiert werden. In Konflikten als redundant markierte Beziehungen werden nicht berücksichtigt, da diese bei der Anwendung der Föderation nicht zur Zugänglichkeit der Begriffe beitragen. Abbildung 12.5 zeigt ein Beispiel zur Berechnung der Zugänglichkeit.

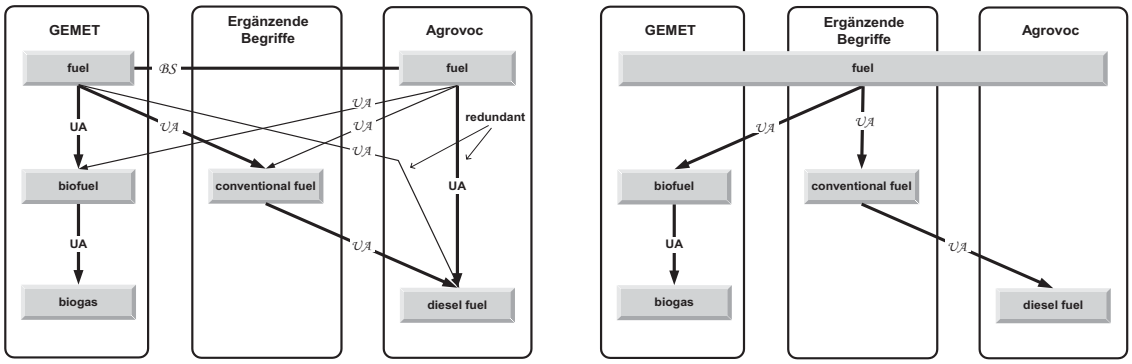

Abbildung 12.5: Bei der Berechnung der Zugänglichkeit für einen föderierten Begriff werden die einen Begriff repräsentierenden Deskriptoren zu einem Knoten und die zugehörigen Beziehungen entsprechend aggregiert und redundante Beziehungen nicht berücksichtigt. Der föderierte Begriff *fuel* besitzt somit die Zugänglichkeit 2. (Links: Betrachteter Teilgraph. Rechts: Reduzierter Graph zur Berechnung der Zugänglichkeit)

Da die Thesaurusföderation als ein großer Thesaurus betrachtet werden kann, gelten die gleichen Idealwerte (2 bis 5 berechnet über Hierarchie- und Assoziationsbeziehungen). (**EE**)

Föderierte Begriffe, die diese Idealwerte unter- bzw. überschreiten, sollten daraufhin untersucht werden, ob weitere Inter-Thesaurus-Beziehungen etabliert oder vorhandene Beziehungen entfernt werden können (**HZ**). Ausgenommen werden können Deskriptoren, für die bereits besondere Randbedingungen festgestellt wurden, etwa Fachtermini eines Thesaurus ohne Entsprechungen in anderen Thesauri. Besonders intensiv betrachtet werden sollen hingegen Oberbegriffe von großen Hierarchien. Denn eine gute Zugänglichkeit dieser Begriffe verbessert zumindest die indirekte Zugänglichkeit der Knoten des untergeordneten Teilgraphen. (**HZ**)

Inter-Thesaurus-Relationsverhältnisse: Einen guten Eindruck über den Stand der Integration vermitteln die Verhältnisse der Anzahlen der Inter-Thesaurus-Beziehungen der verschiedenen Typen untereinander (wie bei der Analyse von Thesauri sind Hierarchie-/Assoziationsrelationsverhältnis und Abstraktions-/Bestandsverhältnis sowie zusätzlich Äquivalenz-/Hierarchieverhältnis relevant). (**B**)

Bei unterschiedlichen Präkoordinationsgraden der Komponententhesauri wird aufgrund der größeren Anzahl von erwarteten Hierarchiebeziehungen ein kleineres Äquivalenz-

/Hierarchieverhältnis erwartet (**EE**). Ist dies nicht der Fall und sind bisher erst wenige Inter-Thesaurus-Hierarchiebeziehungen zwischen den Thesauri mit großen und kleinen Präkoordinationsgraden etabliert, soll in den Thesauri mit großem Präkoordinationsgrad nach Unterbegriffen zu Begriffen in Thesauri mit großem Präkoordinationsgrad gesucht werden. (**HZ**)

Bei großer Flexibilität berechnet über alle Komponententhesauri wird ebenfalls eine große Anzahl an Hierarchiebeziehungen erwartet (**EE**). Ist deren Anteil bisher gering, soll insbesondere in Teilgraphen mit wenigen Hierarchiebeziehungen gezielt nach weiteren Hierarchiebeziehungen gesucht werden. (**HZ**)

Anteil polyhierarchischer Begriffe: Auch für die Thesaurusföderation sollte der Anteil polyhierarchischer Begriffe, also solcher mit mehr als einem Oberbegriff, klein bleiben. Insbesondere sollen Begriffe nicht mehr als 2 Abstraktionsoberbegriffe haben. Ist dies dennoch der Fall, sollen solche Begriffe daraufhin untersucht werden, ob die Bedeutung der Begriffe trotz der mehreren Oberbegriffe spezifisch genug bleibt (**HZ**). Um dies festzustellen, wird jedoch ein menschlicher Integartionsexperte benötigt, der ggf. Restrukturierungsmaßnahmen an der Föderation vornimmt. Das Hypothesenziel muss also eine solche Aufforderung an den menschlichen Integrationsexperten enthalten.

12.2.3 Quantitative Analyse der Struktur

Homogenität der Gruppen: Als Kriterium zur Beurteilung der Veränderung an der Gruppenstruktur hat die Homogenität der Gruppen eine besondere Bedeutung. Relevant ist insbesondere, inwiefern die Homogenität verbessert werden konnte (Indikator für eine gleichmäßigere thematische Abdeckung) oder verschlechtert wurde (weitere Kristallisation thematischer Schwerpunkte). (**B**)

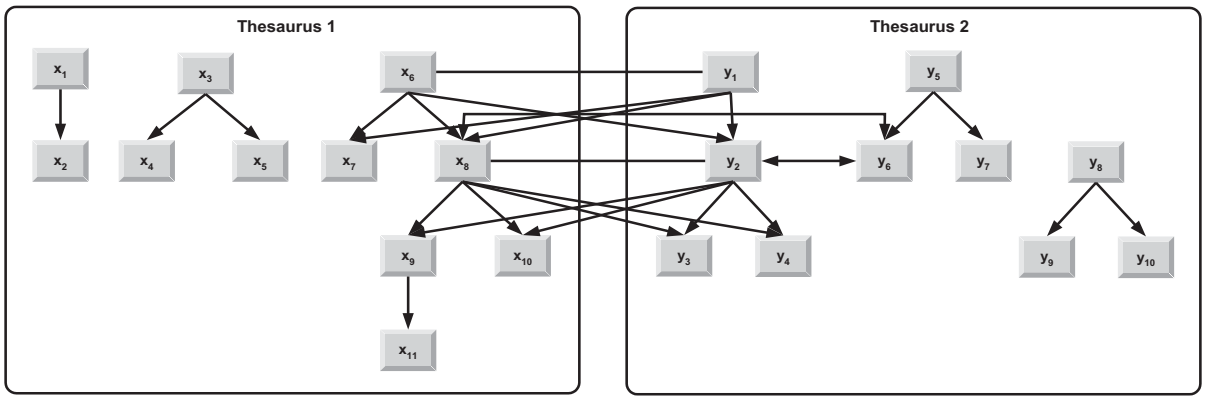

Abbildung 12.6: Beispiele für eine gute Inter-Thesaurus-Konnektivität (0.476) bei gleichzeitig nur geringfügig verbessertem Wert für die Verbindungseinheit gegenüber dem isolierten Zustand (Verbindungseinheit von 0.8 bei isolierten Komponententhesauri und 0.85 bei der dargestellten Integration). Begründet werden diese Werte durch eine Reihe von Inter-Thesaurus-Beziehungen zwischen zwei in den jeweiligen Komponententhesaurus verbundenen Teilgraphen, jedoch fehlenden Inter-Thesaurus-Verbindungen zwischen den weiteren isolierten Teilgraphen (die Verbindungseinheit ist mit 0.8 bzw. 0.889 in beiden Komponententhesauri deutlich unter dem Idealwert vom 1).

Verbindungseinheit: Mittels der Verbindungseinheit können isolierte Strukturen aufgedeckt werden. Eine schlechte Verbindungseinheit fordert eine weitere Betrachtung der isolierten

Strukturen, um deren Integration in die Föderation zu verbessern (vgl. Abbildung 12.6). Besonders berücksichtigt werden sollen dabei isolierte Strukturen, die aus wenigen Knoten bestehen. (**HZ**)

Mittlere Höhe: Nach der Integration ist eine Auswertung der mittleren Höhe insbesondere im Vergleich mit den entsprechenden Werten der Komponententhesauri zu betrachten. Wurde die mittlere Höhe deutlich vergrößert, bedeutet dies eine weitere Differenzierung der Begriffe. (**B**)

Höhenvarianz: Auch die Höhenvarianz muss zusammen mit den entsprechenden Werten der Komponententhesauri betrachtet werden. Eine größere Höhenvarianz deutet auf eine zunehmende Inhomogenität hinsichtlich des Detaillierungsgrades der Begriffe hin. (**B**)

12.2.4 Quantitative Analyse der Konflikte

Bei der Komponententhesaurusanalyse spielen Konflikte keine Rolle, da ausschließlich die isolierten, konfliktfreien Thesauri betrachtet werden. Hinsichtlich der Bewertung und des Vergleichs einer Thesaurusföderation kann jedoch anhand der Analyse von Konflikten die Kompatibilität von Komponententhesauri bewertet werden. Eine sehr hohe Konfliktdichte deutet auf eine Vielzahl von Widersprüchen in den Thesauri und somit eine geringe Kompatibilität der Thesauri hin.

Mittels der folgenden Kriterien wird eine konfliktbasierte Analyse sowie die Herleitung von Hypothesenzielen anhand dieser Bewertung ermöglicht.

Anzahl der Zyklen: Wir unterscheiden bei Zyklen zwischen der Anzahl von Abstraktions-, Bestands- und Hierarchiezyklen. (**B**)

Den deutlichsten Hinweis auf in den Thesauri ausgedrückte unterschiedliche Weltsichten geben Abstraktionszyklen. Wenn Begriffe in den Thesauri gegensätzlich in Bezug auf direkte oder indirekte Abstraktionsbeziehungen angeordnet wurden, sind verschiedene Ursachen möglich: Es kann sich sowohl um tatsächlich auseinandergehende wissenschaftliche Ansichten handeln[3], als auch um Erfassungsfehler in einem der Komponententhesauri[4] oder schließlich um Integrationsfehler[5]. Um die Ursache festzustellen und bei Integrationsfehlern diese zu beseitigen, sollte der menschliche Experte auf entsprechende Zyklen aufmerksam gemacht werden. (**HZ**)

Auch Bestands- und Hierarchiezyklen geben ähnliche Hinweise. Aufgrund der fehlenden Transitivität der Bestandsrelation werden Bestandszyklen und Hierarchiezyklen (als gemischte Abstraktions-/Bestandszyklen) jedoch weniger Gewicht für die weitere Optimierung gegeben. Daher kann eine Vorlage solcher Zyklen für den menschlichen Integrationsexperten mit geringerer Priorität geschehen. Ausgenommen hiervon jedoch sind direkt entgegengesetzte Beziehungen, die mit gleicher Priorität wie Abstraktionszyklen gehandhabt werden sollen. (**HZ**)

[3] Gerade in neuen wissenschaftlichen Gebieten ist die Terminologie häufig noch nicht etabliert sondern wird uneinheitlich definiert und verwendet.

[4] Besonders in Randgebieten des vom Thesaurus abgedeckten Fachgebietes kann fehlende Kompetenz oder Sorgfalt der Thesaurusersteller zu solchen Abweichungen vom eigentlich fachüblichen Verständnis führen.

[5] Integrationsfehler können z.B. aufgrund fälschlicherweise mittels einer Inter-Thesaurus-Äquivalenzbeziehung verbundener Deskriptoren entstehen.

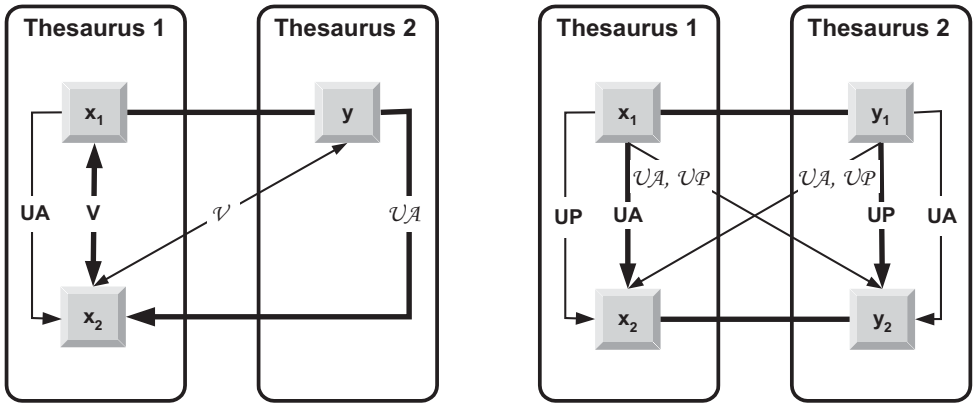

Abbildung 12.7: Werden in Thesaurus 1 innerhalb der Assoziationsrelation versteckte Hierarchien vermutet, wird ein wie links dargestellter Fall des Verstoßes gegen die Eindeutigkeit einer Kante erwartet und benötigt keine weitere Aufmerksamkeit. Der rechts dargestellte Verstoß, der aufgrund zweier unterschiedlicher Hierarchiebeziehungen hervorgerufen wird, sollte jedoch genauer untersucht werden.

Anzahl der Verstöße gegen Eindeutigkeit einer Kante: In Abschnitt 6.3.7.1, S. 95, hatten wir im Falle eines Verstoßes gegen die Eindeutigkeit einer Kante zwischen zwei Knoten zwischen verschiedenen Kantentypen (Beziehungstypdifferenz) und entgegengesetzten Kantenrichtung (Zyklen mit der Pfadlänge 2) unterschieden. Da Zyklen bereits zuvor betrachtet wurden, ist als weitere Kennzahl nur die Anzahl der Beziehungstypdifferenzen relevant. (**B**)

Wurde bereits während der Komponententhesaurusanalyse ein unterschiedlicher Gebrauch der Beziehungen festgestellt (z.B. aufgrund eines unterschiedlichen Hierarchie-/Assoziationsverhältnisses), kann anhand der festgestellten Ursachen (z.B. enthalten in AGROVOC die Bestandsbeziehungen versteckte Hierarchien, vgl. Abschnitt 9.3.2, S. 163) über das weitere Vorgehen entschieden werden. Bei erwarteten Konfliktmarkierungen (**EE**) besteht kein Handlungsbedarf. Ansonsten sollte der menschliche Experte auf das Auftreten hingewiesen werden, so dass er ggf. manuell Korrekturen vornehmen kann (vgl. auch Abbildung 12.7). (**HZ**)

Anzahl der Abstraktionsredundanzen: Eine große Anzahl von Abstraktionsredundanzen deutet darauf hin, dass häufig in vorhandene Hierarchien weitere Begriffe eingefügt wurden (und somit die ursprüngliche Beziehung redundant wurde). Wenn aufgeschlüsselt wird, in welchen Thesauri wieviele Redundanzmarkierungen im Verhältnis zu Intra-Thesaurus-Abstraktionsbeziehungen vorgenommen wurden, kann erkannt werden, ob es Schwerpunkte bei der Verfeinerung der Abstraktionshierarchien gab. (**B**)

Anzahl der Schwesternassoziationen: Zur Interpretation der Anzahl der Schwesternassoziationen ist wie bei den Beziehungstypdifferenzen eine Berücksichtigung der Analyse der Komponententhesauri sowie der Erwartungen erforderlich. Ob und welcher Handlungsbedarf vorhanden ist, kann entsprechend entschieden werden.

Anteil AP/PA-Schwestern: Der AP-/PA-Schwesternanteil wird anhand der Anzahl der Begriffe, die vor der Integration ausschließlich Bestands- bzw. Abstraktionsunterbegriffe hatten und nach der Integration Unterbegriffsbeziehungen beiden Typs eingehen im Verhältnis zu der gesamten Anzahl von Begriffen, für die Inter-Thesaurus-Unterbegriffsbeziehungen etabliert wurden, bestimmt. (**B**)

Ein großer AP/PA-Schwesternanteil deutet darauf hin, dass in einem Thesaurus zu Begriffen häufig ausschließlich Bestands- oder ausschließlich Abstraktionsbeziehungen betrachtet wurden, die durch die Föderation jedoch um Beziehungen des jeweils anderen Typs ergänzt wurden. Falls es das Ziel ist, solche AP/PA-Schwestern möglichst zu vermeiden, kann zur Konfliktauflösung das Finden von Ergänzenden Begriffen vorgeschlagen werden, vgl. Abbildung 11.6, S. 205. (**HZ**)

12.3 Qualitative Analyse einer Thesaurusföderation

Bereits bei der Komponententhesaurusanalyse haben wir die Schwierigkeiten einer umfassenden qualitativen Analyse aufgezeigt (vgl. Abschnitt 9.2.1, S. 153). Entsprechend kann auch eine Thesaurusföderation nur von einem menschlichen Experten umfassend hinsichtlich der Qualität der Integration bewertet werden. Um dennoch zumindest für einzelne Deskriptoren eine Aussage über die Güte ihrer Integration machen zu können, werden in diesem Abschnitt entsprechende Maße entwickelt.

Zur Bestimmung der Güte der inhaltlichen Integration eines Deskriptors ist eine semantische Bewertung der Beziehungen, die dieser Deskriptor eingeht, erforderlich. Eine solche Bewertung beinhaltet zwei Aspekte:

Korrektheit: Sind die etablierten Beziehungen zu recht etabliert?

Vollständigkeit: Sind die erforderlichen Beziehungen vollständig etabliert?

12.3.1 Korrektheit

Aus dem Information Retrieval ist die klassische Kennzahl für die Beurteilung der Korrektheit (Genauigkeit) die Precision (vgl. Anhang D.2, S. 273). In [CL92] wird diese Kennzahl als *ConceptPrecision* (an anderer Stelle auch $TermPrecision$) bereits auf die Bewertung von Thesauri übertragen.[6] Ausgehend von einem initialen Deskriptor im Thesaurus wird die Anzahl der durch Beziehungen verbundenen, relevanten Deskriptoren bestimmt und das Verhältnis zur Anzahl aller verbundenen Deskriptoren berechnet:

$$ConceptPrecision = \frac{\text{Anzahl der verbundenen relevanten Deskriptoren}}{\text{Anzahl aller verbundenen Deskriptoren}}$$

Um dieses Maß auch auf Thesaurusföderationen anwenden zu können, ist es erforderlich zu definieren, wie die Verbundenheit und wie die Relevanz bestimmt werden:

Verbundenheit: Zur Feststellung der Verbundenheit gilt es zu definieren, welche Beziehungstypen und welche Pfadlängen berücksichtigt werden sollen. Um eigenständige Aussagen über die Korrektheit der verschiedenen Beziehungstypen zu erhalten, ist es erforderlich die *ConceptPrecision* weiter in $ConceptPrecision_{Äquivalenzrelation}$, $ConceptPrecision_{Benutze-Kombination-Relation}$, $ConceptPrecision_{Abstraktionsrelation}$ und $ConceptPrecision_{Bestandsrelation}$ zu unterteilen. Nur Beziehungen des entsprechenden Typs sollen berücksichtigt werden, gerichtete Beziehungen nur in eine Richtung.

[6] In [Rug92] und [CYFS95] wird die *ConceptPrecision* zur Bewertung der Korrektheit von automatisch erstellten Paaren von ähnlichen Begriffen verwendet. Viegener nutzt dieses Maß darüberhinaus für die Bewertung automatisch erstellter Thesauri [Vie97].

Entscheidend bei einer Bewertung der korrekten Einordnung eines Deskriptors ist ausschließlich die Betrachtung der direkt von diesem Deskriptor ausgehenden Beziehungen. Als verbunden gelten somit nur Deskriptoren die durch einen Pfad der Läge 1 mit dem Ausgangsdeskriptor verbunden sind. Sind Beziehungen innerhalb von Konfliktmarkierungen als für die Standard-Konfliktauflösung zu entfernen markiert, dürfen diese Beziehungen nicht berücksichtigt werden.

Relevanz: Als relevant gilt ein Deskriptor dann, wenn auch in der idealen Thesurusföderation bestehend aus denselben Komponententhesauri der untersuchten Föderation, der Deskriptor als mit identischem Beziehungstyp verbundener Deskriptor gefunden wird.

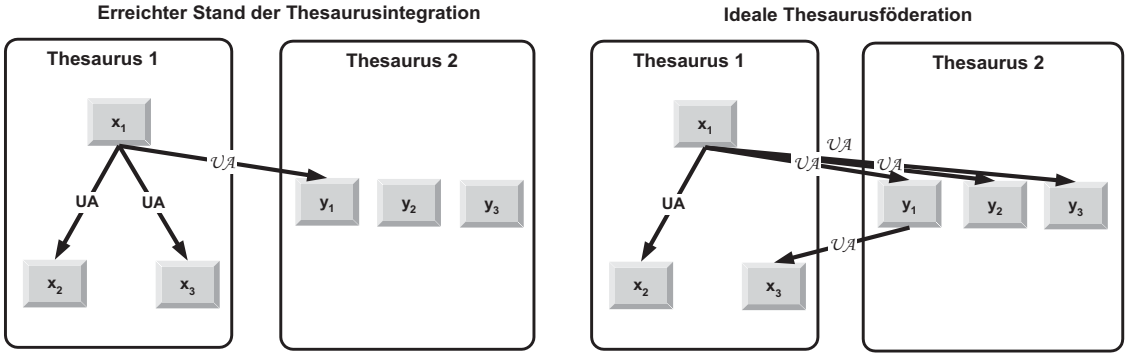

Abbildung 12.8: Beispiel zur Bestimmung der verbundenen und relevanten Deskriptoren: Bei Betrachtung des Deskriptors x_1 eines erreichten Standes der Thesaurusintegration sind die Deskriptoren x_2, x_3 und y_1 verbundene Deskriptoren, x_3 der einzige nicht relevante Deskriptor. y_2 und y_3 wären relevante Deskriptoren, sind jedoch nicht verbunden.

Beispiel 12.1 *Das Maß* $ConceptPrecision_{Abstraktionsrelation}$ *berechnet sich für das in Abbildung 12.8 dargestellte Beispiel wie folgt:*

$$ConceptPrecision_{Abstraktionsrelation} = \frac{2}{3} = 0.667$$

Die Schwierigkeit der Bewertung der Relevanz wird sofort deutlich: Wie gelangt man an die ideale Thesaurusföderation, die als Referenzquelle benötigt wird? Eine manuelle Erstellung einer vollständigen, idealen Thesaurusföderation ist nur für sehr kleine Föderationen mit wenigen Dutzend Deskriptoren sinnvoll. Bei einer größeren Anzahl von Deskriptoren ist eine Beschränkung auf Ausschnitte der Föderation erforderlich. Dazu wird eine Menge von Deskriptoren aus den Komponententhesauri bestimmt, für die menschliche Experten alle als notwendig betrachteten Beziehungen, Ergänzende Begriffe und Konfliktmarkierungen einfügen.

Warum aber sollte dieser „ideale“ Ausschnitt besser sein als die Thesaurusföderation, die unter Verwendung der in dieser Arbeit vorgestellten Methoden und Werkzeuge ebenfalls mit menschlicher Expertise erstellt wurde? Würde das nicht zu einem Vergleich *Mensch* (Erstellung der idealen Thesaurusföderation) vs. *Mensch und Maschine* führen? Handelt es sich um verschiedene menschliche Experten, würden diese Experten als entscheidene Instanz stärker bewertet als das maschinelle System.

Als sinnvoll erscheint somit nur eine qualitative Bewertung einer *rein maschinell erstellten* Thesaurusföderation im Vergleich zu einer mit maschineller Unterstützung und menschlicher Expertise erstellten Föderation. Wir gehen also davon aus, dass der Zustand vor dem Eingriff

des menschlichen Experten nicht vollständig korrekt, nach dem korrigierenden Eingreifen jedoch vollständig korrekt ist und als Referenz dienen kann. Somit werden Zwischenergebnisse hinsichtlich der Korrektheit bewertbar.

Um Korrektheitsaussagen allgemeiner Gültigkeit machen zu können, ist es erforderlich, für die verschiedenen Beziehungstypen jeweils für eine Menge von Deskriptoren (Stichproben) die entsprechende *ConceptPrecision* zu berechnen und die Werte zu mitteln.

12.3.2 Vollständigkeit

Die Bewertung der vollständigen Integration eines Deskriptors erfolgt durch die Überprüfung, ob *alle erforderlichen* Beziehungen etabliert sind. Das Kriterium der Vollständigkeit ist ein für die Integration allgemeingültiges Kriterium, das z.B. in [BLN86] für die Schemaintegration erläutert wird.

Analog zur Bewertung der Korrektheit kann für die Vollständigkeit das aus dem Information Retrieval bekannte Kriterium des Recall (vgl. Anhang D.2, S. 273) übertragen werden (vgl. wiederum [CL92, Rug92, CYFS95, Vie97]). Dieses Maß wird dann *ConceptRecall* benannt und berechnet sich ausgehend von einem Ursprungs-Deskriptor aus dem Verhältnis der Anzahl der verbundenen relevanten Deskriptoren zur Anzahl aller relevanten Deskriptoren:

$$ConceptRecall = \frac{\text{Anzahl der verbundenen relevanten Deskriptoren}}{\text{Anzahl aller relevanten Deskriptoren}}$$

Verbundenheit und Relevanz können wie bei der Berechnung von *ConceptPrecision* betrachtet werden. Es werden jeweils Mittelwerte für die verschiedenen Typen der *ConceptRecall* berechnet.

Beispiel 12.2 *Das Maß* $ConceptRecall_{Abstraktionsrelation}$ *berechnet sich für das in Abbildung 12.8 dargestellte Beispiel wie folgt:*

$$ConceptRecall_{Abstraktionsrelation} = \frac{2}{4} = 0.5$$

12.3.3 Berücksichtigung von Ergänzenden Begriffen

Es bleibt die Frage offen, inwiefern Ergänzende Begriffe bei der qualitativen Analyse (Korrektheit und Verbundenheit) eine Rolle spielen. Da wir uns für die qualitative Analyse auf die Analyse von rein maschinell erstellten Zwischenergebnissen beschränkt haben und für das Finden von Benennungen für Ergänzende Begriffe menschliche Expertise erforderlich ist, würden Ergänzende Begriffe vollständig unberücksichtigt bleiben. Sollen sie jedoch, um einen umfassenderen Eindruck zu erhalten, berücksichtigt werden, können von maschinellen Experten vorgeschlagene Positionen für Ergänzende Begriffe ebenfalls in die Analyse einbezogen werden. Damit diese noch unbenannten Knoten eine interpretierbare Bedeutung bekommen, kann der menschliche Experte vor der Analyse aufgefordert werden, eine Benennung zu vergeben. Wird ansonsten keine Änderung an der Föderation vorgenommen, können wir die Föderation immer noch als rein maschinell erstellt betrachten.

Ergänzende Begriffe unterscheiden sich von Deskriptoren in Komponententhesauri dadurch, dass sie erst während der Begriffsintergration erzeugt werden und somit ein Identitätsvergleich zweier Ergänzender Begriffe keine von vornehinein bekannten Identifikatoren verwenden kann. Ein und derselbe Ergänzende Begriff kann etwa zu verschiedenen Zeitpunkten erzeugt worden sein, so

dass seine automatischen Identifikatoren im untersuchten Zwischenergebnis und in der Referenzquelle verschieden sind. Um dies zu vermeiden, ist vor der Berechnung der Maße zu prüfen, ob Ergänzende Begriffe in der Referenzquelle zur Berechnung verwendet werden und ob für diese Äquivalente im Zwischenergebnis existieren. Falls ja, sind gemeinsame Identifikatoren für den Vergleich zu verwenden.

Die Auswahl der Stichprobe zu Bestimmung der verschiedenen Typen von *ConceptPrecision* und *ConceptRecall* sollte Ergänzende Begriffe als Ursprungs-Deskriptoren und als relevante verbundene Deskriptoren berücksichtigen, damit diese mit den entsprechenden Vorbereitungen in die Berechnung der jeweilig gemittelten Werte für *ConceptPrecision* und *ConceptRecall* eingehen können.

12.4 Exemplarische Evaluierung eines Zwischenergebnisses

Nachdem wir in den vorangegangenen Abschnitten eine Reihe von Kennzahlen für die Analyse und Bewertung von Thesaurusföderationen entwickelt haben, zeigen wir nun anhand ausgewählter Kennzahlen, wie diese in einem konkreten Beispiel zur Verbesserung der Begriffsintegration beitragen. Wir greifen dazu die in Abschnitt 9.3, S. 161ff, zur Analyse von Komponententhesauri verwendeten Thesauri AGROVOC und GEMET wieder auf und bewerten nun den Zwischenstand nach einer initialen Integration und einer ersten zwischenergebnisbasierten Optimierung. Folgende Verfahren (vgl. Tabelle 11.3, S. 194) wurden bereits angewandt:

- Zum Finden und Bewerten von Äquivalenzhypothesen das Überprüfen auf lexikalische Gleichheit, den Übereinstimmungsgrad der Gruppen, den Abstand von Äquivalenzhypothesen, sowie das Auffinden gemeinsamer Unterbegriffe.
- Zum Finden und Bewerten von Benutze-Kombination-Hypothesen das Feststellen der lexikalischen Gleichheit zwischen Deskriptoren und Benutze-Kombination-Deskriptoren sowie des Übereinstimmungsgrades der Gruppen.
- Und schließlich zum Finden und Bewerten von Hierarchiebeziehungen die Analyse von Mehrwertbenennungen und des Übereinstimmungsgrades der Gruppen.

In den folgenden Abschnitten wird beispielhaft gezeigt, wie anhand der Analyse von Kennzahlen die bereits erreichte Güte der Begriffsintegration bewertet und wie mit Hilfe des menschlichen Experten Hypothesenziele zur Verbesserung aufgestellt werden. Wir zeigen, welche tatsächlichen Verbesserungen mit welchen Verfahren erzielt werden können und wie menschlicher und maschinelle Experten diese Verbesserungen gemeinsam erreichen.

12.4.1 Anzahl Äquivalenzbeziehungen

Ein erster Eindruck über die bisher erstellte Thesaurusföderation kann durch die Betrachtung der Anzahl der Äquivalenzbeziehungen gewonnen werden.

Ergebnis: Es sind bereits 2902 Inter-Thesaurus-Äquivalenzbeziehungen etabliert (somit ergeben sich $5398 + 16394 - 2902 = 18890$ Föderierte Begriffe).

Bewertung: Die minimal erwartete Anzahl von Inter-Thesaurus-Äquivalenzbeziehungen $IT\ddot{A}_{min}$, die mit 756 festgelegt wurde, konnte bereits deutlich übertroffen werden.

Angestrebte Verbesserung: Eine Verbesserung ist nicht erforderlich und somit auch kein Eingriff des menschlichen Integrationsexperten.

12.4.2 Benutze-Kombination-Beziehungs-Anteil

Aufgrund der semantischen Ähnlichkeit zu Äquivalenzbeziehungen betrachten wir auch Benutze-Kombination-Beziehungen

Ergebnis: Benutze-Kombination-Beziehungs-Anteil von 3.1 % (das entspricht 93 Benutze-Kombination-Beziehungen).

Bewertung: Da die Komponententhesauri AGROVOC und GEMET sehr unterschiedliche Äquivalenzverhältnisse auszeichnen und das Integrationsziel eine Information-Retrieval-unterstützende Föderation ist, wird der Benutze-Kombination-Beziehungs-Anteil als zu klein bewertet. Eine solche Bewertung kann nur dann maschinell erfolgen, wenn der erwartete Anteil zuvor quantifiziert wurde. Ansonsten erfolgt die Bewertung durch den menschlichen Experten.

Angestrebte Verbesserung: Der Benutze-Kombination-Beziehungs-Anteil soll deutlich erhöht werden. Als erfolgsversprechend werden Gruppen mit einer unterdurchschnittlichen Inter-Thesaurus-Konnektivität angesehen. Daraus ergibt sich das Hypothesenziel, weitere Benutze-Kombination-Beziehungen zu finden und dazu diese Gruppen zu betrachten. Da bereits die Bewertung durch den menschlichen Experten erfolgte, stellt dieser auch das Hypothesenziel auf und bestimmt die zu untersuchenden Gruppen anhand der Kennzahlen.

Erreichte Verbesserung: Alleine durch eine neue Bewertung durch die maschinellen Experten von bisher zu niedrig bewerteten Hypothesen in den angegebenen Bereichen werden 87 zusätzliche Fakten generiert, das entspricht annähernd einer Verdopplung der ursprünglichen Fakten. Diese neuen Fakten werden dem menschlichen Experten präsentiert. Er entscheidet, dass über 80 % dieser Beziehungen (70 Beziehungen) zu Recht etabliert werden. Somit kann der Benutze-Kombination-Beziehungs-Anteil auf 5.6 % deutlich gesteigert werden. In den verbleibenden Fällen schlägt er statt der Benutze-Kombination-Beziehung eine Hierarchie- oder Assoziationsbeziehung im nahen Umfeld der Integrationsstelle vor.

12.4.3 Inter-Thesaurus-Interkonnektivität

Ein Beispiel für eine relationsübergreifende Betrachtung der Föderation ist die Inter-Thesaurus-Interkonnektivität.

Ergebnis: Die Inter-Thesaurus-Konnektivität über sämtliche Deskriptoren der Föderation ist 0.36. Eine gruppenweise Betrachtung der Inter-Thesaurus-Interkonnektivität ergibt eine Spannbreite von Werten unterhalb von 0.05 bis oberhalb von 0.7.

Bewertung: Aufgrund der großen Unterschiede der Anzahl der Deskriptoren in beiden Thesauri ist – bei Verbindung aller GEMET-Deskriptoren – der maximal mögliche Wert für die Inter-Thesaurus-Konnektivität 0.50. Der Wert von 0.36 erscheint hinsichtlich der thematischen Nähe der Thesauri noch zu klein. Bei dem kleinsten Wert von 0.05 gilt es zu berücksichtigen, dass dieser für die Gruppe *Latin terms* gilt, die, wie zuvor festgestellt, kein Schwerpunkt der Begriffsintegration sein soll.

Angestrebte Verbesserung: Um die Inter-Thesaurus-Konnektivität zu verbessern, soll bevorzugt in Gruppen mit schlechter Qualität – die Gruppe *Latin terms* ausgenommen – nach Beziehungen zwischen bisher nicht durch Inter-Thesaurus-Beziehungen verbundenen Deskriptoren gesucht werden. Für unser Experiment beschränken wir uns auf die Gruppe *chemistry, substances, processes* mit einer Inter-Thesaurus-Interkonnektivität von 0.11 (maximale Inter-Thesaurus-Konnektivität in dieser Gruppe: 0.40). Es werden nun durch den menschlichen Experten für jeden Beziehungstyp Hypothesenziele aufgestellt, um Beziehungen zwischen Deskriptoren dieser Gruppe in beiden Thesauri zu finden. Wir betrachten jedoch nur Hypothesenziele zum Finden von Hierarchiebeziehungen.

Erreichte Verbesserung: Als einziges Verfahren zum Finden von Hierarchiebeziehungen wurde eine sehr einfache Reverse Suche in Definitionen angewandt. Wurde ein Wort in der Definition eines GEMET-Deskriptors als Bestandteil einer Benennung eines AGROVOC-Deskriptors in derselben Gruppe gefunden, wurde eine Abstraktionsoberbeziehung vom GEMET-Deskriptor zum AGROVOC-Deskriptor vorgeschlagen. Die Hypothese wurde verworfen, wenn bereits eine Beziehung zwischen den Deskriptoren bestand. Ansonsten wurde der menschliche Experte aufgefordert, die Hypothesen zu bewerten.

Zuvor bestanden ingesamt 78 Inter-Thesaurus-Beziehungen zwischen den Deskriptoren der Gruppe. Es konnten 32 neue Hypothesen erzeugt werden. 16 % dieser Hypothesen wurden vom menschlichen Experten bestätigt, 38 % abgelehnt und in den restlichen 46 % der Fällen wurde eine andere als die vorgeschlagene Beziehung etabliert (Vertauschung der Richtung Ober-/Unterbegriff, Bestands- oder Assoziationsbeziehungen).

Für die untersuchte Gruppe konnte die Inter-Thesaurus-Interkonnektivität somit von 0.11 auf 0.14 verbessert werden. Mit einer weiteren Verbesserung des eingesetzten Verfahrens, das dann nicht nur lexikalische Wortvergleiche ausführt, sondern versucht, Zusammenhänge im Satz zu berücksichtigen, könnten die maschinellen Vorschläge bzgl. der vorgeschlagenen Relationen weiter verbessert und somit der menschliche Aufwand verringert werden.

12.4.4 Zugänglichkeit und polyhierarchische Begriffe

Weitere Kennzahl für eine Bewertung der (Inter-) Thesaurus-Relationen ist die Zugänglichkeit, die hier zusammen mit thesaurusübergreifenden polyhierarchischen Begriffen betrachtet wird.

Ergebnis: Die durchschnittliche Zugänglichkeit für alle föderierten Thesaurusbegriffe beträgt 3.74. Es konnten mehr als 200 föderierte Begriffe mit Oberbegriffen aus zwei verschiedenen Thesauri (thesaurusübergreifend polyhierarchische Begriffe) identifiziert werden.

Bewertung: Die durchschnittliche Zugänglichkeit liegt deutlich über der Zugänglichkeit von 2.31 in GEMET und – wie anhand der deutlich mehr Deskriptoren in AGROVOC erwartet – nahe an der Zugänglichkeit von AGROVOC (3.51). Sie liegt damit bereits innerhalb des Zielbereichs. Die Anzahl thesaurusübergreifend polyhierarchischer Begriffe wird als zu hoch betrachtet.

Angestrebte Verbesserung: Es ist keine Verbesserung der Zugänglichkeit erforderlich. Jedoch soll die Anzahl der thesaurusübergreifend polyhierarchischen Begriffe reduziert werden. Dafür wird eine Bewertung dieser Beziehungen durch den menschlichen Experten mittels Hypothesenziel spezifiziert.

Erreichte Verbesserung: Bei einer Stichprobe von 30 untersuchten thesaurusübergreifend polyhierarchischen Begriffen konnte der menschliche Experte in 26 Fällen die Begriffsintegration durch Umstrukturierungen so verbessern, dass die Polyhierarchie aufgelöst wurde. Es konnte für diese Stichprobe somit eine erhebliche Verbesserung erzielt werden.

Es sei angemerkt, dass trotz der durchschnittlichen guten Zugänglichkeit eine feingranularere Untersuchung der Zugänglichkeit zumindest für die Gruppenoberbegriffe von uns empfohlen wird. Im Rahmen der Experimente haben wir sie jedoch nicht durchgeführt, so dass keine Ergebnisse vorliegen.

12.5 Resümee

In diesem Kapitel haben wir gezeigt, wie anhand von Kennzahlen für die quantitative und qualitative Analyse für Thesaurusföderationen während des Integrationsprozesses Zielsetzungen in Form von Hypothesenzielen für den weiteren Verlauf der Integration hergeleitet werden können. Damit wurde die Voraussetzung geschaffen, das Integrationsergebnis schrittweise optimieren zu können. Diese Hypothesenziele können zum Teil von den Lösungsverfahren bearbeitet werden. In anderen Fällen wiederum dienen sie dazu, die Aufmerksamkeit des menschlichen Integrationsexperten auf Schwachstellen der Integration zu lenken, um diese mit seinem Eingreifen beseitigen zu können. So wird sichergestellt, dass die Föderation zielgerichtet weiter entwickelt wird.

Exemplarisch haben wir den konkreten Nutzen dieser Kennzahlen für die Verbesserung der Begriffsintegration in der Praxis zeigen können. Durch die Kennzahlen aufgedeckte Schwachstellen konnten zum einen direkt durch entsprechende maschinelle Integrationsverfahren und zum anderen durch das Fokussieren der Aufmerksamkeit des menschlichen Integrationsexperten auf entsprechende Föderationsausschnitte gezielt angegangen werden. Durch die von uns eingesetzten sehr einfachen maschinellen Verfahren konnte der menschliche Experten die Begriffsintegration fokussiert verbessern.

Des Weiteren wurden quasi als Nebenprodukt Kennzahlen entwickelt, die einen Vergleich verschiedener Thesaurusföderationen ermöglichen. Werden identische Komponententhesauri mit unterschiedlichen Verfahren zu einer Föderation integriert, können so die Ergebnisse der verschiedenen Ansätze komparativ evaluiert werden. Bei verschiedenen Thesaurusföderationen helfen die Kennzahlen einen ersten Eindruck von diesen zu erhalten und können auch die Basis für eine Integration von Thesaurusföderationen sein.

Kapitel 13

Ausführungsmaschine

Die in den vorangegangenen Kapiteln dargestellte Begriffsintegration stellt das Integrationswissen in Form von Inter-Thesaurus-Beziehungen, Ergänzenden Begriffen und Konfliktmarkierungen bereit. Zum Einsatz zur Anfrageformulierung, -bearbeitung und -erweiterung in Information Retrieval-Systemen wird zusätzlich eine Laufzeitumgebung benötigt, die anhand dieses Integrationswissens sowie des Anfrage-Kontextes thesaurusübergreifende Sichten auf das Vokabular ermöglicht. In diesem Kapitel entwickeln wir eine solche Ausführungsmaschine. Abbildung 13.1 zeigt die Bedeutung der Ausführungsmaschine als Schnittstelle bei der Verwendung der Thesaurusföderation durch den Benutzer.

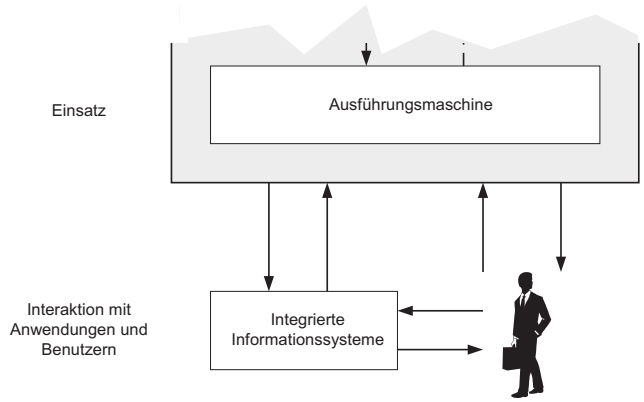

Abbildung 13.1: Ausführungsmaschine als Schnittstelle zwischen Benutzern und Integrationswissen / Komponentensystemen bei der Verwendung der Thesaurusföderation

13.1 Analyse

Bei Anfragen an ein übergeordnetes Informationssystem, das auf mehrere Komponenten-Informationssysteme zugreift (die verschiedene Thesauri zur Indexierung verwenden), kommen die Vorteile der Thesaurusföderation zum Tragen: Die übergeordnete Anfrageformulierungskomponente kann dem Benutzer bei der Auswahl von Begriffen das Gesamtvokabular (einschließlich der Nicht-Indexierungsthesauri, da diese das dem Benutzer vertraute Vokabular enthalten können) zur Verfügung stellen. Anhand der durch die Inter-Thesaurus-Relationen gegebenen Verknüpfungspunkte können die jeweiligen Anfragebearbeitungskomponenten der

Komponenten-Informationssysteme entsprechenden Deskriptoren als Eingabeparameter erhalten. In kritischen Situationen (sehr wenige oder keine Ergebnisse bzw. sehr viele Ergebnisse) kann eine übergeordnete Anfrageerweiterung basierend auf den Relationen der Thesaurusföderation (Inter- und Intra-Thesaurus-Relationen) stattfinden.

Das Ziel der Ausführungsmaschine ist es somit, im Netz an verschiedenen Orten zugängliche, heterogene Thesauri mithilfe von Informationen über den Kontext des Benutzers sowie dem bei der Begriffsintegration erstellten Integrationswissen dem Benutzer so anzubieten, dass er diese zur Formulierung von Suchanfragen an ein Informationsssytem sowie bei der Anfrageerweiterung nutzen kann. Die aus dieser Zielsetzung resultierenden Anforderungen wurden bereits in Abschnitt 2.3.4 identifiziert. Wir unterscheiden diese Anforderungen in Anforderungen

- an die Bereitstellung einer einheitlichen Zugriffsschnittstelle auf die Komponententhesauri (Überwindung der Entfernung und der Heterogenität) und
- an die eigentliche Anfragebearbeitung.

Bevor wir diese Bereiche näher analysieren, werden die Voraussetzungen und Annahmen definiert.

13.1.1 Voraussetzungen und Annahmen

Wir gehen davon aus, dass zum Zeitpunkt der Bearbeitung der Benutzeranfrage bereits bekannt ist, welche *Informationssysteme* angefragt werden sollen und welche Priorisierungen diese Informationssysteme durch den Benutzer erhalten[1]. Anhand der Metainformationen, die von der Thesaurusföderation verwaltet werden, kann somit die *geordnete Menge der Indexierungsthesauri* hergeleitet werden.

Weitere Voraussetzung ist, dass der Benutzer angibt, welche Thesauri er verwenden möchte. Dies können sowohl Indexierungsthesauri als auch Nicht-Indexierungsthesauri sein. Er ordnet diese Menge nach seinen Präferenzen, d.h. die *geordnete Menge aller benutzerrelevanten Thesauri* ist bekannt.

Zur Anfragebearbeitung ist es zudem erforderlich, dass durch den Benutzer festgelegt ist, ob ein guter Recall einer guten Precision bevorzugt wird, oder umgekehrt. Diese im Information Retrieval am häufigsten verwendeten Maße zur Beurteilung der Güte von Retrievalergebnissen (vgl. etwa [Sal87, Gau95] sowie die Definitionen in Anhang D.2) werden hier als Tendenz vorgegeben: Eine *Recall-Präferenz* bedeutet, dass der Benutzer tendenziell eher mehr relevante Dokumente geliefert bekommen möchte und dafür auch eine größere Anzahl nichtrelevanter Dokumente in Kauf nimmt. Umgekeht bedeutet eine *Precision-Präferenz*, dass der Benutzer tendenziell eher weniger nichtrelevante Dokumente geliefert bekommt und dafür auch das Fehlen von relevanten Dokumenten in Kauf nimmt.

Wir bezeichnen die Recall- bzw. Precision-Präferenz, die geordnete Menge der Informationssysteme, die geordnete Menge der Indexierungsthesauri sowie die geordnete Menge der benutzerrelevanten Thesauri auch als den *Kontext einer Anfrage*.

Weiterhin wird die technische Annahme getroffen, dass eine Vernetzung der unterschiedlichen Lokalisationen der Komponententhesauri und gemeinsame Kommunkikationsprotokolle bereits

[1] Diese Annahme mag in der Praxis einschränkend wirken. Jedoch gilt es zu bedenken, dass der Zugriff auf die Thesaurusföderation zum Zwecke der Information-Retrieval-Unterstützung eingebettet ist in ein Information-Retrieval-System, dem zumindest zum Zeitpunkt der Anfrage bekannt sein muss, welche Informationsquellen angefragt werden.

gegeben sind. Diese grundlegende Voraussetzung für verteilte Informationssysteme ist aktueller Stand der Technik.

13.1.2 Anfragebearbeitung

Jede Anfrage an die Thesaurusföderation muss ihren Kontext mitliefern. Aus dieser globalen Anfrage sowie dem Kontext müssen dann die Teilanfragen an die Subsysteme (Komponententhesauri, Thesaurus der Ergänzenden Begriffe, Integrationsdatenbank) hergeleitet werden. Es können folgende Anfragetypen unterschieden werden:

Suchanfragen beschreiben eine Suche nach Teilzeichenketten innerhalb der Benennungen (Deskriptoren und Nicht-Deskriptoren) der Thesaurusföderation.

Detailanfragen liefern zu einem Deskriptor alle Informationen (wie alle äquivalenten Deskriptoren aus anderen Thesauri, alle Nicht-Deskriptoren, Definition).

Navigationsanfragen beschreiben von einem Deskriptor ausgehend die Anforderung aller verwandten Begriffe bzw. aller Abstraktions- oder Bestandsober- oder -unterbegriffe bzw. der Gruppen.

Abbildungsanfragen beschreiben die Anforderung, eine Menge von Deskriptoren aus Nichtindexierungs- und Indexierungsthesauri auf eine Menge von Deskriptoren aus ausschließlich Indexierungsthesauri abzubilden.

Die von den Subsystemen erhaltenen Ergebnisse müssen unter Berücksichtigung des Integrationswissens (Inter-Thesaurus-Relationen, Konfliktmarkierungen) und des Kontextes schließlich zusammengefasst und als Ergebnis geliefert werden.

Die globalen Anwendungen der Thesaurusföderation greifen ausschließlich lesend auf die an der Föderation beteiligten Komponententhesauri zu. Konflikte durch Änderungen, die bei föderierten Datenbanksystemen prinzipiell berücksichtigt und vermieden bzw. aufgelöst werden müssen (vgl. z.B. [Con97]) treten deshalb nicht auf. In dieser Hinsicht ist die Handhabung von Thesaurusföderationen also einfacher als die Handhabung allgemeinerer föderierter Datenbanksysteme.

13.1.3 Einheitliche Zugriffsschnittstelle

Die im vorangegangenen Abschnitt 13.1.2 identifizierten Anfragetypen müssen von den unterschiedlichen Komponententhesauri beantwortet werden. Dazu sollen diese über eine einheitliche Zugriffsschnittstelle zugänglich gemacht werden, die

- die Heterogenität der Komponententhesauri verdeckt und
- die entfernt aufgerufen werden kann.

Zur Überwindung der Heterogenität der Datenverwaltungswerkzeuge und der Daten selbst sind die Schritte der *Schematransformation* und der *Schemaintegration* erforderlich. Unter Schematransformation wird die aufgrund der potenziell unterschiedlichen Datenmodelle der Komponententhesauri erforderliche Transformation in das Datenmodell der Föderation verstanden. Da die Autonomie der Föderations-Thesauri u.a. bedeutet, dass die lokalen Schemata nicht geändert werden dürfen, bedeutet Schemaintegration eine Homogensierung unter Berücksichtigung der heterogenen Schemata.

Das System soll bei der Überwindung der Heterogenität der Datenmodelle und -schemata so flexibel sein, dass eine Einbindung von Thesauri, die nur über eine HTML-Schnittstelle zugreifbar sind (z.B. AGROVOC über `http://www.fao.org/agrovoc/`), ebenso möglich sein soll wie die Integration von Thesauri in relationalen oder objektorientierten Datenbankmanagementsystemen.

Die Überwindung heterogener Betriebssystem-Plattformen erfordert plattformunabhängige Werkzeuge bzw. Middleware-Technologien.

13.2 Architektur

Die Basis einer Architektur für Thesaurusföderationen kann auf aktuelle Methoden und Technologien aufgesetzt werden. In unserem Ansatz greifen wir im Wesentlichen auf die Methoden und Technologien aus dem Bereich Föderierte Datenbanken [Con97] und des I^3-Projektes [Wie94, AHK+95, Wie96] zurück, die auf die lose Kopplung heterogener, autonomer Thesauri übertragen werden.

13.2.1 Übersicht

Um die Anforderungen an eine Ausführungsmaschine in einem modularen System umzusetzen wurde eine Mediator-/Kapsel-Architektur entwickelt (vgl. Abbildung 13.2): Die Aufgaben der Schematransformation und der Schemaintegration wird dabei durch die Kapseln wahrgenommen. Diese sehr generische Architektur erlaubt es, auf Komponententhesauri zuzugreifen, die z.B. in einem relationalen Datenmodell verwaltet werden und über eine SQL-Schnittstelle erreichbar sind oder auf Daten, die im Dateisystem gehalten werden und über eine HTTP/HTML-Schnittstelle abgefragt und präsentiert werden. Die Kapsel kann entweder direkt am Ort des Komponententhesaurus platziert werden (dies ist unbedingt erforderlich, wenn dieser bisher über keinen entfernten Zugriff verfügt) oder aber am Ort der Mediatoren (z.B. wenn auf einen Thesaurus über eine HTTP/HTML-Schnittstelle zugegriffen werden soll).

Die Kapseln bieten eine einheitliche Diensteschnittstelle an, auf die entfernt zugegriffen werden kann. Eine Implementierung kann hierzu inzwischen auf eine ganze Reihe von standardisierten Technologien zugreifen. Es aus Gründen der Einfachheit und Handhabbarkeit des Systems naheliegend, die hier ausgewählte Technologie nicht nur für die Kommunikation zwischen Mediatoren und Kapseln zu verwenden sondern auch für die Kommunikation zwischen weiteren Integrationsdiensten sowie der eigentlichen Anwendung.

Ein Thesaurusföderationsmediator bietet die gemeinsame Sicht auf die gesamte Thesaurusföderation an. Dazu kann er zum einen als Dienstnehmer auf die Kapseln der Komponententhesauri zugreifen, zum anderen auf das in der Föderationsdatenbank gespeicherte Integrationswissen (Metadaten über die beteiligten Thesauri, Inter-Thesaurus-Relationen, Konflikte). Der Thesaurusföderationsmediator ist verantwortlich für das Zerlegen der globalen Anfrage an die Thesaurusföderation in Anfragen an die Subsysteme sowie das Zusammenführen der Teilergebnisse.

Wie bereits in Abbildung 13.2 dargestellt, muss die technische Integration der Komponententhesauri zu einer Thesaurusföderation im gesamten Umfeld der Integration von Informationssystemen zu föderierten Informationssystemen betrachtet werden. Daher beinhaltet die Architektur nicht nur einen Thesaurusföderationsmediator sondern auch einen Informationssystemmediator, der die Integration der eigentlichen Informationssysteme bewerkstelligt. Als Beispiel für ein solch föderiertes Informationssystem – ohne aber die Herausforderung der Thesaurusintegration

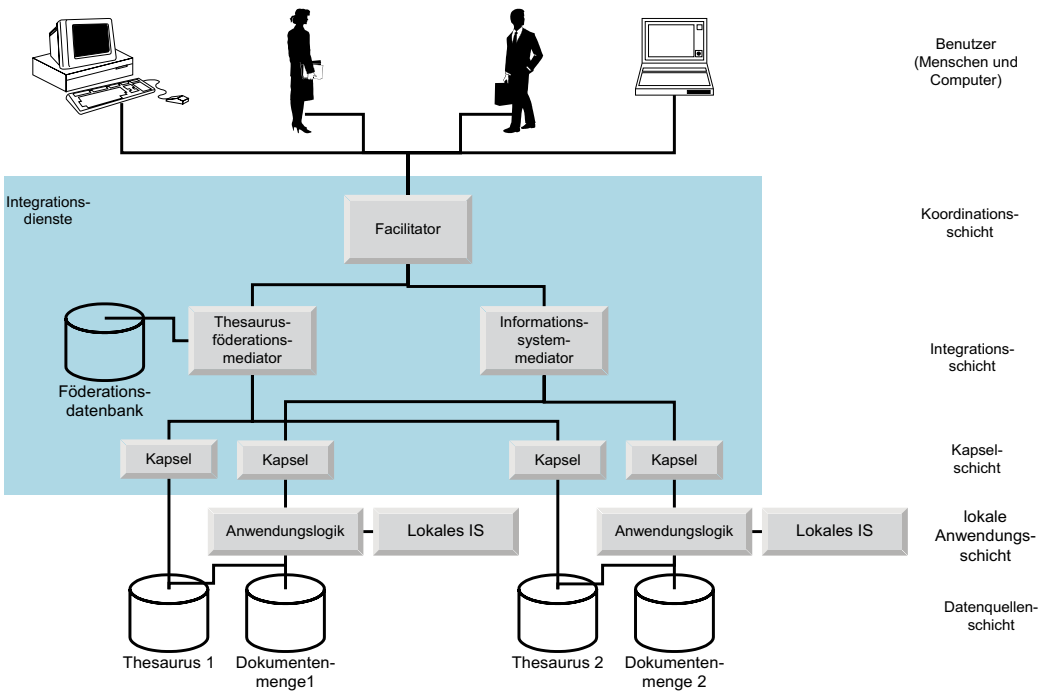

Abbildung 13.2: Architektur für föderierte Informationssysteme (IS) mit Thesaurusföderation

angegangen zu sein – sei WebCDS+ genannt [RKK99]. Der Informationssystemmediator verteilt die Anfragen nach den Anfragekriterien entsprechenden Dokumenten an die verschiedenen Informationssysteme und liefert eine gemeinsame Ergebnismenge.

Die Koordinierung des Thesaurusföderationsmediators und des Informationssystemmediators übernimmt in der entwickelten Architektur ein Facilitator. Dieser nimmt die Anfragen des Benutzers oder anderer Systeme entgegen, entscheidet, ob eine Anfrage an die Thesaurusföderation erforderlich ist (z.B. um bei einer semantischen Suche entsprechende Föderationsbegriffe zu suchen und deren Entsprechungen in den Komponententhesauri) und reicht ggf. die Ergebnisse der Anfrage an die Thesaurusföderation weiter an den Informationssystemmediator.

Da der Facilitator und der Informationssystemmediator nur am Rande dieser Arbeit interessant sind, werden sie im Folgenden nur kurz erläutert. Der Thesaurusföderationsmediator und die Thesauruskapseln werden ausführlich dargestellt.

13.2.2 Kommunikationsschnittstellen und -formate

Als plattformübergreifender, offener Standard zum Informationsaustausch zwischen verschiedenen Systemen hat sich XML [Wor98] etabliert. Als Basis für interoperable Software wird daher auch in unserer Architektur XML als Kommunikationsformat eingesetzt. Die als genereller Nachteil von XML – aus Gründen der erwünschten Lesbarkeit von XML-Daten – betrachtete „Gesprächigkeit von XML“ (großer syntaktischer Overhead) fällt bei unserer Architektur nicht ins Gewicht, da zum einen die übertragene Datenmenge wenig umfangreich ist, zum anderen inzwischen auch Kompressionstechniken entwickelt wurden, die eingesetzt werden können [Cov01].

Für XML wurden und werden auch eine Reihe von Anfragesprachen entwickelt [FSW00]. Die bedeutendste ist die vom World Wide Web-Konsortium standardisierte XML-Anfragesprache XQuery [Wor01]. XQuery ist eine mächtige, dabei aber konzeptionell einfach gehaltene Sprache

auf gutem theoretischen Fundament[2]. Weiterer Vorteil ist die gute Integration mit anderen XML-Standards wie XPath oder XML-Schema.

Die Formulierung von Anfragen mit XQuery erfordert aber – wie auch bei anderen XML-Anfragesprachen – die Kenntnis aller beteiligten Schemata. Die Schemata der Komponententhesauri sollen bei der Formulierung einer Anfrage aber nicht bekannt sein. Nun könnte stattdessen die Anfrage gegen ein globales Schema formuliert werden, und die Aufgabe der Kapseln wäre es, diese Anfragen zu transformieren in Anfragen gegen die lokalen Schemata. Das setzte aber Voraus, dass die Komponententhesauri bereits in der Lage wären, Anfragen z.B. in XQuery zu beantworten, also als XML-Quelle zugänglich wären. Diese Voraussetzung kann aufgrund der Anforderung, mit einem möglichst breitem Spektrum von Thesauri umgehen zu können, die in unterschiedlichsten Modellen vorliegen (z.B. relationale Modelle und objektorientierte Modelle), nicht erfüllt werden. Die XQuery-Anfrage müsste daher in eine – oder eine Folge von – Nicht-XML-Anfragen übersetzt werden, womit die Vorteile von XQuery nicht zum Tragen kommen.

Da in den nächsten Jahren auch weiter nicht davon auszugehen ist, dass die überwiegende Anzahl von Thesauri über eine XML-Anfragesprache verfügbar ist, verzichten wir auf eine Formulierung der Anfragen in einer XML-Anfragesprache. Statt direkt in den Anfragen zu formulieren, woher (aus welchen Schema-Bereichen) die Ergebnisse kommen, spezifizieren wir nur, was geliefert werden soll (Anhand des Anfragetyps und der Ergebnisstruktur) und was die Eingabeparameter sind, die u.a. die Quellen in Form von Komponententhesauri spezifizieren. Diese abstrakte Beschreibung der Anfrage kann mit einem Funktionsaufruf verglichen werden – die Interna der Funktion, wie und woher die Ergebnisse erzeugt werden, werden jedoch durch die Funktion gekapselt. In unserer Architektur geschieht dies durch den Mediator, der die Anfrage an die relevanten Komponententhesauri weiterreicht und durch die Kapseln der Komponententhesauri, die die abstrakte Anfrage in konkrete Anfragen an die Komponententhesauri umsetzen sowie aus den Ergebnissen die Ergebnisse in der global definierten Ergebnisstruktur erzeugen. Nur diese Kapseln benötigen daher Kenntnis der internen Strukturen und Schnittstellen der Komponententhesauri.

Die Definition der Ergebnisstrukturen kann mittels XML-Schema erfolgen. XML-Schema erlaubt sowohl die Festlegung einer Struktur, Typisierungen und Kardinalitäten als auch einfache Vererbung. Zu berücksichtigen ist jedoch, dass XML-Schema keine Modellierungssprache auf semantischer Ebene, sondern eine Schema-Definitionssprache auf logischer Ebene ist. Zur Modellierung wird daher weiterhin auf UML zurückgegriffen und aus diesem UML-Modell dann das XML-Schema erzeugt. Bei einfachen Schemata wie Rückgabelisten von Deskriptoren kann das XML-Schema aus dem UML-Modell weitestgehend automatisch erzeugt werden.

13.2.3 Protokolle

Als Protokoll für die XML-Kommunikation wird SOAP (Simple Object Access Protokoll, [Wor00b]) eingesetzt, das sich als Standard für entfernte Funktionsaufrufe auf XML-Basis z.B. gegenüber XML-RPC durchgesetzt hat. SOAP erhält entsprechende Unterstützung aller wichtigen Plattform- und Systemhersteller und kann daher sprach- und systemübergreifend eingesetzt werden. SOAP ist es gelungen, durch Nutzung existierenden Standards wie HTTP und XML ein einheitliches und erweiterbares Protokoll zu schaffen, welches auch in Umgebungen funktioniert (und gerade dort), wo herkömmliche Entfernte-Funktionsaufruf-Mechanismen (RPC-Mechanismen) z.B. aufgrund der Filterung durch Firewalls scheitern. SOAP erfüllt somit die Anforderung, dass sehr unterschiedliche Systeme über Weitverkehrsnetze miteinander kom-

[2]XQuery basiert wie SQL auf einer Algebra.

munizieren sollen. Die potenziell kritische Geschwindigkeit der Datenübertragung von SOAP spielt in unserer Architektur keine vorrangige Rolle, da das Datenvolumen verhältnismäßig klein ist.

Abbildung 13.3 zeigt den Einsatz von SOAP in unserer Architektur.

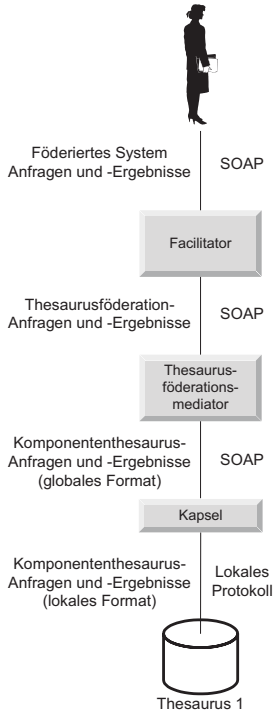

Abbildung 13.3: SOAP als Protokoll zwischen den Komponenten des föderierten Informationssystem

13.3 Thesaurusföderationsmediator

Der Thesaurusföderationsmediator bietet die Dienste an, die erforderlich sind, um die Menge der Komponententhesauri als Föderation darzustellen. Somit nimmt der Thesaurusföderationsmediator eine zentrale Rolle ein. Neben den eigentlichen Anfragediensten (vgl. Abschnitt 13.1.2) stellt er Auskunftsdienste bereit, die Informationen über die beteiligten Informationssysteme und Thesauri liefern. Innerhalb der Auskunftsdienste wird ausschließlich auf das Integrationswissen – einschließlich eines Repositoriums, das die Zugangsparameter zu den Komponententhesauri enthält – innerhalb der Föderationsdatenbank zugegriffen. Die Anfragedienste benötigen über die Kapseln zusätzlich Zugriff auf die Komponententhesauri.

13.3.1 Schnittstellen

Der Thesaurusföderationsmediator soll also Anfragen *über* und *an* die Thesaurusföderation beantworten können. Dazu besitzt er eine Auskunftsschnittstelle und eine Anfrageschnittstelle, die jeweils eine Reihe von Diensten beinhalten, vgl. Tabelle 13.1 (Notation: <Dienstname>(<Parametertyp> $\star$):<Ergebnistyp>, [] hinter einem Typbezeichner bedeutet eine Liste von Elementen des Typs, durch | getrennte Aufzählungen notieren entweder-oder-Konstrukte).

Auskunftsdienste	
getFederation():Federation	liefert die Föderation (Name und Identifikator)
getAllThesauri():Thesaurus[]	liefert alle Thesauri der Föderation
getIndexingThesauri():Thesaurus[]	liefert alle Indexierungsthesauri der Föderation
getIndexingThesauri(InfoSystem[]): Thesaurus[]	liefert zu einer Menge von Informationssystemen alle Indexierungsthesauri
getNonIndexingThesauri():Thesaurus[]	liefert alle Nicht-Indexierungsthesauri der Föderation
getInfoSystem():InfoSystem[]	liefert alle Informationssysteme der Föderation
getInfoSystem(Thesaurus[]):InfoSystem[]	liefert zu einer Menge von Thesauri alle Informationssysteme
Anfragedienste	
lookupTerm(String, Thesaurus[]): Term[]	Suchanfrage, die alle Bezeichnungen (Terme) innerhalb der gegebenen Menge von Thesauri liefert, die die gegebene Zeichenkette enthalten
getDetail(Term, Thesaurus[]): Descriptor[]	Detailanfrage, die zu einer Bezeichnung den oder die (bei Benutze-Kombination-Beziehungen) zugehörige(n) Deskriptor(en) aus den gegebenen Thesauri liefert. Der Deskriptor enthält alle Informationen wie zugehörige Nicht-Deskriptoren, äquivalente Deskriptoren aus anderen Thesauri der gegebenen Thesaurusmenge, Definitionen und Identifikatoren. Wir bezeichnen diese Struktur auch als Föderierten Deskriptor.
getBroader(Descriptor|Group, Thesaurus[]): Descriptor[]|Group[]|Federation	Navigationsanfrage, die übergeordnete Elemente eines Deskriptors bzw. einer Gruppe innerhalb der gegebenen Menge von Thesauri liefert. Die Ergebnisse selber wiederum können Deskriptoren, Gruppen oder die Föderation sein.
getBroaderAbstract(Descriptor, Thesaurus[]): Descriptor[]	Navigationsanfrage, die Abstraktionsoberbegriffe eines Deskriptors liefert.
getBroaderPartitive(Descriptor, Thesaurus[]): Descriptor[]	Navigationsanfrage, die Bestandsoberbegriffe eines Deskriptors liefert.
getNarrower(Descriptor|Group| Federation, Thesaurus[]): Descriptor[]|Group[]	Navigationsanfrage, die untergeordnete Elemente eines Deskriptors bzw. einer Gruppe bzw. der Föderation innerhalb der gegebenen Menge von Thesauri liefert. Die Ergebnisse selber wiederum können Deskriptoren, Gruppen oder die Föderation sein.
getNarrowerAbstract(Descriptor, Thesaurus[]): Descriptor[]	Navigationsanfrage, die Abstraktionssunterbegriffe eines Deskriptors liefert.
getNarrowerPartitive(Descriptor, Thesaurus[]): Descriptor[]	Navigationsanfrage, die Bestandsunterbegriffe eines Deskriptors liefert.
getRelated(Descriptor, Thesaurus[]): Descriptor[]	Navigationsanfrage, die Assoziationsbegriffe eines Deskriptors liefert.
getGroup(Descriptor, Thesaurus[]): Group[]	Navigationsanfrage, die Gruppen eines Deskriptors liefert.
map(Descriptor, Thesaurus[]): DescriptorExpression	Abbildungsanfrage, die einen Deskriptoren auf einen Ausdruck von Deskriptoren ausschließlich aus den gegebenen Thesauri abbildet. Als Ausdruck werden UND- und ODER-Verknüpfungen zugelassen, die geklammert werden können.

Tabelle 13.1: Dienste der Schnittstellen des Thesaurusföderationsmediators

Anhang C zeigt ein Beispiel für die Repräsentation einer solchen Anfrage als SOAP-Anfrage.

13.3.2 Anfragebearbeitung

Alle Auskunftsdienste können ausschließlich mit den Metainformationen der Föderation beantwortet werden, die Anfragebearbeitung ist daher trivial.

Die Suchanfrage *lookupTerm* erfordert das Weiterleiten der Anfrage an alle gegebenen Komponententhesauri sowie den Thesaurus der Ergänzenden Begriffe sowie das Vereinigen aller Ergebnisse. Da die Ergebnisteilmengen disjunkt sind und Integrationswissen per Spezifikation der Funktion nicht erforderlich ist, ist auch hier die Anfragebearbeitung trivial.

Detailanfragen, Navigationsanfragen und Abbildungsanfragen erfordern eine komplexere Anfragebearbeitung die wir in den folgenden Abschnitten 13.3.2.1 bis 13.3.2.3 vorstellen. Ein wichtiger Aspekt ist hier die Art und Weise, in der markierte Konflikte behandelt werden. Diese kann in einer Konfliktbehandlungspolitik festgelegt werden. Wir verfolgen im weiteren folgende Konfliktbehandlungspolitik:

- Abstraktionsredundanz, Zyklen, Beziehungstypdifferenzen, Schwesternassoziationen und Abstraktionsniveaudifferenzen sollen aufgelöst werden.
- Gemischte Abstraktions- und Bestandsunterbegriffe werden toleriert und erfordern keine Konfliktauflösung.

Diese exemplarisch verfolgte strikte Konfliktbehandlungspolitik wird gewählt, da die Konfliktbehandlung weniger strikte Politiken daraus durch Reduktion der Aktionen abgeleitet werden kann. Als Beispiel für eine solche Duldung von Konflikten lassen wir die Heterogenität von Hierarchiebeziehungen (gemischte Abstraktions- und Bestandsunterbegriffe) zu.

Hauptvorteil einer solchen expliziten Konfliktbehandlungspolitik ist die gewonnene Flexibilität. Die Politik kann für verschiedene Nutzergruppen oder von jedem Nutzer individuell spezifiziert werden. Die gewählte Politik entspricht der Standardeinstellung, die die üblichen Erwartung an einen Thesaurus ausdrückt (vgl. [Wer85]).

13.3.2.1 Detailanfragen

Um zu einer Bezeichnung aus einem Thesaurus den entsprechenden Föderierten Deskriptor zu ermitteln sind folgende Schritte erforderlich:

1. Falls es sich um einen Nicht-Deskriptor aus einem Komponententhesaurus handelt, wird dieser mithilfe der Intra-Thesaurus-Äquivalenzrelation bzw. der Intra-Thesaurus-Benutze-Kombination-Relation auf den bzw. die entsprechenden Deskriptor(en) abgebildet. Dazu wird eine *getDetail*-Anfrage an den entsprechenden Komponententhesaurus gestellt.
2. Zu dem bzw. den ermittelten Deskriptor(en) werden anhand der Inter-Thesaurus-Äquivalenzrelation und der Inter-Thesaurus-Benutze-Kombination-Relation die äquivalenten Begriffe aus der Menge der gegebenen Thesauri ermittelt:
   - Anhand der Föderationsdatenbank können die Identifikatoren der Deskriptoren ermittelt werden.
   - Mit den Deskriptor-Identifikatoren können durch *getDetail*-Anfragen an die Komponententhesauri die vollständigen Deskriptor-Informationen ermittelt werden.

Wir aus Abschnitt 11.2.2 ersichtlich wird, sind Inter-Thesaurus-Äquivalenz- bzw. -Benutze-Kombination in keinem Fall Bestandsteil einer Konfliktmarkierung. Bei der Bearbeitung von Detailanfragen ist es daher nicht erforderlich, die Konfliktmarkierungen zu betrachten.

13.3.2.2 Navigationsanfragen

Navigationsanfragen beschaffen zu einem Föderierten Deskriptor alle Föderierten Deskriptoren die durch den gegebenen Relationstyp mit diesem in Beziehung stehen ohne dass diese Beziehung zur Auflösung eines Konfliktes entfernt werden soll. Navigationsanfragen für die verschiedenen Relationstypen können ähnlich behandelt werden. Wir betrachten exemplarisch die zur Beantwortung einer *getNarrowerAbstract*-Anfrage erforderlichen Schritte der Anfragebearbeitung:

1. Die Ergebnismengen V_1 und V_2 werden als leere Menge erzeugt.

2. Aufgrund der implizierten Beziehungen ist es ausreichend, *einen* Deskriptor aus der Menge der Komponententhesaurus-Deskriptoren, die den Föderierten Deskriptor repräsentieren, herauszugreifen, und diesen als Basisdeskriptor x_1 zu verwenden. Von diesem Basisdeskriptor ausgehend werden alle in Beziehung stehenden Deskriptoren bestimmt.

3. Zu dem Basisdeskriptor werden alle Abstraktionsunterbegriffe $x_2, \ldots, x_m$ innerhalb des Komponententhesaurus bestimmt, indem eine *getNarrowerAbstract*-Anfrage an diesen gestellt wird.

4. Für jedes Tupel (x_1, x_i), $1 < i \leq m$, wird geprüft, ob es Teil einer Konfliktmarkierung, also Element in einer v-Menge (Konfliktverursacher) einer Konfliktmarkierung, ist. Eine Konfliktmarkierung ist nur dann relevant, wenn auch alle anderen Tupel der entsprechenden v-Menge aus den benuzterrelevanten Thesauri stammen[3].

   - Ist das Tupel nicht an einem Konflikt beteiligt, wird x_i der Ergebnismenge V_1 hinzugefügt.
   - Ist das Tupel an einem oder mehreren Konflikten beteiligt, wird x_i nur dann der Ergebnismenge V hinzugefügt, wenn die entsprechende Beziehung oder der Deskriptor nicht zur Auflösung des Konfliktes entfernt wird. Falls der Deskriptor entfernt wird, wird für x_i *getNarrowerAbstract* aufgerufen und diese Ergebnisse werden ebenfalls daraufhin überprüft, ob die Beziehungen oder Deskriptoren zur Konfliktauflösung entfernt werden. Die Überprüfung wird detailliert im Anschluss an diese Erläuterung der Schritte der Anfragebearbeitung dargestellt.

5. Jedes Element aus V_1 wird durch eine Mediator-Detailanfrage *getDetail* (vgl. Abschnitt 13.3.2.1) um die äquivalenten Begriffe aus allen benutzerrelevanten Thesauri ergänzt.

6. Anhand des Integrationswissens der Föderationsdatenbank werden alle Inter-Thesaurus-Abstraktionsunterbegriffe sowie alle implizierten Intra-Thesaurus-Abstraktionsunterbegriffe $y_1, \ldots, y_n$ aus den benutzerrelevanten Thesauri bestimmt.

[3] Markierte Konflikte, die Konfliktverursacher aus anderen Thesauri als den benutzerrelevanten Thesauri enthalten sind bei Betrachtung des durch die benutzerrelevanten Thesauri gegebenen Ausschnitts der Thesaurusföderation keine Konflikte. Wird in der Konfliktbehandlungspolitik spezifiziert, gemischte Abstraktions- und Bestandsunterbegriffe nicht zu tolerieren, sind Konflikte dieses Typs relevant, wenn Deskriptoren aus benutzerrelevanten Thesauri zugleich in r_1- und r_2-Mengen aufgeführt sind.

7. Nach den gleichen Regeln wie für Intra-Thesaurus-Abstraktionsunterbegriffe (vgl. 4.) wird für jedes y_i, $1 \leq i \leq n$, geprüft, ob es in die Ergebnismenge V_2 aufgenommen wird oder nicht.

8. Jedes Element aus V_2 wird durch eine Mediator-Detailanfrage um die äquivalenten Begriffe aus allen benutzerrelevanten Thesauri ergänzt.

9. Die Vereinigung von V_1 und V_2, bei der Duplikate (in diesem Fall Föderierte Abstraktionsunterbegriffe, die durch Deskriptoren aus mindestens zwei verschiedenen Thesauri repräsentiert werden, die zugleich durch eine Inter-Thesaurus-Äquivalenzbeziehung miteinander verbunden sind) entfernt werden, wird schließlich als Ergebnis geliefert.

Entscheidender Teil der Anfragebearbeitung ist also die Feststellung, ob eine Beziehung zwischen den Deskriptoren eines Tupels (x, y) zur Auflösung eines Konfliktes entfernt werden muss oder nicht. Die Konfliktauflösung abhängig vom Typ des Konfliktes durchgeführt:

Abstraktionsredundanz: Als redundant markierte Beziehungen werden auf jeden Fall entfernt. Alternativen werden nicht betrachtet, da die ergänzten (längeren) Pfade als Erfolg der Begriffsintegration verstanden werden.

Zyklen: Zur Auflösung von Zyklen genügt es, eine einzige Beziehung zu entfernen. Da den Inter-Thesaurus-Beziehungen eine so große Bedeutung zugemessen wurde, dass zu markierende Konflikte in Kauf genommen wurden, soll die zu entfernende Beziehung möglichst eine Intra-Thesaurus-Beziehung sein.

Bezeichnet das Tupel in der Konfliktauflösermenge s der Konfliktmarkierung eine Intra-Thesaurus-Beziehung, die nicht aus dem Komponententhesaurus mit höchster Priorität stammt, wird die entsprechende Beziehung entfernt (vgl. Beispiel 13.1, Alternative 1). Das bedeutet, dass die Standard-Konfliktauflösung greift, wenn nicht die Beziehung innerhalb des vom Benutzer als bedeutendsten eingeordnetem Thesaurus entfernt werden muss.

Andernfalls gilt es, eine neue zu entfernende Beziehung zu bestimmen. Hierzu wird aus der Menge v das Tupel bestimmt, das aus dem Thesaurus mit der niedrigsten Priorität stammt, wobei die niedrigste und die höchste Priorität nicht identisch sein dürfen (d.h. am Zyklus ist mindestens eine Intra-Thesaurus-Beziehung beteiligt, die nicht aus dem höchstpriorisierten Thesaurus stammt). Kommen mehrere Tupel in Frage, wird das Tupel ausgewählt, dass im Hierarchiepfad dem Tupel aus s, der Standardauflösung, als nächstes folgt. Kann keine andere Intra-Thesaurus-Beziehung entfernt werden, wird die Inter-Thesaurus-Beziehung entfernt, die im Hierarchiepfad der Standardauflösung als nächstes folgt und nicht durch eine Intra-Thesaurus-Beziehung des am höchsten priorisierten Thesaurus impliziert wird (vgl. Beispiel 13.1, Alternative 2 sowie Beispiel 13.2).

Die Entscheidung für die zu entfernende Beziehung ist bei gleichem Kontext somit deterministisch.

Beziehungstypdifferenzen: Beziehungstypdifferenzen werden aufgelöst, indem die in r_1 oder r_2 angegebene Beziehung entfernt wird (vgl. Abschnitt 6.3.7.1). Es wird diejenige Beziehung ausgewählt, die aus dem niedriger priorisierten Komponententhesaurus stammt bzw. falls dieser mit dem höchst priorisierten Komponententhesaurus identisch ist (also an dem Konflikt keine Intra-Thesaurus-Beziehung aus einem anderen Thesaurus beteiligt ist), die Inter-Thesaurus-Beziehung (vgl. Abbildung 6.12, S. 96).

Sollte der Konflikt ausschließlich durch Inter-Thesaurus-Beziehungen verursacht werden (vgl. Abbildung 6.11, S. 95), wurde die Konfliktauflösung bereits in der Realisierungsphase gefordert.

Schwesternassoziationen: Die Assoziationsbeziehung zwischen den durch Inter-Thesaurus-Hierarchie-Beziehungen zu Schwestern gewordenen Deskriptoren wird auf jeden Fall entfernt, da die somit vorhandenen Hierarchiebeziehungen als stärker als die Assoziationsbeziehung betrachtet werden.

Abstraktionsniveaudifferenzen: Unterschiedliche Abstraktionsnievaus, die auf Redundanzen zurückgeführt werden können, sind ausschließlich als Abstraktionsredundanzen markiert (vgl. Abschnitt 6.3.6.3, S. 92ff). Somit verbleiben als markierte Abstraktionsniveaudifferenzen solche Fälle, bei denen es zwischen zwei Deskriptoren zwei unterschiedlich lange Abstraktionspfade gibt, wobei die Länge beider Pfade größer als eins ist (vgl. Abbildung 6.10 links, S. 92). Es sind somit folgende Alternativen zu betrachten:

Alternative 1: Bereits während der Thesaurusföderationserstellung wurden zur Auflösung der Niveaudifferenz ein oder mehrere Ergänzende Begriffe eingeführt. Damit ist über den oder die Ergänzenden Begriffe ein Pfad der Länge entstanden, die identisch mit der Länge des längeren Pfades ohne die Ergänzenden Begriffe ist. Der oder die redundant gewordenen Beziehungen sind entsprechend markiert und werden entfernt (vgl. Beispiel 13.3).

Alternative 2: Die Abstraktionsniveaudifferenz wurde während der Föderationserstellung nicht durch einen oder mehrere Ergänzende Begriffe aufgelöst, sondern ausschließlich markiert. Das es ohne die Unterstützung eines menschlichen Experten nicht möglich ist, Ergänzende Begriffe einzufügen, gibt es somit zum Zeitpunkt der Anfragebearbeitung nur die Möglichkeit, die Länge des längeren Pfades der Länge des kürzeren Pfades anzupassen, d.h. *Deskriptoren werden entfernt*. Die Entfernung eines Deskriptors b aus einem Abstraktionspfad (a, b), (b, c), (c, d) bedeutet, dass als Unterbegriffe von a alle Unterbegriffe von b geliefert werden. Die Konfliktmarkierungen bestimmen den oder die zu entfernenden Deskriptor(en) durch die Beziehungen zu und von diesem bzw. diesen (vgl. Beispiel 13.4).

Alternative 2a: In der Konfliktauflösermenge r ist kein Deskriptor aus dem höchstpriorisierten Thesaurus als zu entfernen markiert, d.h. Deskriptoren aus diesem Thesaurus sind in allen Tupeln entweder als erstes oder als zweites Element oder gar nicht aufgeführt. Somit werden die durch die Standardauflösung spezifizierten Deskriptoren entfernt.

Alternative 2b: In der Konfliktauflösermenge r ist ein Deskriptor aus dem höchstpriorisierten Thesaurus als zu entfernen markiert, d.h. Deskriptoren aus diesem Thesaurus sind als erstes und als zweites Element der Tupel aufgeführt. Falls alternative Deskriptoren gefunden werden können, werden diese entfernt, ansonsten müssen die Deskriptoren aus dem höchstpriorisierten Thesaurus entfernt werden. Ein Deskriptor x ist ein alternativ zu entfernender Deskriptor, wenn er folgende Bedingungen erfüllt:

- x stammt nicht aus dem höchstpriorisierten Thesaurus.
- x ist ausschließlich Teil des längeren Pfades.
- x steht in keiner Inter-Thesaurus-Äquivalenzbeziehung mit einem Deskriptor des höchstpriorisierten Thesaurus.

Kommen mehr als erforderliche Anzahl alternativ zu entfernender Deskriptoren in Frage, werden diejenigen ausgewählt, die aus dem niedrigstpriorisierten Thesauri stammen und in der Hierarchie möglichst weit unten sind, da dies als geringerer Eingriff in die Struktur des Thesaurus betrachtet wird.

Gemischte Abstraktions- und Bestandsunterbegriffe: Laut der in Abschnitt 13.3.2, S. 237, dargestellten Konfliktbehandlungspolitik wird diese Art von Konflikt toleriert. Das bedeutet, die Beziehung zwischen x und y soll nicht entfernt werden.

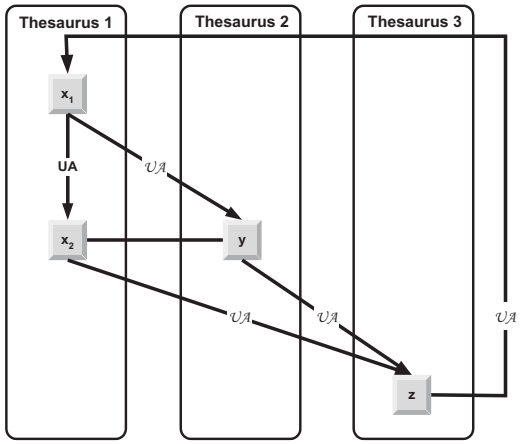

Abbildung 13.4: Ausschnitt aus Thesaurusföderationsgraph mit zwei Abstraktionszyklen

Beispiel 13.1 *Abbildung 13.4 zeigt einen Ausschnitt aus einem Thesaurusföderationsgraphen der zwei Abstraktionszyklen enthält: Zum einen den Pfad* (x_1, x_2), (x_2, z), (z, x_1) *und zum anderen den Pfad* (x_1, y), (y, z), (z, x_1). *Wir nehmen folgende Konfliktmarkierungen* k_1 *und* k_2 *an:* $k_1.v$ *und* $k_2.v$ *enthalten die bereits aufgeführten Konfliktverursacher, die* r_1*-Mengen enthalten die Inter-Thesaurus-Abstraktionsbeziehungen ohne den Standard-Konfliktauflöser, d.h.* $k_1.r_1 = \{(x_2, z), (z, x_1)\}$, $k_2.r_1 = \{(y, z), (z, x_1)\}$, *und die* r_2*-Mengen enthalten die von* x_1 *ausgehenden Beziehungen* $k_1.r_2 = \{(x_1, x_2)\}$, $k_2.r_2 = \{(x_1, y)\}$. *Die s-Mengen sind jeweils mit den* r_2*-Mengen identisch, d.h. die Standardkonfliktauflösung besteht darin, die von* x_1 *ausgehende Beziehung zu entfernen*[4].

Alternative 1: *Thesaurus 1 ist nicht der vom Benutzer am höchsten priorisierte Thesaurus. Damit greift die Standardkonfliktauflösung, so dass die in s angegebene Beziehung entfernt wird.*

Alternative 2: *Thesaurus 1 ist der vom Benutzer am höchsten priorisierte Thesaurus. Damit greift die Standardkonfliktauflösung nicht. Da bei beiden Konflikten keine andere Intra-Thesaurus-Beziehung in v angegeben ist, wird die der Standardkonfliktauflösung nächste Inter-Thesaurus-Beziehung entfernt, die nicht durch eine Intra-Thesaurus aus dem am höchsten priorisierten Thesaurus impliziert wird. Das bedeutet für Konflikt 1 wird die Beziehung zwischen* (x_2, z) *und für Konflikt 2 die Beziehung zwischen* (y, z) *entfernt.*

[4]Voraussetzung ist eine konsistente Konfliktmarkierung. Das bedeutet, dass zur Auflösung von verschiedenen Konflikten des gleichen Typs, an denen ein Deskriptor beteiligt ist, die von ihm ausgehenden Beziehungen zu äquivalenten Begriffen entweder immer oder nie Bestandteil der Konfliktauflöser-Mengen s sind. Diese Konsistens ist während der Realsierungsphase sicher zu stellen.

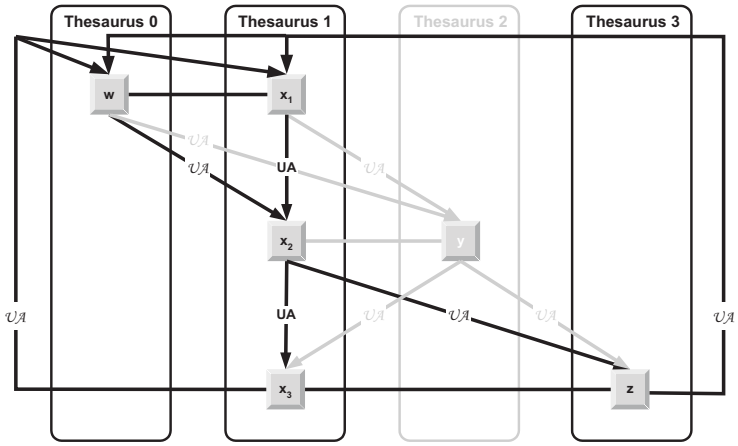

Abbildung 13.5: Ausschnitt aus Thesaurusföderationsgraph mit ausgeblendetem Thesaurus

Beispiel 13.2 *Ein komplexeres Beispiel mit einem ausgeblendetem, d.h. nicht benutzerrelevanten Thesaurus 2, zeigt Abbildung 13.5. Durch das Ausblenden des Thesaurus wird die Anzahl der Abstraktionszyklen von sechs auf vier reduziert. Prinzipiell wird bei der Konfliktauflösung vorgegangen, wie im vorangegangenen Beispiel beschrieben. Falls der Thesaurus 1 der präferierte Thesaurus ist, wird weder eine Intra-Thesaurus-Beziehung aus Thesaurus 1 entfernt, noch die Inter-Thesaurus-Beziehungen zwischen* (x_2, z) *und* (w, x_2)*, die durch eine Intra-Thesaurus-Beziehung in Thesaurus 1 impliziert werden. Stattdessen werden zur Auflösung der Konflikte die Beziehungen zwischen* (z, w)*,* (z, x_1)*,* (x_3, w) *bzw. die implizierte Intra-Thesaurus-Beziehung zwischen* (x_3, x_1) *entfernt. Implizierte Intra-Thesaurus-Beziehungen, die daran erkannt werden, dass sie statt in den Komponententhesauri in der Föderationsdatenbank gespeichert werden, werden also wie Inter-Thesaurus-Beziehungen behandelt.*

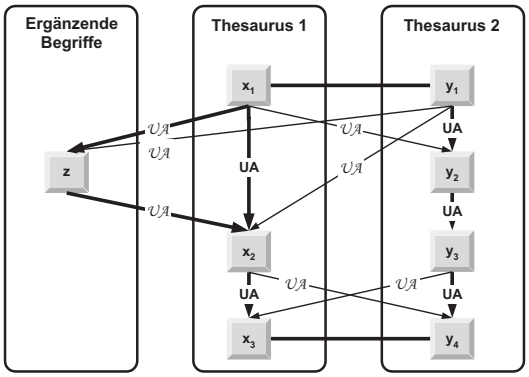

Abbildung 13.6: Auflösung einer Abstraktionsniveaudifferenz durch Ergänzenden Begriff

Beispiel 13.3 *Abbildung 13.6 zeigt eine Abstraktionsnieveaudifferenz, die durch Einfügen eines Ergänzenden Begriffs aufgelöst wurde. Durch den Ergänzenden Begriff sind zugleich die Beziehungen zwischen* (x_1, x_2) *und* (y_1, x_2) *als redundant markiert worden und werden daher von der Anfragebearbeitung entfernt. Weitere Aktionen zur Konfliktauflösung sind nicht erforderlich.*

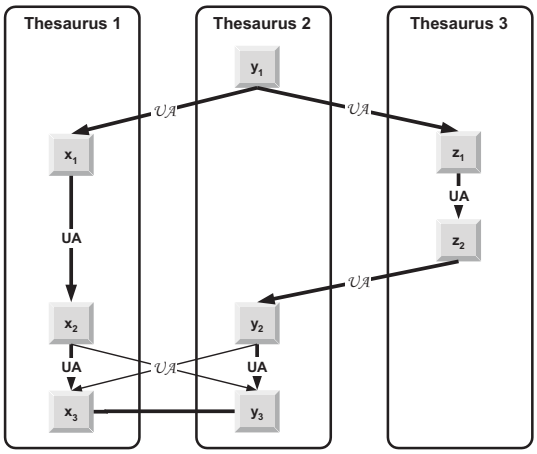

Abbildung 13.7: Abstraktionsniveaudifferenz ohne Auflösung durch Ergänzenden Begriff

Beispiel 13.4 *Ein Beispiel für eine markierte Abstraktionsniveaudifferenz ohne Auflösung durch einen Ergänzenden Begriff zeigt Abbildung 13.9. Die Abstraktionsniveaudifferenz ist zweifach markiert: Zum einen ist der Pfad von* y_1 *über* x_1, x_2 *nach* y_3 *um eins kürzer als der Pfad von* y_1 *über* z_1, z_2, y_2 *nach* y_3, *zum anderen ist ebenso der Pfad von* y_1 *über* x_1, x_2 *nach* x_3 *um eins kürzer als der Pfad von* y_1 *über* z_1, z_2, y_2 *nach* x_3. *Zur Konfliktauflösung sind durch die Tupel* (y_1, z_1), (z_1, z_2) *jeweils alle Beziehungen zu* z_1 *angegeben, d.h. der Deskriptor* z_1 *soll entfernt werden. Ist Thesaurus 3 nicht der höchstpriorisierte Thesaurus wird dieser Standardkonfliktauflösung gefolgt, d.h.* getNarrowerAbstract *liefert für* y_1 *neben* x_1 *statt* z_1 *dessen Unterbegriff* z_2. *Andernfalls wird* y_2 *als alternativer zu entfernender Deskriptor bestimmt, d.h.* getNarrower-Abstract *liefert für* y_1 *neben* x_1 *auch* z_1 *und für* z_2 *die Deskriptoren* x_3 *und* y_3.

Die Konfliktauflösung geschieht somit kontextabhängig und ermöglicht eine an die Anforderung des Benutzers angepasste Darstellung der Thesaurusföderation:

- Zur Feststellung der aufzulösende Konflikte wird die Menge der benutzerrelevanten Thesauri herangezogen. Konflikte, die durch andere (zusätzliche) Thesauri entstehen, werden nicht als aufzulösende Konflikte betrachtet.
- Die Konflikte werden entsprechend der Konfliktbehandlungspolitik, die generell festlegt welche Konflikte aufgelöst werden sollen, sowie der Priorität, die der Benutzer den verwendeten Thesauri gibt, aufgelöst. Falls möglich werden Beziehungen in als bedeutender eingestuften Thesauri Beziehungen in als weniger bedeutend eingestuften Thesauri bevorzugt. Damit kann erreicht werden, dass die höher priorisierten und dem Benutzer in der Regel vetrauteren Thesauri in der Zusammenschau mit den anderen Thesauri möglichst wenig modifiziert werden.Der vom Benutzer höchstpriorisierte Thesaurus bleibt weitgehend unmodifiziert.

In allen Fällen wird dasselbe konsistente Bild der Föderation geliefert unabhängig davon, zu welchem Deskriptor eine *getNarrower*-Anfrage abgesetzt wird.

Die Entscheidung, die zu entfernende Beziehung zu bestimmen, wird lokal getroffen[5], d.h. es werden nicht gleichzeitig alle Konfliktmarkierungen bei denen es nichtleere Schnittmengen mit

[5]Ausnahme ist die Konfliktbehandlung bei Abstraktionsniveaudifferenzen, falls diese zur Entwicklungszeit durch Ergänzende Begriffe aufgelöst wurden. In diesem Fall müssen zusätzlich Redundanzmarkierungen, die beim Einfügen der Ergänzenden Begriffe stattgefunden haben, berücksichtigt werden.

den Konfliktverursachern der untersuchten Konflikte gibt, untersucht. Die dadurch evtl. nicht optimalen Ergebnisse (evtl. werden mehr Beziehungen und Deskriptoren entfernt, als unbedingt erforderlich wäre) werden aus Gründen der Komplexitätsreduktion und der damit zur Laufzeit besseren Performanz in Kauf genommen.

13.3.2.3 Abbildungsanfragen

Nachdem der Benutzer seine Anfrage durch die Auswahl von Föderierten Deskriptoren aus der Menge der benutzerrelevanten Thesuri beschrieben hat, wird diese an das Föderierte Informationssystem gesendet, das enstprechend indexierte Dokumente liefern soll. Das Föderierte Informationssystem benötigt dazu statt der Deskriptoren aus den benuzterrelevanten Thesauri Deskriptoren aus den Indexierungsthesauri. Die entsprechende Abbildung wird durch die Funktion *map()* vorgenommen.

Enthält eine Anfrage mehrere Deskriptoren sind diese durch Operatorn wie *und*, *oder*, *nicht* miteinander verknüpft. Damit die *map()*-Funktion diese Operatoren nicht zu berücksichtigen braucht, wird sie zum Abbilden einzelner Deskriptoren verwendet. Für das Zerlegen der Anfrage in Einzelanfragen und das Zusammenfügen der Teilergebnisse zu einem Gesamtergebnis ist der Facilitator verantwortlich.

Ziel der Abbildungsanfrage ist es also, einen Föderierten Deskriptor aus den benutzerrelevanten Thesauri auf einen oder mehrere Föderierte Deskriptoren aus den Indexierungsthesauri abzubilden und die Semantik, d.h. den Begriffsinhalt, dabei möglichst wenig zu verändern. Ein erster Beitrag zu Bearbeitung der Abbildungsanfragen wurde in der Diplomarbeit [Tra97] geliefert und technisch umgesetzt. Dieses Konzept wird im Folgenden grundlegend erweitert.

13.3.2.3.1 Verwendung der Intra- und Inter-Thesaurusrelationen Wenn es Inter-Thesaurus-Äquivalenz- bzw. Benutze-Kombinationsbeziehungen zwischen Föderiertem Deskriptor des benutzerrelevanten Thesaurus und Deskriptoren der Indexierungsthesauri gibt, kann der Begriffsinhalt unverändert bleiben (die durch Benutze-Kombinationsbeziehungen gefundenen Deskriptoren werden dazu konjunktiv verknüpft). Um jedoch auch bei Nichtvorhandensein dieser Beziehungen die Deskriptoren übersetzen zu können, so dass sie in die Anfrage an das Föderierte Informationssystem einfließen können, wird die Modifikation des Begriffsinhaltes akzeptiert. Das bedeutet, das auch Inter- und Intra-Thesaurus-Hierarchie- und Assoziationsbeziehungen zur Abbildung herangezogen werden können. Schließlich ist es vorteilhaft, den ungefähren Inhalt eine Begriffes berücksichtigen zu können statt diesen Begriff ganz zu missachten.

Bei der Abbildung von Deskriptoren wird bei Nichtvorhandensein von Äquivalenz- bzw. Benutze-Kombinationsbeziehungen die Abstraktionsrelation bevorzugt gehandhabt. Dies wird begründet durch die Subsumption der Begriffe durch übergeordnete Abstraktionsbegriffe: Ein durch einen Deskriptor x repräsentierter Begriffes subsumiert per Definition der Abstraktionsrelation alle direkten Abstraktionsunterbegriffe. Das bedeutet, der Begriffsinhalt (vgl. Definition in Anhang D.3) des durch x repräsentierten Begriffes kann durch die konjunktive Verknüpfung seiner Abstraktionsoberbegriffe bzw. durch die disjunktive Verknüpfung seiner Abstraktionsunterbegriffe annähernd ausgedrückt werden (vgl. auch [Men98], S. 126ff). Im Falle der Abstraktionsoberbegriffe wird der Begriffsinhalt verkleinert, also kann ein Retrievalverfahren durch die unpräzisere Beschreibung mehr – bezogen auf den durch x repräsentierten Begriff sowohl relevante als auch nichtrelevante – Ergebnisse liefern. Im Falle der Abstraktionsunterbegriffe wird der Begriffsinhalt erweitert, also können weniger Ergebnisse geliefert werden, die aber bezogen auf den Inhalt des durch x repräsentierten Begriffs relevant sind.

Aus dieser Betrachtung folgt, dass bei Recall-Präferenz des Benutzers eine Abbildung durch Berücksichtigung der konjunktiven Verknüpfung der Abstraktionsoberbegriffe versucht wird. Bei Precision-Präferenz wird hingegen die Abbildung über die disjunktive Verknüpfung der Abstraktionsunterbegriffe versucht. Das Verfahren kann rekursiv fortgeführt werden: Jeder nicht auf einen Indexierungsthesaurus abgebildete Deskriptor kann versucht werden, durch Berücksichtigung seiner Abstraktionsober- resp. -unterbegriffe abzubilden. Konnte ein Deskriptor in die eine Richtung nicht auf mindestens einen Deskriptor aus einem Indexierungsthesaurus abgebildet werden, kann versucht werden, ihn in die andere Richtung abzubilden. Dabie gilt zu berücksichtigen, dass aufsteigend stets gefundene Deskriptoren konjunktiv verknüpft werden udn absteigend disjunktiv. Sobald eine erste Abbildung gefunden wird, kann bei der konjunktiven Verknüpfung für leere Mengen, die von *getBroader*- bzw. *getNarrower*-Anfragen geliefert wird die Konstante *wahr* und bei disjunktiven Verknüpfungen die Konstante *falsch* in den Ausdruck eingesetzt und dieser anschließend vereinfacht werden.

Gelingt anhand der Äquivalenz-, Benutze-Kombination und Abstraktionsbeziehungen keine Abbildung auf mindestens einen Deskriptor eines Indexierungsthesaurus, können entsprechend der Abstraktionsrelation auch sukzessive die Bestandsrelation und die Assoziationsrelation hinzugezogen werden. Hier gilt es aber zu berücksichtigen, dass die Begriffsinhalte stärker verändert werden und die entsprechend unscharfen Abbildungen zu weniger guten Anfrageergebnissen führen. Diese Unschärfe in der Übersetzung wird jedoch in Kauf genommen, um überhaupt aus dem Vokabular der benutzerrelevanten Thesauri in das Vokabular der Indexierungsthesauri übersetzen zu können. Ist der Benutzer mit den Ergebnissen nicht zufrieden, kann er über Anfrageverfeinerung und Anfrageerweiterung versuchen zu besseren Ergebnissen zu gelangen (vgl. Abschnitt 13.3.3).

In folgenden Fällen werden Bestandsrelation und Assoziationsrelation zur Abbildung berücksichtigt:

Hoher Recall und aufsteigend keine Abbildung gefunden: Wenn der Benutzer einen hohen Recall verlangt und über die Äquivalenrelation, Benutze-Kombination-Relation und Abstraktionsoberbegriffe keine Abbildung gefunden werden konnte, wird zuerst aufsteigend die gesamte Hierarchierelation (d.h. Abstraktions- vereinigt mit der Bestandsrelation) verwendet. Führt auch das nicht zu einer erfolgreichen Abbildung wird die gesamte Hierarchierelation absteigend (in Richtung der Unterbegriffe) verwendet, um eine Abbildung zu finden. Gelingt auch das nicht, wird vom ursprünglichen Deskriptor über Assoziationsbeziehungen versucht eine Abbildung zu finden. Konnte noch immer nicht ein Pfad zu einem Deskriptor aus einem Indexierungsthesaurus gefunden werden, kann keine Abbildung stattfinden.

Hohe Precision und absteigend keine Abbildung gefunden: Wenn der Benutzer eine hohen Precision verlangt und über die Äquivalenrelation, Benutze-Kombination-Relation und Abstraktionsoberbegriffe keine Abbildung gefunden werden konnte, wird zusätzlich absteigend die gesamte Hierarchierelation verwendet. Führt auch das zu nicht mindestens einer erfolgreichen Abbildung, wird aufsteigend erst ausschließlich die Abstraktionsrelation dann die Hierarchierelation verwendet, um eine Abbildung zu finden. Kann wiederum keine Abbildung gefunden werden, wird über Assoziationsbeziehungen, die dierekt vom ursptünglichen Deskriptor ausgehen, versucht eine Abbildung zu finden. Ist auch das nicht erfolgreich, kann keine Abbildung stattfinden.

Pfade, die erst über Abstraktions- bzw. Bestandsbeziehungen und dann über Assoziationsbeziehungen gehen, werden nicht verfolgt, da die Bedeutungsänderung, die der Begriff bei einer

solchen Abbildung erfahren würde, als nicht mehr akzeptabel hingenommen wird.

13.3.2.3.2 Thesauruspriorisierung und Konfliktbehandlung Bei der Bestimmung der Ober-/-unter- bzw. Assoziationsbegriffe stellt sich die Frage, welche Thesauri zu berücksichtigen sind. Diese Frage kann einfach beantwortet werden: Da die Indexierungsthesauri die Zielthesauri sind, müssen alle Indexierungsthesauri berücksichtigt werden. Über unterschiedliche benutzerrelevante Thesauri können ggf. unterschiedliche Begriffe in potenziell verschiedenen Indexierungsthesauri erreicht werden. Daher müssen auch alle benutzerrelevante Thesauri berücksichtigt werden.

Somit muss angegeben werden wie die Indexierungsthesauri im Verhältnis zu den benutzerrelevanten Thesauri zu priorisieren seien und welche Konfliktauflösungspolitik verfolgt wird. Da bei Abbildungsanfragen die Zielsetzung eine andere ist (Abbildung zu möglichst nahem Deskriptor aus einem Indexierungsthesaurus) als bei den Navigationsanfragen innerhalb der benutzerrelevanten Thesauri (liefern eines konsistenten Bildes unter Berücksichtigung der Benutzerpräferenzen), spiegelt sich das in einer anderen Konfliktauflösungspolitik wider:

- Abstraktionsredundanz sollen weiterhin aufgelöst werden, damit das durch Redundanzen mehrfache Besuchen eines Deskriptor vermieden wird. Um die Abbruchbedingung der Rekursion zu garantieren, müssen Zyklen ebenfalls aufgelöst werden.
- Neben gemischten Abstraktions- und Bestandsunterbegriffen werden auch Beziehungstypdifferenzen, Schwesternassoziationen und Abstraktionsniveaudifferenzen toleriert und erfordern keine Konfliktauflösung. Dies ist möglich, da jeweils nur ein Beziehungstyp berücksichtigt wird.

Bei dieser Konfliktbehandlungspolitik hat die Priorisierung der Thesauri ausschließlich auf die Zyklenauflösung einen Einfluss. Um möglichst keinen Pfad zu verhindern, der von einem benutzerrelevanten Thesaurus in einen Indexierungsthesaurus führt, werden alle Indexierungsthesauri als geringer als alle benutzerrelevanten Thesauri eingeordnet. Indexierungsthesauri die zugleich benutzerrelevante Thesauri sind behalten die Priorisierung, die sie als benutzerrelevanter Thesaurus besitzen. Als Ordnung der Indexierungsthesauri wird die Ordnung der entsprechenden Informationssysteme verwendet.

13.3.2.3.3 Beispiele

Beispiel 13.5 *Ein konkretes Beispiel soll die Auswirkungen der Precision- bzw. Recall-Präferierung des Benutzers zeigen. Gegeben sei der in Abbildung 13.8 dargestellte Ausschnitt einer Thesaurusföderation mit dem benutzerrelevanten Thesaurus GEMET und dem Indexierungsthesaurus AGROVOC. Für den Deskriptor* G.climate *soll eine Abbildung in den Indexierungsthesaurus gefunden werden. Äquivalenz-, Benutze-Kombinations- und Oberbegriffsbeziehungen existieren für diesen Deskriptor nicht. Wir betrachten sowohl den Fall der Precision- als auch der Recall-Präferierung durch den Benutzer:*

Precision-Präferierung: *Erster Schritt ist die Abbildung auf* G.climate type *und anschließend die disjunktive Verknüpfung der Deskriptoren* G.arid climate, G.temperate climate, A.temperate climate *und* A.desert climate *anhand der Abstraktionsuntebegriffsbeziehungen. Im nächsten Schritt kann der Ausdruck aufgrund der Inter-Thesaurus-Äquivalenzbeziehungen vereinfacht werden zu* A.temperate climate $\vee$ A.desert climate. *Die entsprechenden Begriffsinhalte sind umfangreicher als die Begriffsinhalte des urspünglichen Begriffs* G.climate*, daher kann ein guter Precision-Wert erwartet werden.*

Recall-Präferierung: *Da keine Abstraktionsoberbegriffe existieren, muss auch bei einer Recall-Präferierung mit Unterbegriffen versucht werden, den Deskriptor* G.climate *in den Indexierungsthesaurus abzubilden. Als erster Ausdruck wird* G.climate type $\vee$ A.climatic change $\vee$ A.meteorological phenomens *erzeugt. Da ein guter Recall-Wert präferiert wird, wird bei Betrachtung der Unterbegriffe auch die Bestandsrelation berücksichtigt (laut Konfliktbehandlungspolitik werden gemischte Abstraktions- und Bestandsbeziehungen toleriert). Im weiteren wird der Teilausdruck* G.climate type *wie im vorangegangenen Fall beschrieben abgebildet, so das als Endergebnis der Ausdruck* A.temperate climate $\vee$ A.desert climate $\vee$ A.climatic change $\vee$ A.meteorological phenomens *erzeugt wird. Die durch die Bestandsrelation gefundene Teilabbildung hat die Bedeutung des urprünglichen Begriffes zwar verändert, jedoch so, dass zwar eine schlechtere Precision aber ein besserer Recall erwartet werden können, da zusätzlich semantisch nahe Begriffe berücksichtigt werden.*

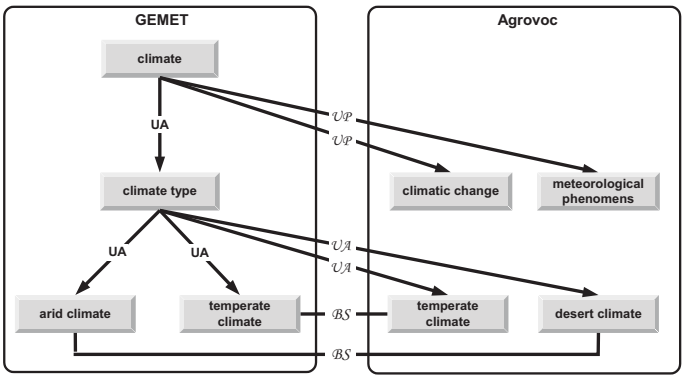

Abbildung 13.8: Vereinfachter Ausschnitt der Thesaurusföderation bestehend aus GEMET und AGROVOC

Beispiel 13.6 *Gegeben sei die in Abbildung 13.9 dargestellte Situation: Der Abstraktionspfad zwischen den Begriffen x_1, x_2, x_3 aus dem höchstpriorisierten benutzerrelevanten Thesaurus BR_1 und y_1, y_2 aus dem niedriger priorisierten benutzerrelevanten Thesaurus BR_2 ergibt einen Zyklus. Als Standardkonfliktauflösung ist die Kante zwischen x_2 und x_3 als zu entfernen markiert. Aufgrund der Priorisierung wird jedoch bei Navigationsanfragen die Kante zwischen y_1 und y_2 entfernt. Eine Abbildungsanfrage erfordert nun die Abbildung von y_1 auf Deskriptoren in die Indexierungsthesauri IT_1 bzw. IT_2. Äquivalenz- oder Benutze-Kombination-Beziehungen in einen Indexierungsthesaurus existieren nicht. Da der Benutzer eine gute Precision bevorzugt, wird versucht, über* getNarrower*-Anfragen Abstraktionsunterbegriffe zu finden. Aufgrund der Konfliktauflösung werden jedoch keine Unterbegriffe geliefert. Als direkter Oberbegriff wird x_3 geliefert. Da wiederum keine Äquivalenz- oder Benutze-Kombination-Beziehung in einen Indexierungsthesaurus exitistiert, werden als direkte Oberbegriffe von x_3 x_2 und v konjunktiv verknüpft geliefert. v stammt aus einem Indexierungsthesaurus und braucht daher nicht mehr abgebildet werden. x_2 kann durch die Inter-Thesaurus-Äquivalenzrelation auf v abgebildet werden, der entstehende Ausdruck $v \wedge v$ wird zu v vereinfacht und als Ergebnis zurückgeliefert.*

13.3.3 Anfragereformulierung und -erweiterung

Der Benutzer des Föderierten Systems fordert insbesondere in kritischen Situationen, d.h. es wurden zu wenige oder zu viele oder zu viele nicht-relevante Dokumente gefunden, Möglichkeiten zur Anfragereformulierung und -erweiterung, vgl. z.B. [Fid91, Kri93, PE94, PSBT96,

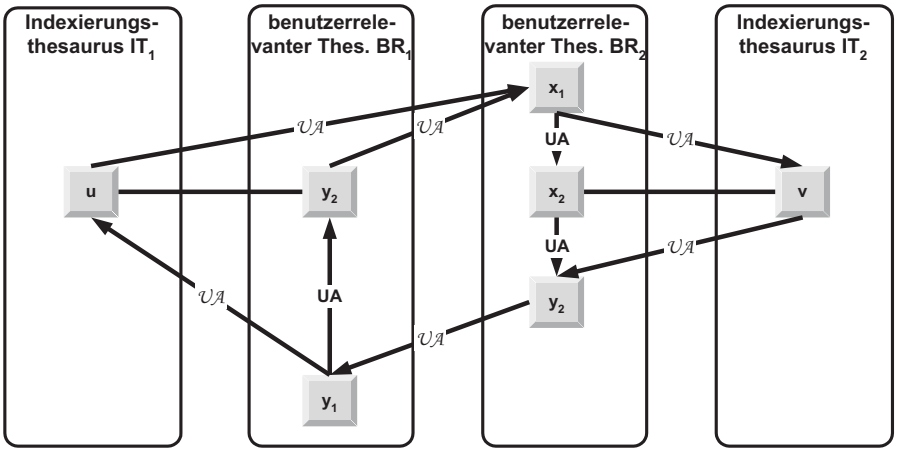

Abbildung 13.9: Abstraktionszyklen zwischen Deskriptoren in benutzerrelevanten Thesauri und Indexierungsthesauri

Eft94, Yon95, BMS98]. Das Ziel einer Anfragereformulierung oder -erweiterung besteht in der Verbesserung der Retrieval-Performance.

Bei Verwendungung einer Thesaurusföderation kann dem Benutzer durch die Anfragereformulierung die Möglichkeit gegeben werden, die Genauigkeit der Abbildung seiner Deskriptoren aus den benutzerrelevanten Thesauri auf die Indexierungsthesauri zu verbessern. Dazu kann er einzelne Deskriptoren aus der Anfrage entfernen oder durch Suchen in den Indexierungsthesauri neue hinzufügen. Somit ist der Benutzer nicht ausschließlich auf die Genauigkeit der automatischen Abbildung angewiesen.

Der Benutzer kann manuell oder das System automatisch oder der Benutzer mit Unterstützung des Systems oder das System mit Unterstützung des Benutzers versuchen, die ursprüngliche Anfrage anzupassen, um das Retrieval-Ergebnis zu verbessern. Prinzipiell unterschieden wird zwischen Anfrageerweiterung basierend auf den Suchergebnissen und Anfrageerweiterung basierend auf Wissensstrukturen wie Thesauri [Eft96]. Wir betrachten hier nicht den ersten Fall, der Bestandteil des Information-Retrieval-Verfahren ist und in keinem direkten Zusammenhang mit unserer Arbeit steht (vgl. hierzu etwa [Rol01]). Für den zweiten Fall werden die Möglichkeiten bei Verwendung einer Thesaurusföderation aufgezeigt.

Automatische Anfrageerweiterung mit Thesaurusföderationen: Bei der automatischen Anfrageerweiterung werden entlang der Beziehungen der Thesaurusföderation Deskriptoren aufgesucht und in die Anfrage mit einbezogen. Ein Beispiel ist die Berücksichtigung aller Hierarchieunterbegriffe, solange diese nicht mehr als n Ebenen von dem ursprünglichen Begriff entfernt sind. Grundlage der automatischen Anfrageerweiterung sind die Indexierungsthesauri, da Präferenzen des Benutzers für bestimmte Thesauri nicht berücksichtigt werden müssen. Vorteil ist die Nicht-Erforderlichkeit menschlichen Eingriffs, Nachteil die Gefahr des Anfrageabtriebs (engl. query drift), d.h. die erweiterte Anfrage drückt nicht mehr den ursprünglichen Informationsbedarf aus.

Manuelle Anfrageerweiterung mit Thesaurusföderationen: Auch wenn eine rein manuelle Anfrageerweiterung stattfinden soll, können Deskriptoren vom Benutzer nur aus den Indexierungsthesauri ausgewählt werden. Vorteil ist, dass der Benutzer durch Suche und Navigation in den Indexierungsthesauri gezielt seinen Informationsbedarf besser ausdrücken kann. Nachteil ist jedoch, dass er sich dazu in die Strukturen ihm unbekannter

Thesauri einarbeiten muss.

Semi-automatische Anfrageerweiterung mit Thesaurusföderationen: Eine Anfragerweiterung durch den Benutzer mit Unterstützung des Systems kann eine zielgerichtete Anfrageerweiterung bei Verwendung des für den Benutzer vertrauten Vokabulars ermöglichen. Dazu wählt der Benutzer aus den benutzerrelevanten Thesauri Deskriptoren aus, die vom System auf die Indexierungsthesauri abgebildet werden. Den genannten Vorteilen steht als Nachteil gegenüber, dass nicht offensichtlich ist, welche neuen Deskriptoren aus den benutzerrelevanten Thesauri zu neuen Deskriptoren aus den Indexierungsthesauri führen. Eine weitere Möglichkeit besteht daher in der gleichzeitigen Verwendung der benutzerrelevanten Thesauri und der Indexierungsthesauri. Nachteil ist die Mitverwendung des fremden Vokabulars, Vorteile sind jedoch, dass auch das bekannte Vokabular verwendet wird und bereits während des Prozesses der Anfrageerweiterung ersichtlich gemacht werden kann, welche neuen Deskriptoren aus Indexierungsthesauri für die Anfrageerweiterung verwendet werden können.

Welche dieser Alternativen verwendet wird, kann der Benutzer abhängig von den Retrieval-Zwischenergebnissen und den erzielten Verbesserungen entscheiden. Es wird jedoch offensichtlich, dass die Schnittstellen des Thesaurusföderationsmediators (vgl. Abschnitt 13.3.1) ausreichen, um alle aufgeführten Arten der Anfrageerweiterung zu unterstützen.

13.4 Kapseln

Die Kapseln verbergen die Heterogenität der Komponententhesauri und machen diese über eine einheitliche Diensteschnittstelle verfügbar, auf die entfernt zugegriffen werden kann. Das bedeutet, dass alle Anfragen des Mediators an die Komponententhesauri Bestandteil der Diensteschnittstelle der Kapseln sein müssen.

Der Thesaurusföderationsmediator reicht ausschließlich Anfragen an die Komponententhesauri weiter, wenn Thesaurusföderationsmediator-Anfragedienste aufgerufen werden. Die Dienste der Kapseln können daher vollständig aus diesen Diensten des Mediators hergeleitet werden: Es handelt sich bis auf *getGroup* und *map* um alle Anfragedienste des Meditors. Als Parameter werden statt Föderationsdeskriptoren Komponententhesaurus-Deskriptoren übergeben, auf die Angabe des Tehsaurus kann verzichtet werden, da dieser durch die aufgerufene Kapselinstanz eindeutig ist.

Die Funktionen werden üblicherweise bereits von den Komponententhesaurussystemen unterstützt, da sie für allgemeine Thesaurusbrowser-Funktionalität erforderlich sind (vgl. etwa [NKS+99, PSBT96]). Eine Abbildung der allgemeinen Schnittstelle ist daher verhältnismäßig einfach. Wir betrachten die beiden häufigsten Fälle ausführlicher: Ein Thesaurus verfügt bereits über Anfrageschnittstellen bzw. er steht als HTML-Thesaurus im Web zur Verfügung und soll eingebunden werden.

13.4.1 Kapseln für Thesauri mit Anfrageschnittstellen

Verfügt der Thesaurus bereits über eine Anfrageschnittstelle (Programmierschnittstelle) ist abhängig davon, ob diese entfernt aufgerufen werden kann, eine Installation der Kapsel am Ort des Thesaurus erforderlich (Kapsel stellt entfernte Aufrufbarkeit sicher) oder nicht (Anfrageschnittstelle ist entfernt aufrufbar).

Die Anfrageschnittstelle des Thesaurus wird im Allgemeinen eine sehr ähnliche Struktur zur Kapselschnittstelle besitzen. Selbst wenn dies nicht der Fall ist, weil die Anfrageschnittstelle z.B. direkt über SQL abgefragt werden muss, ist es für einen Entwickler der mit der nativen Schnittstelle des Komponententhesaurus sowie dem Austauschformat XML vertraut ist, aufgrund der ähnlichen Grundstruktur von Thesauri sowie der vertrauten Anfragen einfach, die Kapseln zu implementieren. Der Aufwand wird, wie wir anhand der Entwicklung von Kapseln für GEMET und Valids nachgewiesen haben, nur wenige Stunden betragen – verglichen mit der eigentlichen Begriffsintegration ist das vernachlässigbar.

13.4.2 Kapseln für Thesauri mit HTML/HTTP-Anfrageschnittstellen

HTML-Kapseln werden immer am Ort des Mediators installiert. Ihre Aufgabe ist es, Anfragen über das Web-Protokoll http in Form von parametrisierten URL-Aufrufen an den HTML-Thesaurus (ggf. über einen Proxy) zu senden, die eine statische oder eine dynamische HTML-Seite als Antwort liefern. Aus dieser HTML-Seite sind dann durch Parsen die gewünschten Informationen zu generieren. Unterschiede zu den oben aufgeführten Anfrageschnittstellen betreffen die Zustandslosigkeit des http-Protokolles, das zudem keinen verlässlichen Kommunikationskanal bietet, d.h. die Häufigkeit von Verbindungsfehlern ist signifikant höher. Auch Fehlermeldungen werden als HTML-Seiten geliefert und müssen von den eigentlichen Nutzdaten unterschieden werden. Die hohe Dynamik im Web bedeutet häufige Redesigns, die Änderungen sowohl in der URL-Generierung als auch im HTML-Parser nach sich ziehen.

Aus diesen Gründen sowie einer beinahe beliebigen Struktur, die HTML-Seiten besitzen können, ist das Extrahieren von Informationen aus dynamischen Web-Seiten keine einfache Aufgabe. Vorteil bei der Konstruktion von Kapseln zur Exktraktion von Informationen aus HTML-Thesauri ist jedoch, dass die meisten HTML-Thesauri ähnliche Strukturen besitzen, womit ähnliche Parser eingesetzt werden können. Zudem liegen inzwischen eine Reihe von Forschungsergebnissen zum semi-automatischen Entwurf von Parsern anhand der Identifikation von spezifischen Strukturen vor (vgl. etwa [AK97, KWD97, Coh99, DPA00]), die bereits in erste Produkte umgesetzt sind (z.B. equero [equ01]). Mit der Verfügbarkeit dieser Produkte kann somit eine Kapsel für einen HTML-Thesaurus auch unter den oben genannten Bedingungen mit wenig Aufwand realisiert werden.

Steht bereits ein Rahmen zur Verfügung (XML-Parser, HTML-Parser, URL-Generator) kann auch ohne Werkzeugunterstützung innerhalb weniger Stunden bis Tage eine Kapsel für HTML-Thesauri entwickelt werden. Wir haben das durch die Entwicklung einer Kapsel für den Astronomie-Thesaurus der Internationalen Astronomischen Vereinigung `http://msowww.anu.edu.au/library/thesaurus/` gezeigt, die im Verlaufe der Diplomarbeit [Tra97] implementiert wurde.

13.5 Facilitator und Informationssystemmediator

Benutzer und Systeme, die mit dem Föderierten System interagieren, benötigen, um die Komplexität der Kommunikation zu reduzieren, *einen* Zugangspunkt (engl. single point of access). Diese Aufgabe erfüllt in unserer Architektur der Facilitator.

Die Funktionalität Föderierter Informationssysteme ist, um den spezifischen Anforderungen der Anwender und der jeweiligen Anwendungsdomäne gerecht zu werden, sehr unterschiedlich. Ohne diese Funktionalität näher zu analysieren, können die Teilaufgaben des Facilitators auf abstraker Ebene definiert werden:

Feststellung relevanter Mediatoren: Anhand einer Analyse der Anfrage muss festgestellt werden, welche Mediatoren (aus Sicht des Faciltators sind das die Informationsquellen) die zur Beantwortung der Anfrage erforderlichen Informationen liefern.

Entwurf einer Abarbeitungsstrategie: Ein Anfrage an den Facilitator kann sowohl das Absetzen nur einer Anfrage an einen Mediator bedeuten (Beispiel: Liefern von Abstraktionsunterbgriffen für die Darstellung in einem Thesaurusföderationsbrowser) als auch eine Sequenz von Abfragen an verschiedene Mediatoren und das Zusammenfügen der Teilergebnisse erfordern (Beispiel: boolsche Anfrage mit mehreren Deskriptoren aus benutzerrelevanten Thesauri erfordert mehrere Zugriffe auf den Thesaurusföderationsmediator, um die Deskriptoren in Deskriptoren aus Indexierungsthesauri zu übersetzen und anschließend das Absetzen der so modifizierten Anfrage an den Informationssystemmediator).

Transformation der Ergebnisse: Schließlich können Facilitatoren zusätzlich die Aufgabe übernehmen, die im allgemeinen Format vorliegenden Ergebnisse in ein benutzerspezifisches Format zu transformieren. Dieser Aspekt soll hier aber als nicht weiter betrachtet werden, da dies der Funktionalität einer Kapsel entspricht.

Zur Feststellung relevanter Mediatoren und zum Entwurf einer Abarbeitungsstrategie reichen in unserem Szenario einfache Regeln: Alle Anfragen über und an Thesauri werden direkt an den Thesaurusföderationsmediator weitergeleitet. Dazu erhält der Facilitator eine Schnittstelle, die der des Thesaurusföderationsmediators entspricht. Suchanfragen werden daraufhin analysiert, ob sie Deskriptoren aus Nicht-Indexierungsthesauri enthalten[6]. Ist dies der Fall, werden diese jeweils durch Anfragen an den Thesaurusföderationsmediator ersetzt, bevor sie an den Informationssystemmediator weitergeleitet werden. Anfragen, die die Reformulierung und Erweiterung von Anfragen betreffen, müssen je nach Reformulierungs- und Erweiterungsstrategie individuell behandelt werden. Etwa kann eine Anfrageerweiterung erfordern, die Unterbegriffe der aktuellen Deskriptoren, die durch einen Pfad der maximalen Länge drei erreicht werden können zur Anfrage hinzuzunehmen (Sequenz von Anfragen an Thesaurusföderationsmediator erforderlich). Oder aber es sollen von Ergebnissen, die der Benutzer als relevant markierte, Thesaurusdeskriptoren für weitere Anfragen übernommen werden (keine Anfragen an Thesaurusföderationsmediator erforderlich).

Weitere Mediatoren sind möglich. In unserem Szeanria aus dem Bereich der Umwelt/Landwirtschaft etwa Geo-Mediatoren, die entsprechend der Abbildung von Thesaurusbegriffen die Abbildung von räumlichen Begriffen und Koordinaten vornehmen.

Für unterschiedliche Föderierte Informationssysteme, die in der Funktionalität und Mächtigkeit der Anfrageformulierungen, -reformulierungen und -erweiterungen, der Information Retrieval-Verfahren, der Benuzeranalysen, der Ergebnispräsentation und des Benutzerfeedbacks sehr verschieden sein können, ist sowohl der Informationssystemmediator als auch der Facilitator entsprechend anzupassen. Gestützt werden kann sich dabei sowohl auf den hier angerissenen Rahmen als auch eine Reihe weiterer Forschungsergebnissen, die das Erzeugen und Pflegen von Facilitatoren und Mediatoren betreffen (z.B. [GKD97, Rie99, KAM+01])

[6] Um nicht versehentlich Freitext-Eingaben des Benutzers, die bei einer Volltextsuche genutzt werden sollen, zu übersetzen, ist die Markierung von Thesaurus-Deskriptoren für eine semantische Suche innerhalb der Anfrage erforderlich.

13.6 Resümee

Wir haben in diesem Kapitel gezeigt, dass Thesaurusföderationen, die wie in den vorangegangenen Kapiteln beschrieben konstruiert wurden, durch eine mehrschichtige Architektur in die Anfrageformulierung, -bearbeitung und -erweiterung in integrierten Informationssysteme eingebracht werden können. Durch Berücksichtigung des Kontextes sowie der Konfliktmarkierungen wird eine an die Präferenzen des Benutzers angepasste übergreifende Sicht auf das Vokabular der Komponententhesauri bereitgestellt. Durch die Modularität der Architektur sowie die Verwendung verbereiteter Internet-Standards ist die einfache Erweiterbarkeit sowohl hinsichtlich weiterer Informationsquellen als auch hinsichtlich der Funktionalität gewährleistet.

Kapitel 14

Zusammenfassung und Ausblick

In diesem Kapitel werden die wesentlichen wissenschaftlichen Beiträge dieser Arbeit zusammengefasst sowie ein Ausblick auf weiterführende Arbeiten gegeben.

14.1 Zusammenfassung

14.1.1 Ausgangssituation

Der Wandel von einer industriellen Gesellschaft zu einer post-industriellen „Informationsgesellschaft“ geht mit einem zunehmendem Bedarf nach einfachem Zugriff auf heterogene, global verteilt vorgehaltene Informationen einher. Mit offenen Computernetzen wie dem World Wide Web [Con02] steht die erforderliche technische Infrastruktur zur Verfügung. Von besonderer Bedeutung in solchen großen „verteilten Informationsumgebungen“ sind Dienste zum gezielten Wiederauffinden von Informationen (Information Retrieval). Die heute verwendeten Suchmaschinen wie AltaVista [Alt02] oder Google [Goo02] ermöglichen eine einfache Volltextsuche mit unterschiedlichen Verfahren zur Ordnung der Ergebnisse (Ranking). In klassischen „Fachinformationssystemen“, z.B. den vom FIZ Karlsruhe [FIZ00] angebotenen Informationsdiensten, dem FORIS Forschungsinformationssystem Sozialwissenschaften [GES01] oder AGRIS als Agrar-Informationssystem [ZRSJ+92] oder auch Bibliotheks-/Museenkatalogen wie dem Getty-Katalog [The01]) hingegen geht die Unterstützung des Benutzers deutlich darüber hinaus. Insbesondere Thesauri mir ihrem einheitlichen und konsistenten Vokabular sind ein bewährtes Werkzeug. Thesauri ermöglichen über die reine Volltextsuche hinausgehend semantisches Suchen.

Jedoch sind traditionelle Fachthesauri – bereits 1990 wurden von der Europäischen Union über 1.000 häufig verwendeten Thesauri weltweit identifiziert [Rad90] – für den Zugriff von sehr unterschiedlichen Benutzergruppen auf fachübergreifende Datenbestände nicht mehr ausreichend. Ein einheitlicher Zugang zu verschiedenen Informationssystemen erfordert eine Integration dieser mit jeweils großem Aufwand erstellten entsprechenden Thesauri, die eine gemeinsame Benutzung der Terminologie ermöglicht.

Existierende Ansätze betrachteten bisher nur isolierte Aspekte einer Integration von Thesauri. Ein ganzheitliches Rahmenwerk, das von den Informationsmodellen für Thesauri und Thesaurusföderationen über ein Vorgehensmodell und die Architektur sowie Verfahren zur Gewinnung des Integrationswissens bis hin zur Bewertung und Anwendung der integrierten Thesauri reicht, war weder in einzelnen Arbeiten noch in der Gesamtheit der Arbeiten vorhanden.

14.1.2 Lösungsansatz

Das Ziel dieser Arbeit war es, ein umfassendes Rahmenwerk für die flexible Integration von heterogenen, autonomen Thesauri zu schaffen. Dabei sollte insbesondere die Autonomie der beteiligten Thesauri erhalten bleiben und mit ggf. aus dieser Autonomie enstehenden Konflikten umgegangen werden können. Schwerpunkte waren einerseits die Skalierbarkeit, die eine wesentliche Erleichterung des Prozesses der Thesaurusintegration aber auch die entsprechende Unterstützung des Benutzers beim Umgang mit dem integrierten System fordert, und andererseits die Flexibilität, die das Einbinden heterogener Thesauri sowie die an den Benutzer angepasste Verwendung des integrierten Systems bedeutet.

Ausgangsbasis aller weiteren Arbeiten waren formale Informationsmodelle für Thesauri und Thesaurusföderationen. Das Informationsmodell für Thesauri war erforderlich, da in der Praxis verwendete Thesaurusmodelle z.T. erheblich voneinander abweichen. Das Modell entstand durch Betrachtung verschiedener Thesauri und ist offen für eine Vielzahl von Thesauri. Die mathematische Präzision des mengentheoretischen Modells für Thesauri ermöglicht eine Integration unter Berücksichtigung einer vergleichbaren Syntax und Semantik. Durch Repräsentation des Modells als Graph konnte zudem eine gute Anschaulichkeit gewonnen werden. Wir haben darüberhinaus weitgehend automatisierte Verfahren für die Identifikation und Behandlung von Abweichungen entwickelt, die eine Transformation zuvor nicht konformer Thesauri in unser Informationsmodell ermöglichen.

Das Informationsmodell für Komponententhesauri wiederum ist Ausgangsbasis für das Informationsmodell der von uns entwickelten Thesaurusföderation. Thesaurusföderationen tragen der Autonomie der beteiligten Komponententhesauri Rechnung und unterscheiden sich daher grundlegend von bereits bekannten Multi-Thesaurus-Systemen: Beziehungen zwischen Begriffen verschiedener Thesauri werden durch Inter-Thesaurus-Relationen repräsentiert, deren Semantik sich an die der Intra-Thesaurus-Relationen anlehnt. Konflikte, die nicht aufgelöst werden können, werden zugelassen, wenn sie gegen Invarianten festgestellt werden, die eine Konfliktmarkierung erlauben. Das ermöglicht erstmalig das situationsabhängige Auflösen der Konflikte im Moment der Benutzeranfrage. Das Informationsmodell für Thesaurusföderationen sieht nicht nur explizite Beziehungen vor, sondern auch implizierte Beziehungen, also solche, die aus vorhandenen Beziehungen und den Relationseigenschaften abgeleitet werden können. Die explizite Aufführung implizierter Beziehungen verhindert, dass diese bei der Begriffsintegration, dem Auffinden von Konflikten oder auch dem Anwenden der Föderation wiederholt berechnet werden müssten.

Als komplexe Aufgabe hat sich die eigentliche Begriffsintegration, also das Auffinden von Inter-Thesaurus-Beziehungen einschließlich dem Auffinden von Konflikten und deren mögliche Bereinigung, erwiesen. In der Literatur werden bereits eine Reihe möglicher Integrationsverfahren beschrieben. Jedoch existiert für das Problem der Begriffsintegration kein geschlossener Lösungsansatz, sondern es gilt, die flexible und iterative Lösungssuche zu unterstützen. Flexibilität bedeutet an dieser Stelle, existierende sowie neu entwickelte und zukünftige Lösungsverfahren einbringen sowie unterschiedliche Problemlösungsstrategien verfolgen zu können. Wir haben daher eine auf dem Blackboard-Modell basierende Architektur mit Ansätzen aus dem Workflow-Management-System-Bereich vereint. Diese neue Architektur erlaubt es, Lösungsverfahren und Lösungsstrategien strikt voneinander getrennt zu betrachten. So kann die Lösungsstrategie jederzeit – auch während der Abarbeitung einer Strategie – geändert werden und weiterentwickelte oder neue Lösungsverfahren können einfach eingebracht werden. Die Expertise des menschlichen Experten kann dabei sowohl in die Lösungsstrategie als auch zur Unterstützung der Problemlösungsverfahren eingebracht werden. Die Lösungsstrategie wird als

Workflow (Aufgaben-Agenda) modelliert. Die Lösungsverfahren hingegen werden von maschinellen Experten implementiert, die ihre Ergebnisse über das Blackboard in Form von Hypothesen und Hypothesenbewertungen austauschen. Hypothesenbewertungen erlauben es den Experten, Vertrauen in einzelne Hypothesen auszudrücken. Expertenbewertungen wiederum drücken das Vertrauen in Experten aus und ermöglichen die angemessene Berücksichtigung sowohl von Verfahren, die qualitativ sehr hochwertige Ergebnisse liefern, als auch solcher, die weniger hochwertige Ergebnisse liefern, jedoch z.B. interessante neue Vorschläge generieren. Ein Bewertungsmodell ermöglicht es, zu aggregierten Gesamtbewertungen zu gelangen.

Mit dieser Wissensakquisitionsarchitektur haben wir also die Basis für eine Begriffsintegration geschaffen, die sich durch ihre Flexibilität von den starren zuvor existierenden Ansätzen unterscheidet.

Um die bestmögliche Integration der Thesauri zu erreichen, ist eine eingehende Analyse dieser Thesauri erforderlich. Dieser Aspekt ist in bekannten Ansätzen bisher beinahe vollkommen vernachlässigt worden. Wir haben daher ein Kennzahlensystem entwickelt, das es erlaubt, durch Interpretation dieser Kennzahlen und einiger weniger stichprobenartiger Untersuchungen mit geringfügigem Aufwand des menschlichen Integrationsexperten wesentliche inhaltliche Eigenschaften von Thesauri zu erkennen. Wie wir gezeigt haben, können aus diesen Ergebnisse wichtige Folgerungen sowohl für Verfahren der Begriffsintegration als auch hinsichtlich der Erwartungen an das Ergebnis gewonnen werden.

Für die Begriffsintegration haben wir eine Lösungsstrategie entwickelt, die erstmalig explizit als Prozess dargestellt wird und unabhängig von den Lösungsverfahren ist. Dieser Prozess sieht als Phasen ähnlich einem Wasserfallmodell die initiale Integration, die Optimierung anhand der Zwischenergebnisse und schließlich die bewertungsbasierte Integration vor. Somit wird auch innerhalb dieses Prozesses die iterative Lösungsfindung unterstützt.

Die Lösungsstrategie wird innerhalb der Realisierungsphase umgesetzt. Die aus der Strategie resultierende Prozessbeschreibung in Form der Aufgaben-Agenda stellt die Basis für die Prozessablaufsteuerung durch den mit einem Workflow-Ausführungsdienst vergleichbaren Ausführungsagenten dar. Das Problemlösungswissen wird durch die Problemlösungsverfahren in Form von (maschinellen) Experten eingebracht. Wir haben diese Verfahren auf Basis bekannter Verfahren weiterentwickelt und Verfahrensvorschläge für Bereiche, die in der Literatur bisher kaum beachtet wurden, vorgestellt. Mit diesen Verfahren wird ein umfangreicher Grundstock zur Gewinnung des Integrationswissens bereit gestellt. Eine Bewertung der Verfahren (Voraussetzungen, Aufwand, Interaktionen mit dem menschlichen Experten) ermöglicht eine Einordnung dieser und potenziell weiterer Verfahren in die entsprechenden Phasen der Lösungsstrategie. Eine Anpassung der Verfahren an eine konkrete Situation kann durch Konfigurieren der Konfidenzfaktoren für die zugrundeliegenden Regeln oder Algorithmen geschehen.

Um anhand der von den Lösungsverfahren erzeugten Hypothesen Fakten zu erhalten, ist – wie wir gezeigt haben – neben der Berechnung einer Gesamtbewertung eine qualitative Überprüfung der Hypothesen hinsichtlich Widerspruchsfreiheit gegen die Modelleigenschaften erforderlich. Eine vollständige qualitative Überprüfung hat dabei unterschiedliche Ausschnitte der Hypothesenmenge (einzelne Hypothese, alle Hypothesen, Fakten) und des bereits vorhandenen Integrationswissens (persistente Fakten, implizierte Beziehungen) zu betrachten. Nur wenn dies der Fall ist, kann auch bei Einbezug von Verfahren, die nicht zwingend alle Modelleigenschaften kennen und überprüfen und die von den verschiedenen Verfahren erzeugten Hypothesen nicht auf eine Gesamtkonsistenz abstimmen können, gewährleistet werden, dass die Föderation – ggf. unter Zuhilfenahme von Konfliktmarkierungen – konsistent bleibt.

Werden schließlich Hypothesen zu Fakten, wurden Algorithmen vorgestellt, um diese Fakten mit

sämtlichen Implikationen festzuhalten (z.B. ist beim Entfernen einer Inter-Thesaurus-Beziehung das Entfernen aller implizierten Kanten sowie aller Konfliktmarkierungen, die ohne diese Kante nicht länger erforderlich sind, notwendig).

Basis der bewertungsbasierten Integration ist die Bewertung eines Zwischenstandes der Begriffsintegration hinsichtlich des Grades der Erfüllungen der bei der Analyse der Thesauri formulierten Erwartungen. Um zu dieser Bewertung zu gelangen, haben wir die Kennzahlen zur Analyse von Thesauri auf die Analyse und Bewertung von Thesaurusföderationen übertragen. Damit wurde sowohl die Basis für die bewertungsbasierte Optimierung als auch erstmals ein umfangreiches Kennzahlensystem für die Bewertung und den Vergleich von Multi-Thesaurus-Systemen geschaffen.

Wurde das Integrationswissen akquiriert, wird schließlich eine Laufzeitumgebung benötigt, um die Thesaurusföderation in übergreifenden Informationssystemen zum gezielten Wiederauffinden von Informationen verwenden zu können. Inbesondere müssen bei Anfragen an die Föderation markierte Konflikte aufgelöst werden. Um diese Konfliktauflösung situationsgerecht und benutzerpräferenzgesteuert durchführen zu können, wird der Kontext (nach Bedeutung geordnete Menge der Informationssysteme mit den entsprechenden Indexierungsthesauri, geordnete Menge relevanter Thesauri, Recall- bzw. Precision-Präferenz) berücksichtigt. Wir können somit erstmals eine benutzerspezifische Sicht auf das Vokabular eines Multi-Thesaurus-Systems geben und unterstützen dadurch auch heterogene Benutzergruppen mit unterschiedlichen Vorkenntnissen und Vorlieben.

Mit den zentralen vier Bausteinen unserer Lösung (Informationsmodelle, Wissensakquisitionsarchitektur, Begriffsintegration (inkl. aller Phasen des Vorgehensmodells), Ausführungsmaschine) haben wir die Voraussetzungen geschaffen, um Thesauri semi-automatisch unter weitmöglichster Verwendung der Thesaurus-inhärenten Informationen zu einer Thesaurusföderation zu integrieren und diese Thesaurusföderation benutzer- und aufgabengerecht für das gezielte Wiederauffinden von Informationen verwendbar zu machen. Wir haben somit unser Ziel der Verbesserung der Skalierbarkeit und Flexibilität bei der Erstellung und Anwendung eines integrierten Thesaurus-Systems erreicht.

14.1.3 Realisierung des Ansatzes

Für die Validierung wurden weite Teile des Ansatzes realisiert (vgl. auch die entsprechenden Studienarbeiten und Diplomarbeiten [Hab96, Mor98, Sch99b, Vas99, Fis00, Geb00, Gut00, Str01, Sun01, Wei01]): Die Informationsmodelle wurden als ER-Modelle modelliert und in relationale Datenbankschemata umgesetzt. Zur Überprüfung der Invarianten wurden eigene Programm-Module entwickelt. Die Thesauri AGROVOC, GEMET und GCMD Valids wurden mittels ETL-Algorithmen (Extraction, Transformation, Loading) und den implementierten Verfahren der Vorbereitungsphase in die Informationsmodelle transformiert. Zur Analyse der Komponententhesauri sowie der Thesaurusföderation wurden eine Reihe von Algorithmen zur Berechnung der Kennzahlen realisiert. Vollständig implementiert wurde die Wissensakquisitionsarchitektur. Eine Reihe von ausgewählten, konfigurierbaren Integrationsverfahren für die verschiedenen Phasen sowie die qualitative Bewertung der von diesen Verfahren generierten und bewerteten Hypothesen wurde ebenfalls implementiert. Der besonderen Bedeutung des Benutzeragenten zum Einbringen der menschlichen Expertise sowie zur Beurteilung des Ergebnisses durch den menschlichen Experten gerecht werdend, wurde dieser auf Basis von Fischaugenansichten zur Visualisierung von Ausschnitten des Thesaurusföderationsgraphen einschließlich relevanter Konflikte implementiert (vgl. Abbildung 14.1).

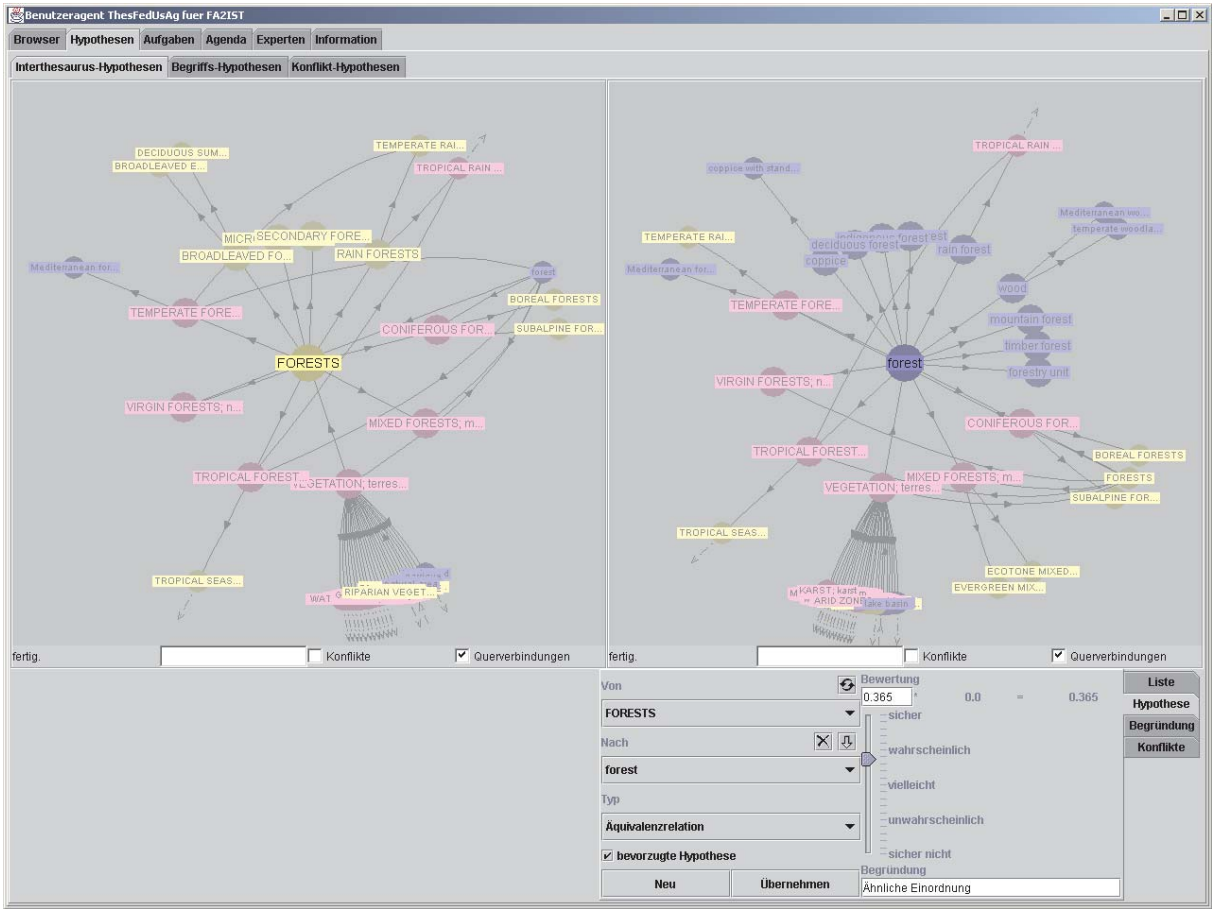

Abbildung 14.1: Visualisierung eines Ausschnittes der Thesaurusföderation durch den Benutzeragenten

Die Implementierung erfolgte in der objektorientierten Sprache Java. Sämtliche Daten wurden in der relationalen Datenbank Oracle verwaltet. Auf die Datenbank wurde mittels SQL und über JDBC (Java Database Connectivity) zugegriffen. Schnittstellen zu WordNet wurden mittels C und JNI (Java Native Interface) geschaffen.

Insgesamt umfasst das gesamte System über 300.000 Zeilen Java-Quellcode. Für die Realisierung und die Experimente stand ein Pool verschiedener Sun UltraSparcs zur Verfügung.

Mit den durchgeführten Experimenten konnte die Praxistauglichkeit des entwickelten Ansatzes gezeigt werden. Da dies im Vordergrund stand, wurde eine aufwendige Performance-Optimierung nicht vorgenommen, weder hinsichtlich der Datenhaltung noch hinsichtlich der Algorithmen. Aufgrund der Größe der betrachteten Thesauri hatten einige der Algorithmen zur Vorverarbeitung der Thesauri aber auch zum Finden und Bewerten von Inter-Thesaurus-Beziehungen und Konflikten sehr lange Laufzeiten, manche davon länger als 24 Stunden. Sollten jedoch die in dieser Zeit durchgeführten umfangreichen Analysen und Bewertungen solch komplexer Begriffssysteme durch menschliche Integrationsexperten durchgeführt werden, würde dies ein Vielfaches an Zeit und Geld kosten. Bereits durch unser prototypisches System wurde der menschliche Intergrationsexperte ganz wesentlich entlastet.

Mit dem von uns geschaffenen Rahmenwerk für die flexible Integration von heterogenen und autonomen Thesauri haben wir einen in dieser Art einzigartigen ganzheitlichen Ansatz entwickelt. Über die Einzelexperimente hinaus, die wir durchgeführt haben, müsste nun dieser Ansatz mit

anderen Ansätzen in seiner Gesamtheit verglichen werden. Eine solche Art der Evaluierung wäre jedoch eine eigene Arbeit, die weit über den Rahmen unserer Arbeit hinausgeht. Neben dem erforderlichen sehr großen Aufwand für eine solche Evaluierung kommt als Schwierigkeit hinzu, überhaupt einen vergleichbaren ganzheitlichen Ansatz zu finden.

14.2 Ausblick

Das entwickelte Rahmenwerk für Thesaurusföderationen bietet eine Reihe von Ansatzpunkten sowohl für die Weiterentwicklung als auch für die Verwendung in anderen Anwendungsfeldern.

14.2.1 Weiterentwicklung

Mögliche Ansatzpunkte für die Weiterentwicklung sind:

Problemlösungsverfahren: Wir haben bereits eine Reihe von Problemlösungsverfahren zum Auffinden und Bewerten von Vorschlägen für die Thesaurusintegration vorgestellt. Jedoch sind die Möglichkeiten solcher Verfahren, die über die klassischen Verfahren der Überprüfung auf lexikalische Gleichheit/Ähnlichkeit sowie einfache Auswertungen der Graphenstrukturen hinausgehen, u.E. längst noch nicht ausgeschöpft. Da diese Verfahren das Kernstück jeder Integration sind, sollte versucht werden, hier Verfahren weiter und neu zu entwickeln. Ansatzpunkte können etwa lernende Verfahren sein, die anhand etablierter Inter-Thesaurus-Beziehungen eigenständig weitere Beziehungen finden oder zumindest ihre Bewertungen von Hypothesen anhand der Rückkopplung der Annahme/Ablehnung der Hypothesen selbst verbessern. Die Fortschritte in der Verarbeitung natürlicher Sprache können zu einer weiteren Auswertung der natürlichsprachigen Definitionen und Erläuterungen herangezogen werden.

Wir haben im Rahmen dieser Arbeit ausschließlich Begriffsintegrationsverfahren angewandt, die die Informationen in den Thesauri selber sowie aus externen Wissensquellen analysieren. Zusätzlich sind, wenn entsprechende Korpora vorhanden sind, Verfahren möglich, die hieraus weitere Informationen auswerten.

Schließlich können weitere fachabhängige Verfahren entwickelt werden, die etwa Formeln in chemischen Thesauri oder lateinische Benennungen in biologischen/medizinischen Thesauri besonders berücksichtigen.

Kooperative Integration: Über die von uns entwickelte Architektur arbeiten die maschinellen Experten bereits kooperativ an der Thesaurusintegration. Decken die Thesauri jedoch mehrere Fachbereiche ab, ist auch ein einzelner menschlicher Experte als oberste Entscheidungsinstanz und dem Einbringen seiner Expertise rasch überfordert. Optimal wäre es, eine Plattform bereitzustellen, die mehreren Experten die Zusammenarbeit über Zeit und Raum hinweg ermöglicht. Denkbar wäre z.B. die Spezifikation von Teilbereichen für die jeweils ein Experte die Hauptverantwortung trägt. Des Weiteren könnten Protokolle von Änderungen, Diskussionsforen mit Ergebnisprotokollen, die gleichzeitige Ansicht von identischen Föderationsausschnitten, Integrationsvorschlägen und Konflikten (Application Sharing) diese kooperative Erstellung durch Experten jeweils in ihren Büros unterstützen.

Änderungsmanagement: Wir haben in unserer Arbeit ein Rahmenwerk für die Erstintegration von Komponententhesauri in Thesaurusföderationen geschaffen. Die Komponententhesauri werden jedoch in der Regel weiterentwickelt und den geänderten Anforderungen und

terminologischen Änderungen angepasst. Daher sind Neuerscheinungen von Thesauri im Ein- oder Zwei-Jahres Rhythmus üblich. Eine bestehende Herausforderung ist es, bei der Integration eines solchen modifizierten Thesaurus den Aufwand gegenüber der Erstintegration deutlich zu verringern. Grundlage hierzu wäre ein allgemeines Versionsmanagement für Thesauri, das aber nicht existiert. Aktuelle Arbeiten von Noy und Musen [NM02], die die Unterschiede und Gemeinsamkeiten in verschiedenen Versionen von Ontologien erkennen, könnten jedoch Grundlage für ein umfassendes Änderungsmanagement sein. Aus dem Forschungsbereich der Aktiven Föderierten Datenbanksysteme kommen aktive Mechanismen zum Erkennen lokaler Modifikationen, die ebenfalls für das Änderungsmanagement genutzt werden können, vgl. z.B. [Kos99].

14.2.2 Übertragbarkeit

Prinzipiell kann der von uns entwickelte Ansatz auf eine Reihe von Integrationsproblemen übertragen werden. Am nahe stehendsten ist sicher die Integration von Ontologien. In vielen Ontologien sind die Beziehungstypen ausschließlich Bestands- und Abstraktionsbeziehungen, so dass eine Übertragung des Ansatzes einfach möglich ist. Sind zusätzliche Erweiterungen gegenüber dem von uns verwendeten Thesaurusmodell vorhanden, gilt es, diese in den Informationsmodellen, den Kennzahlen für die Analyse sowie den Verfahren und dem Benutzeragent zu berücksichtigen. So wird etwa in [NM00] ein allgemeines Frame-basiertes Ontologie-Modell vorgestellt. Die Slots in diesem Modell bezeichnen aber im Wesentlichen Bestandsunterbegriffe, Beziehungen zwischen den Frames sind Abstraktionsbeziehungen. Nur die Facets, durch die als tertiäre Relationen zusätzliche Constraints wie Kardinalitaten und Wertebereiche ausgedrückt werden können, müssten zusätzlich berücksichtigt werden.

Die durch unseren generellen Ansatz gewonnene Flexibilität und Skalierbarkeit sollte auch aktuelle Ansätze der Ontologie-Integration (vgl. z.B. [NM00, Cha00, SM01]) deutlich verbessern.

Generelle Ideen unseres Ansatzes, wie die kombinierte Blackboard-/Workflow-Architektur zur Trennung von Lösungsstrategie und Lösungsverfahren, das Zulassen von Konflikten in autonomen Quellen und das kontextabhängige Auflösen dieser Konflikte oder das Gewinnen von Erkenntnissen über die zu integrierenden Komponenten sowie einen Zwischenstand der Integration anhand von berechenbaren Kennzahlen, lassen sich jedoch auch auf vollständig andere Informationsmodelle übertragen und können z.B. die semantische Integration von Katalogen oder Metainformationssystemen sowie Informationssystemen allgemein verbessern.

Anhang A

Aufgaben-Agenda-Definitions-Sprache AADS

Die in Abschnitt 7.3.2.3 eingeführte Aufgaben-Agenda-Definitions-Sprache AADS (s. Seite 120f) wird in diesem Anhang durch die Festlegung einer XML-DTD spezifiziert.

Die DTD für AADS orientiert sich an der WPDL (Workflow Process Definition Language) [Wor99] sowie einer ersten Umsetzung in XML durch Robert Tolksdorf, Amarilis Macedo-Aranya und Marc Stauch [Tol00]. Da für die Spezifikation einer Aufgaben-Agenda nur eine Teilmenge der in WPDL möglichen Sprachkonstrukte benötigt wird, beinhaltet die AADS-DTD keine Anwendungen, Teilnehmer, Organisationseinheiten und Rollen, keine (operativen) Daten und keine Elemente für geschätzte Kosten und Dauer. Ergänzend zu den WPDL-Sprachelementen wurde ein Element zum Ausdrücken des Zustandes einer Aktivität vorgesehen. Weitere Abweichungen vom WPDL-Standard sind in der DTD selbst kommentiert.

```
<!-- Entity-Definitionen zur Angabe der Existenzanforderung an Objekte -->
<!ENTITY %  OptRequired     "#IMPLIED">
<!ENTITY %  Required        "#REQUIRED">
<!ENTITY %  Optional        "#IMPLIED">

<!-- Entitiy-Definitionen zur Angabe spezifischer Datentypen -->
<!ENTITY % Identifier        "ID">
<!ENTITY % IdRef             "IDREF">
<!ENTITY % String            "CDATA">
<!ENTITY % Date              "CDATA">
<!ENTITY % Cardinal          "CDATA">
<!ENTITY % Priority          "CDATA">
<!ENTITY % Classification    "CDATA">
<!ENTITY % Reference         "CDATA">
<!ENTITY % Mode              "(AUTOMATIC|MANUAL)">
<!ENTITY % Version           "%String;">
<!ENTITY % Name              "%String;">
<!ENTITY % ModelId           "%Identifier;">
<!ENTITY % ProcessId         "%Identifier;">
<!ENTITY % ProcessRef        "%IdRef;">
<!-- statt einem Verweis auf einen Ausfuehrenden (Experten)
     wird die Aufgabe, die der Ausfuehrende erledigen
     koennen muss, beschrieben -->
<!ENTITY % Task              "%String;">
<!ENTITY % TransitionId      "%Identifier;">
<!ENTITY % ActivityId        "%Identifier;">
<!ENTITY % ActivityRef       "%IdRef;">
```

```
<!ENTITY % ParticipantAssignment "%IdRef;">
<!-- JAVA-Ausdruck -->
<!ENTITY % Condition        "%String;">

<!ENTITY % RedefinableHeader
 "AUTHOR        %String;             %Optional;
  VERSION       %Version;            %Optional;
  CHARACTERSET  %String;             %Optional;
  CODEPAGE      %String;             %Optional;
  COUNTRY-KEY   %String;             %Optional;
  <!-- RESPONSIBLE entfernt -->
  STATUS        (under-revision|released|under-test)    %Optional;">

<!ENTITY % ExecMode "(ASYNCHR|SYNCH)">

<!ENTITY % LoopKind "(WHILE|REPEAT-UNTIL)">

<!-- ================================================
Gemeinsame Elemente
================================================= -->

<!ELEMENT DESCRIPTION ANY>

<!-- ================================================
Workflow-Modell
================================================= -->
<!ELEMENT MODEL (
  DESCRIPTION?,
  WORKFLOW*
)>
<!ATTLIST MODEL
  ID              %ModelId;      %Required;
  NAME            %Name;         %Optional;
  WPDL_VERSION    %String;       %OptRequired;
  VENDOR          %String;       %OptRequired;
  CREATED         %Date;         %OptRequired;
  DOCUMENTATION   %Reference;    %Optional;
  %RedefinableHeader;
>

<!-- ================================================
Workflow process
================================================= -->
<!ELEMENT WORKFLOW (
   DESCRIPTION?,
   (ACTIVITY|TRANSITION)*
)>
<!ATTLIST WORKFLOW
  ID             %ProcessId;      %Required;
  NAME           %Name;           %Optional;
  CREATED        %Date;           %Optional;
  PRIORITY       %Priority;       %Optional;

  CLASSIFICATION %Classification; %Optional;
  DOCUMENTATION  %Reference;      %Optional;
>

<!ELEMENT ACTIVITY (
  DESCRIPTION?,
  (NOIMPLEMENTATION|APPLICATIONS|SUBFLOW|LOOP),
```

```
  ACCESSRESTRICTION?,
  TRANSITIONRESTRICTION?
)>
<!ATTLIST ACTIVITY
  ID             %ActivityId;     %Required;
  NAME           %Name;           %Optional;
  DOCUMENTATION  %Reference;      %Optional;
>

<!ENTITY  % ImplementationAttributes
  "TASK          %Task;           %Optional;
  START-MODE     %Mode;           %Optional;
  FINISH-MODE    %Mode;           %Optional;
  PRIORITY       %Priority;       %Optional;
  DOCUMENTATION  %Reference;      %Optional;
  INSTANTIATION  (ONCE|MULTIPLE)  %Optional;
">

<!ELEMENT NOIMPLEMENTATION EMPTY>
<!ATTLIST NOIMPLEMENTATION
  %ImplementationAttributes;
>

<!-- keine Angabe von Tools zu applications erforderlich -->
<!ELEMENT APPLICATIONS EMPTY>
<!ATTLIST APPLICATIONS
  %ImplementationAttributes;
>

<!ELEMENT SUBFLOW EMPTY>
<!ATTLIST SUBFLOW
  %ImplementationAttributes;
  REF            %ProcessRef;     %Required;
  EXEC           %ExecMode;       %Optional;
>

<!ELEMENT LOOP EMPTY>
<!ATTLIST LOOP
  %ImplementationAttributes;
  KIND           %LoopKind;       %Required;
  CONDITION      %Condition;      %Required;
>

<!ELEMENT ACCESSRESTRICTION EMPTY>

<!ELEMENT TRANSITIONRESTRICTION EMPTY>
<!ATTLIST TRANSITIONRESTRICTION
  JOIN           (AND|XOR)        %Optional;
  SPLIT          (AND|XOR)        %Optional;
  TRANSITIONS    %String;         %Optional;
>

<!ELEMENT TRANSITION (
  DESCRIPTION?
)>
<!ATTLIST TRANSITION
  ID             %TransitionId;   %Required;
  NAME           %Name;           %Optional;
  FROM           %ActivityRef;    %Required;
  TO             %ActivityRef;    %Required;
```

```
  CONDITION       %Condition;       %Optional;
  LOOP            (FROM|TO)         %Optional;
>
```

Anhang B

Spezifikation der Expertenein-/-ausgaben

Um einen (maschinellen) Experten, der ein Problemlösungsverfahren implementiert, hinsichtlich seinem möglichen Beitrag zur Lösung einer Aufgabe bewerten zu können, ist eine Beschreibung der erforderlichen Eingaben oder Voraussetzungen sowie seiner Ausgaben oder potenziellen Beiträge erforderlich. Unter Vernachlässigung einer Syntax wird in diesem Anhang dargestellt, welche Informationen für eine solche Beschreibung benötigt werden.

B.1 Informationen über erforderliche Eingaben

Relevant für die Beschreibung der Eingaben sind folgende Informationen:

Etablierte Inter-Thesaurus-Beziehungstypen: Falls die Experten bereits etablierte Inter-Thesaurus-Beziehungen benötigen, um Hypothesen zu generieren bzw. zu bewerten, werden die entsprechenden Typen dieser vorausgesetzten Beziehungen aufgeführt.

Hypothesentypen: Die Hypothesentypen sind eine Auflistung der Typen, von denen jeweils mindestens eine Hypothese auf dem Blackboard vorhanden sein muss, damit der Experte Hypothesen generieren bzw. bewerten kann.

B.2 Informationen über mögliche Ausgaben

Zu den Ausgaben sind folgende Informationen erforderlich:

Hypothesengenerierung: Wird eine Hypothesengenerierung als möglich angegeben, bedeutet dies, der Experte ist in der Lage, neue Hypothesen zu generieren und diese initial zu bewerten. Anderenfalls kann der Experte ausschließlich bereits vorhandene Hypothesen bewerten.

Hypothesentyp: Der Hypothesentyp sagt aus, welcher Typ von Hypothese generiert bzw. bewertet wird.

Polarität: Die Polarität ist ein Indikator, der aussagt, ob positive, negative oder sowohl positive als auch negative Bewertungen erzeugt werden.

Die Informationen zur Hypothesengenerierung, zum Hypothesentyp und zur Polarität werden als eine Einheit angegeben. Für einen Experten sind jeweils Mehrfachangaben solcher Einheiten möglich. Dies ermöglicht es, auch komplexere Verfahren, die Hypothesen verschiedenen Typs erzeugen bzw. bewerten, einzubringen. Der Nachteil solcher komplexeren Verfahren ist jedoch, dass diese weniger gezielt zum Lösen bestimmter Aufgaben eingesetzt werden können.

Anhang C

SOAP-Repräsentation einer Anfrage an den Mediator

Für das Serialisieren eines entfernten Funktionsaufrufes in einer SOAP-Nachricht wurden in der SOAP-Spezifikation [Wor00b] strikte Regeln definiert, um die gewünschte Interoperabilität zwischen verschiedenen Systemen sicher zu stellen:

- Ein Funktionsaufruf wird durch einen einzigen XML-Struct repräsentiert, dessen Namen und Typ mit der entsprechenden Funktion übereinstimmen.
- Jeder Eingabe- oder Ein-/Ausgabe-Parameter wird durch ein Element in diesem Struct repräsentiert. Dabei korrespondieren auch hier Name und Typ des Elements mit Namen und Typ des Parameters. Die Parameter-Elemente stehen auch in derselben Reihenfolge wie in der Signatur der Funktion.
- Das Resultat eines Funktionsaufrufs wird ebenfalls durch einen einzigen Struct repräsentiert. Dabei ist der Name des Structs nicht signifikant, es wird jedoch ein Anhängen des Wortes *Response* an den ursprünglichen Funktionsnamen als Konvention vorgeschlagen.
- Jeder Ausgabe- oder Ein-/Ausgabe-Parameter wird durch ein Element im Ergebnis-Struct repräsentiert. Dabei ist das erste Element in dem Struct das Resultat des Aufrufs, wobei sein Name nicht signifikant ist. Es folgen die Parameter in der selben Reihenfolge wie in der Signatur des Funktionsaufrufs.
- Ein evtl. aufgetretener Fehler muss durch Benutzen des SOAP-Fehler-Elementes übermittelt werden. Falls vom Übertragungsprotokoll weitere Regeln zum Übermitteln von Fehlermeldungen definiert werden (wie bei HTTP), müssen diese auch befolgt werden.
- Da ein Resultat ein erfolgreiches Ausführen eines RPCs signalisiert, ein Fehler-Element aber einen Fehler, ist es unzulässig, in der Antwort auf einen RPC sowohl ein Resultat als auch einen Fehler zu übermitteln.

Als Beispiel für die Repräsentation einer Anfrage in SOAP wird die Navigationsanfrage an den Thesaurusföderationsmediator genannt, die alle Abstraktionsunterbegriffe eines Deskriptors innerhalb einer gegebenen Menge von Thesauri zu liefert. Bei diesem Funktionsaufruf gibt es ausschließlich Eingabe-Parameter und die einzige Ausgabe ist das Ergebnis.
Die Anfrage sieht wie folgt aus:

```
POST /Thesaurusfoederation HTTP/1.1
Host: www.thesaurusfoederation.de
Content-Type: text/xml; charset="utf-8"
Content-Length: 977
SOAPAction: "http://www.thesaurusfoederation.de/thesmediator/"

<SOAP-ENV:Envelope
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/envelope/">
    <SOAP-ENV:Body>
       <M:getNarrowerAbstract xmlns:M="http://www.thesaurusfoederation.de/thesmediator/">
         <Descriptor>
            <!-- der foederierte Begriff wird durch die durch
                 Inter-Thesaurus-Aequivalenzbeziehungen verbundenen
                 Begriffe der Komponententhesauri repraesentiert -->
            <CompThesDescriptor thesID='1'   langID='94' descID='3505' term='fuel'/>
            <CompThesDescriptor thesID='982' langID='94' descID='3136' term='FUELS'/>
         </Descriptor>
         <Thesauri>
            <!-- Thesaurus-IDs genügen als Identifikatoren -->
            <Thesaurus thesID='1'/>   <!-- GEMET -->
            <Thesaurus thesID='982'/> <!-- AGROVOC -->
         </Thesauri>
       </M:getNarrowerAbstract>
    </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Die entsprechende Antwort lautet:

```
HTTP/1.1 200 OK
Content-Type: text/xml; charset="utf-8"
Content-Length: 1790

<SOAP-ENV:Envelope
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/envelope/">
    <SOAP-ENV:Body>
      <M:getNarrowerAbstractResponse xmlns:M="http://www.thesaurusfoederation.de/thesmediator/">
      <Descriptors>
        <Descriptor>
          <CompThesDescriptor thesID='1'   langID='94' descID='12011' term='biofuel'>
            <CompDefinition>
              A gaseous, liquid, or solid fuel that contains an energy content derived from a
              biological source. The organic matter that makes up living organisms provides a
              potential source of trapped energy that is beginning to be exploited to supply
              the ever-increasing energy demand around the world. An example of a biofuel is
              rapeseed oil, which can be used in place of diesel fuel in modified engines. The
              methyl ester of this oil, rapeseed methyl ester (RME), can be used in unmodified
              diesel engines and is sometimes known as biodiesel. Other biofuels include biogas
              and gasohol.
            </CompDefinition>
          <CompThesDescriptor>
          <CompThesDescriptor thesID='982' langID='94' descID='27465' term='BIOFUELS'>
            <CompNonDescriptors>
              <CompNonDescriptor term='BIODIESEL'/>
              <CompNonDescriptor term='BIOMASS FUEL'/>
            </CompNonDescriptors>
          <CompThesDescriptor>
        </Descriptor>
        <Descriptor>
```

```
          <!-- Deskriptor aus Thesaurus der Ergaenzenden Begriffe -->
          <CompThesDescriptor thesID='77'   langID='94' descID='2' term='conventional fuel'/>
        </Descriptor>
      </Descriptors>
    </M:getNarrowerAbstractResponse>
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Obwohl in der Anfrage der Thesaurus der Ergänzenden Begriffe nicht explizit aufgeführt war, können Deskriptoren des Ergebnisses (hier: *conventional fuel*) immer auch aus diesem Thesaurus stammen, der somit nicht ausgeblendet werden kann.

Anhang D

Glossar

Innerhalb dieses Glossars werden die im Rahmen dieser Arbeit verwendeten grundlegenden Begriffe nach Bereichen geordnet dargestellt und definiert. Die Definitionen versuchen, den allgemeinen Konsens über Bedeutung und Verwendung dieser Begriffe wiederzugeben. Wo dies aufgrund uneinheitlicher Verwendung nicht möglich ist, wird entsprechend darauf hingewiesen, dass diese Definition im Rahmen dieser Arbeit, nicht aber unbedingt in der gesamten Literatur gilt. Auf formale Definitionen wird in diesem Gloassar verzichtet. Sie werden, insofern es für die Arbeit erforderlich ist, in Kapitel 5 und 6 geliefert. Stattdessen werden wir aber noch ergänzende Erläuterungen liefern, welche Bedeutung den Begriffen derzeit in Wissenschaft und Praxis eingeräumt wird.

D.1 Allgemeine Begriffe

Deskriptive Sprachen: Deskriptive Sprachen (description logics) werden insbesondere im Bereich der künstlichen Intelligenz zur Repräsentation von Begriffen und Begriffsnetzen verwendet (vgl. z.B. [WV99]). In der Regel wird ihnen eine deklarative Semantik nach Tarski gegeben, so dass sie als Unter-Sprache der Prädikatenlogik gesehen werden können.

Deskriptive Sprachen besitzen eine formale Semantik. Grundkonzepte deskriptiver Sprachen sind atomare Begriffe (concepts) und Rollen (roles), Individuen und Konstruktoren. Begriffe (z.B. *Tier*, *Gruen*) beschreiben gemeinsame Eigenschaften einer Menge von Individuen und können als unäre Prädikate betrachtet werden, die als Menge von Individuen interpretiert werden. Rollen werden als binäre Relationen zwischen Individuen interpretiert (z.B. *Farbe*). Die Konstruktoren (z.B. $\wedge, \forall$) können zur Definition neuer Begriffe verwendet werden (z.B. der komplexe Begriff Frosch in Prädikatenlogik erster Ordnung als $Frosch \doteq Tier(x) \wedge \forall y : (Farbe(x, y) \rightarrow Gruen(y))$ mit der freien Variablen x).

Außer der formal repräsentierten Semantik von Begriffen und ihren Zusammenhängen können auf deskriptiven Sprachen basierende Systeme (z.B. KL-ONE [BS85], LOOM [McG91]) Schlussfolgerungen anhand der Wissensbasis ziehen. Die wesentlichen Schlussfolgerungen sind Klassifikation, Erfüllbarkeit (satisfiability), Subsumierung und Instanz-Überprüfung. Die Subsumierung repräsentiert dabei die Abstraktionsbeziehung (Ist-ein-Beziehung). Die Klassifikation ist die Berechnung einer Begriffshierarchie basierend auf der Subsumierung.

Effizienz: Die besonders wirtschaftliche Ausnutzung der eingesetzten Ressourcen nach dem Wirtschaftlichkeitsprinzip, möglichst viel Ertrag je eingesetztem Aufwand zu erzielen.

Effektivität: Wirksamkeit der eingesetzten Mittel in dem Sinn, dass mit ihnen das erwünschte Ziel bestmöglich erreicht wird.

Invariante: Prädikat über den Variablen und Datenstrukturen eines Systems, das immer *wahr* ist.

Konflikt: Verstoß gegen eine →Invariante.

D.2 Begriffe aus den Bereichen Terminologielehre und Information Retrieval

Begriff: Ein *Begriff* ist ein Sachverhalt oder eine Idee selber oder, wie vom Technischen Kommitee 37 (TC37) der ISO in [ISO90] definiert, eine „Einheit des Denkens".

Benennung: Eine *Benennung* – auch Bezeichnung oder Term genannt – stellt ein sprachliches Mittel dar, um einen →Begriff auszudrücken [DIN93b]. Benennungen können →Einwortbenennungen oder →Mehrwortbenennungen sein.

Einwortbenennung: →Benennung, die aus einem Wort besteht (vgl. Anhang D.3). *Elementare Einwortbenennungen* bestehen aus einer bedeutungstragenden Einheit (neben evtl. vorhandenen Flexionselementen, z.B. Dünger, Regel). *Komplexe Einwortbenennungen* bestehen aus zwei oder mehr bedeutungstragenden Einheiten (z.B. Mineraldünger oder EU-Länder als Kompositum oder Regelung als Derivat von Regel mit dem Suffix -ung; vgl. Anhang D.3)[DIN93b].

Index: Menge von verwendeten →Begriffen eines →Indexierungsvokabulars, die auf Dokumente verweisen, die mit diesen Begriffen inhaltlich beschrieben sind.

Indexierung: Entscheidungsprozess, welche →Begriffe eines →Indexierungsvokabular von genügend großer Bedeutung sind, um ein Dokument im Index inhaltlich zu beschreiben [CM63].

Indexierungsvokabular: Kontrollierte Menge von →Begriffen, die zur →Indexierung von Dokumenten eines bestimmten Dokumentenbestandes verwendet werden dürfen. Das Indexierungsvokabular kann z.B. in Form eines →Thesaurus bereitgestellt werden.

Information-Retrieval-System: Ein Information-Retrieval-System ist die vollständige Organisationstruktur, die sich mit der Erlangung, Speicherung und Verfügbarmachung von Informationen befasst [CM63]. Diese Definition umfasst eine Bibliothek genauso wie ein →rechnergestütztes Information-Retrieval-System.

Informationssystem: Der Begriff *Informationssystem* wird in dieser Arbeit abkürzend für →rechnergestütztes Information-Retrieval-System verwendet.

Interlingua: Eine *Interlingua* im Kontext der Thesaurus- oder Ontologie-Integration ist eine Menge von Begriffen, die durch Beziehungen miteinander verbunden sind. Die Interlingua wird konstruiert, um Beziehungen zwischen den Begriffen eines Thesaurus und den Begriffen eines anderen Thesaurus indirekt über die Begriffe der Interlingua ausdrücken zu können. Die Interlingua enthält daher in der Regel sowohl eine Teilmenge der Begriffe und Beziehungen der zu integrierenden Thesauri als auch zusätzliche Begriffe und Beziehungen, die zum Zwecke der Integration ergänzt wurden.

Komponententhesaurus: Ein →Thesaurus, der an einem →Multi-Thesaurus-System teilnimmt.

Mehrwortbenennung: Benennung, die aus mehreren Wörtern (vgl. Anhang D.3) besteht (auch: Wortgruppe). Die Wörter selbst können elementar oder komplex sein (organischer Dünger, ausser haus Verzehr, Auswirkung auf die Umwelt)[DIN93b]. In einem Thesaurus sind solche Mehrwortbenennungen häufig Nominalphrasen (vgl. Anhang D.3).

Monolingualer Thesaurus: Ein *einsprachiger* oder *monolingualer Thesaurus* ist ein →Thesaurus, dessen →Benennungen überwiegend aus einer (natürlichen) Sprache stammen. Fremdsprachige Benennungen können aufgenommen werden, wenn deren Gebrauch in der Sprache des Thesaurus üblich ist oder eine zukünftige Übersetzung angestrebt wird. In der Regel werden die fremdsprachigen Benennungen als Nicht-Deskriptoren (→Thesaurus) aufgeführt.

Multilingualer Thesaurus: Ein *mehrsprachiger* oder *multilingualer Thesaurus* ist ein →Thesaurus, der für die jeweiligen Begriffe äquivalente Deskriptoren (gegebenenfalls auch Nicht-Dekriptoren) in jeder der Thesaurussprachen enthält [DIN93a]. Der Vorteil eines mehrsprachigen Thesaurus ist, dass Benutzer nicht mit den verschiedenen Sprachen vertraut sein müssen, sondern *eine* Sprache zum Retrieval verwenden können, auch wenn eine andere Sprache zur Indexierung verwendet wurde. Aus diesem Grund sind mehrsprachige Thesauri häufig die Basis für einen mehrsprachigen Zugang zu Informationssystemen.

Die DIN-Norm unterscheidet zwischen mehrsprachigen Thesauri mit einer Hauptsprache, nach der sich die Struktur und die Benennungen aller übrigen (Sekundär-)Sprachen richten, und mehrsprachigen Thesauri mit Sprachen, die den gleichen Status haben. Im letzten Fall müssen für alle Begriffe in allen Sprachen äquivalente Begriffe vorhanden sein bzw. gebildet werden.

Ein mehrsprachiger Thesaurus enthält im Vergleich zu einem monolingualen Thesaurus eine weitere Äquivalenzrelation: Die *Intersprach-Äquivalenzrelation*. Diese verbindet die Deskriptoren der verschiedenen Sprachen miteinander. Nicht-Deskriptoren werden in der Regel nicht übersetzt.

Multi-Thesaurus-System: Ein Software-System, dass den gleichzeitigen Zugriff auf mehrere Thesauri erlaubt. Der Integrationsgrad der Thesauri in einem Multi-Thesaurus-System kann von unverbunden bis hin zu vollständig integriert reichen.

Precision: Mit *Precision* wird das Maß für die Genauigkeit des Retrievalergebnisses bezeichnet. Precision ist definiert als das Verhältnis der Anzahl der gefundenen →relevanten Dokumente zur Anzahl aller gefundenen Dokumente [Sal87].

Recall: Als *Recall* wird das Maß für die Vollständigkeit des Retrievalergebnisses bezeichnet. Recall wird definiert als das Verhältnis zwischen der Anzahl der gefundenen →relevanten Dokumenten und der Gesamtanzahl der im Dokumentenbestand vorhandenen →relevanten Dokumente [Sal87].

Relevanz: Ein Dokument ist für einen Benutzer relevant, wenn der Benutzer dieses Dokument als Ergebnis seiner Anfrage haben möchte [CLvRC98].

Rechnergestütztes Information-Retrieval-System: Ein →Information-Retrieval-System, bei dem wesentliche Teilaufgaben von Rechnersystemen ausgeführt werden.

Semantisches Information Retrieval: Da die Dokumente eines →Informationssystems bei der Verwendung eines →Thesaurus nicht mit den in diesen Dokumenten vorkommenden Zeichenketten indexiert werden, sondern mit Begriffen, die diese Zeichenketten ausdrücken, unterstützen Thesauri semantisches Information Retrieval.

Semantische Netze: Semantische Netze wurden von [Qui68] zur Repräsentation der Semantik von natürlicher Sprache entwickelt. Begriffe und Instanzen werden als Knoten in einem Graphen dargestellt. Dabei können die Knoten entweder Eigenschaftsknoten sein (z.B. eine Farbe) oder Begriffs- (z.B. Frosch) bzw. Instanzknoten (z.B. Kermit). Zwischen den Begriffen bzw. Instanzen werden Beziehungen durch gerichtete Kanten ausgedrückt. Werden die Kanten nicht beschriftet, besitzen sie in der Regel eine Semantik im Sinne von „ist ein“ (Abstraktionsbeziehung). Durch Beschriftung der Kanten können andere Beziehungen eingeführt werden.

Semantische Netze werden eher im Bereich des Textverstehens als für die Recherche verwendet (s. z.B. [Vie97, S. 29]).

Term: S. →Benennung.

Thesaurus: Nach [DIN87] ist ein Thesaurus „... eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) →Benennungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient. Er ist durch folgende Merkmale gekennzeichnet:

- →Begriffe und →Benennungen werden eindeutig aufeinander bezogen (terminologische Kontrolle) ...
- Beziehungen zwischen Begriffen (repräsentiert durch ihre Benennungen) werden dargestellt.“

Ontologie: Ursprünglich stammt der Begriff *Ontologie* aus der Philosophie und bezeichnet die Lehre vom Wesen und den Eigenschaften des Seienden, die zu zeigen hat, was allen Seienden als solchen gemeinsam ist [DH90]. Er wurde von der Informatik geborgt und stellt dort ein Mittel zu einer formalen Beschreibung einer abstrakten, vereinfachten, von verschiedener Seite anerkannter Sichtweise eines Ausschnittes der Welt, einer so genannten Konzeption, dar. Diese Definition, die in [Gru92] zusammenfassend als *Eine Ontologie ist eine formale, explizite Spezifikation einer gemeinsamen Konzeption* („An ontology is a formal, explicit specification of a shared conceptualization“) dargestellt wird, ist die am meisten verwendete. Sie lässt aber sehr viele Spielräume für konkrete Ausprägungen von Ontologien, insbesondere, da dem verwendeten Begriff *Konzeption* keine eindeutige Bedeutung zugrunde liegt. Entsprechend vielfältig sind die Systeme, die in der Literatur als Ontologie bezeichnet werden. In der Regel bestehen solche Ontologien aus Begriffen, Relationen zwischen den Begriffen und Axiomen.

→Thesauri können somit als eine Ausprägung von Ontologien angesehen werden, bei der die Relationstypen eingeschränkt sind und explizite Axiome fehlen. Des Weiteren werden in der Regel von Ontologien möglichst formale Definitionen der Begriffe erwartet. Die natürlichsprachigen Definitionstexte der Thesauri erfüllen dieses Kriterium nicht, wohl aber die durch synonyme Benennungen und die Relationen gegebenen Definitionen.

WordNet: WordNet [Mil98, Fel98] ist ein umfangreiches →Wortnetz der englischen Sprache, das an der Princeton Universität entwickelt wurde und wird. WordNet fasst in einem Kontext identische Benennungen zu so genannten Synsets (Synonymmengen) zusammen, die

Begriffe repräsentieren. Zwischen diesen Begriffen sind u.a. Abstraktions- und Bestandsrelationen etabliert. Lexikalische Grundeinheit sind Wörter. Mehrwortbenennungen und Phrasen sind in geringem Umfang enthalten. Inhaltlich deckt WordNet umfangreich das allgemeine Wortgut ab (WordNet Version 1.6 enthält z.B. über 80.000 Substantive, die zu über 60.000 Begriffen zusammengefasst sind), nicht aber fachspezifisches Wortgut.

Der größte Unterschied zwischen WordNet, einem konventionellen Wörterbuch und einem Thesaurus besteht darin, dass WordNet das ganze Lexikon in fünf syntaktische Kategorien oder Wortarten aufteilt (Substantive, Verben, Adjektive, Adverben und funktionale Wörter), während ein konventionelles Wörterbuch alle Einträge alphabetisch darstellt, ohne zwischen Wortarten zu unterscheiden, und ein Thesaurus im Wesentlichen nur Nomen enthält. Die Wortarten werden in separaten Datenbanken abgelegt. Der Preis für diese Aufteilung ist eine gewisse Redundanz, die ein normales Wörterbuch vermeidet: Wörter, die als Substantive oder als Verben auftreten können (im Englischen z.B. das Wort *back*), sind in mehr als einer Wortartdatenbank gespeichert. Ein großer Vorteil besteht jedoch darin, dass die Unterschiede in der semantischen Organisation dieser syntaktischen Wortarten deutlicher gesehen und systematischer erforscht werden können.

Wortgruppe: →Mehrwortbenennung.

Wortnetz: Thesauri speichern Wissen über Begriffe und ihre Beziehungen. Darüber hinausgehend ist für die maschinelle Verarbeitung von Benennungen, die die Begriffe repräsentieren, auch lexikalisches Wissen erforderlich. Wortnetze speichern beide Arten von Wissen. Das bekannteste und umfangreichste Wortnetz ist →WordNet.

D.3 Begriffe aus der Linguistik

Affix: Oberbegriff für →Präfix, →Infix und →Suffix.

Begriffsinhalt: Der *Begriffsinhalt* (auch: Intension) bezeichnet die Gesamtheit der Merkmale oder Eigenschaften eines Begriffes [DIN92].

Begriffsumfang: Der *Begriffsumfang* (auch: Extension) bezeichnet die Gesamtheit der Objekte, die durch den Begriff bezeichnet sind [Bußß90]. In [DIN92] wird statt Begriffsumfang die Bezeichnung Klasse verwendet, der Begriffsumfang anders definiert. Wir verwenden den in der Informatik, →Computerlinguistik und in der Sprachwissenschaft eingeführten Begriff.

Computerlinguistik: „Disziplin zwischen Linguistik und (angewandter) Informatik, die sich mit der maschinellen Verarbeitung natürlicher Sprachen (auf allen Beschreibungsebenen) befasst“ [Bußß90].

Derivat: Ergebnis des Wortbildungsvorgangs durch →Derivation (Ableitung)[Bußß90].

Derivation: Wortbildung durch Versehung eines freien Morphems mit einem oder mehreren →Affixen, Lautveränderung (trinken – Trank), Rückbildung (Schau aus schauen) oder Konversion in eine andere Wortklasse (deutsch – Deutsch) [Bußß90]

Flexionsform: Wortform, die durch Deklination (Nomen), Konjugation (Verb) oder Komparation (Adjektiv) von der →Grundform des Wortes abweicht.

Grundform: Infinitivform eines Verbes (z.B. gehen), Nominativ Singular eines Nomen (z.B. Baum), Positiv (oder Grundstufe) eines Adjektives (z.B. interessant).

Homonyme: Ausdrücke, die über eine gleiche Ausdrucksform hinsichtlich Orthographie und Aussprache verfügen, aber unterschiedliche Bedeutungen, die nichts miteinander zu tun haben, besitzen und unterschiedlicher etymologischer Herkunft sind (Beispiele: Ton, Tau) [Buß90]. Da die etymologische Herkunft nicht immer exakt bestimmt werden kann, verwenden wir die Begriffe Homonyme (bzw. Homonymie) auch als Oberbegriff für Homonyme (bzw. Homonymie) und →Polyseme (bzw. Polysemie).

Infix: Gebundenes →Morphem, das in den →Stamm eingefügt wird (-n- im lateinischen iungere (verbinden) von iugum (Joch)) [Buß90].

Kompositum: Ausdruck, der durch Komposition aus mindestens zwei freien →Morphemen zusammengesetzt ist (z.B. Dorf-kirche, Rind-er-braten). Die Nahtstelle zwischen den Konstituenten können Fugenelemente oder Verbindungen sein (-er- in Rinderbraten)[Buß90].

Lemma: →Grundform eine →Wortes.

Lemmatisierung: Reduktion der Flexionsform eines Wortes auf seine →Grundform.

Linguistik: Teildisziplin der Sprachwissenschaften, die als synchron (d.h. zeitlich fixiert) orientierte, auf die interne Struktur der Sprache bezogene Wissenschaft die sprachlichen Regularitäten auf allen Beschreibungsebenen (Phoneme, Morpheme, Sätze usw.) untersucht und ihre Ergebnisse in expliziter (formalisierter) Beschreibungssprache und in integrierten Modellen ablegt (nach [Buß90]).

Morphem: Kleinstes bedeutungstragendes Element der Sprache. Es wird unterschieden zwischen *freien Morphemen*, die lexikalische (Buch, rot, schnell; Objekte oder Sachverhalte etc. der Welt) oder grammatikalische (aus, und, es; zum Ausdrücken grammatikalischer Beziehungen im Satz) Bedeutung besitzen können und *gebundenen Morphemen*, die entweder lexikalische Stammmorpheme (Sprach- in Sprachanalyse), Flexionsmorhpeme (wie Verbindungen, z.B. -er- in Rinderbraten) oder Ableitungsmorpheme (zer-, -bar, -nis; →Affixe) sind.

Nominalphrase: Zusammengehörige Gruppe von →Wörtern, bestehend aus nominalen Ausdrücken mit entsprechenden attributiven Erweiterungen (Beispiel: kandiertes Obst, Wirkung auf die Umwelt) [Buß90].

Polyseme: Ausdrücke, die über eine gleiche Ausdrucksform hinsichtlich Orthographie und Aussprache verfügen, aber unterschiedliche Bedeutungen, die in engem Zusammenhang stehen, besitzen und gleicher etymologischer Herkunft sind (Beispiel: grün im Sinne von „unerfahren“, „frisch“ oder „roh“) [Buß90]. Vgl. →Homonyme.

Präfix: Dem →Stamm vorausgehendes gebundenes →Morphem (Ur- in Urwald, ent- in entspannen) [Buß90].

Stamm: Der *Stamm* (auch: Wortstamm, Wurzel, Basismorphem) ist das zu den Wörtern einer Wortfamilie (Wörter mit gleicher etymologischer Herkunft) grundlegende →Morphem, das Träger der urprünglichen lexikalischen Grundbedeutung ist (les- als Stamm von lesen, liest, Leser, unlesbar) [Buß90].

Suffix: In der Regel gebundenes →Morphem, das an den →Stamm angehängt wird (-ung in Regelung, -nis in Bildnis, -en in entspannen) [Buß90].

Synonym: Zwei (oder mehr) unterschiedliche Ausdrücke werden als Synonyme bezeichnet, wenn sie bedeutungsgleich sind (Beispiel: Stockwerk und Etage).

Wort: Kleinster, relativ selbständiger Träger von Bedeutung, der im Lexikon kodifiziert ist [Buß90, Definition (d)]. Wörter werden im Schriftbild durch Leerstellen isoliert.

Wortart: Wortarten sind die Klassen, in die die Wörter einer Sprache nach Form- und Bedeutungsmerkmalen eingeteilt werden können. Wir unterscheiden die gängigen Wortarten Nomen, Verb, Adjektiv, Artikel, Pronomen, Präposition, Adverb und Konjunktiv [Buß90].

Wortstamm: →Stamm.

D.4 Begriffe aus dem Bereich der Systemintegration

Kapsel: *Kapseln* (engl. Wrapper) sind Werkzeuge, die bekannte Quellen auf eine einem internen oder externen Standard konforme Art und Weise zugänglich machen und die Objekte dieser Quellen übertragen [AHK+95].

Mediator: Ein *Mediator* ist Teil der →Vermittlungsschicht. Er bietet eine ganze Reihe von wertsteigernden Diensten [Wie94, S. 2]: Den gebündelten Zugang zu einer Vielzahl von Quellen, auf die er über die →Kapseln zugreift, die Auswahl von relevantem Quellmaterial, die Auflösung von nicht zusammenpassenden Anwendungsbereichen, die Abstraktionfähigkeit, Materialien auf die richtige Granularitätsstufe zu bringen, die Integration von Material aus diversen Quelldomänen, die Beurteilung der Qualität des Materials von veschiedenen Quellen, die Weglassung von Replikaten oder bereits bekannten Informationen, das Aufspüren von Ausnahmen von erwarteten Werten oder Trends, die Umformung von Material, um die Darstellung für den Benutzer aufzubereiten, die Optimierung von Antwortzeiten, um geringe Kosten zu ermöglichen.

Vermittlungsschicht: In einer Integrationsarchitektur, ist die *Vermittlungsschicht* die Schicht oberhalb der Informationsquellen (Ressourcenschicht), in der die Integration der Informationsquellen vorgenommen wird. Auf der Vermittlungsschicht basiert schließlich die Benutzerschicht mit den Benutzeragenten.

Literaturverzeichnis

[AB97] Chris Addison und Peter Ballantyne. A 'Web Without Frontiers': Building European Partnerships on the Internet. *Quarterly Bulletin of IAALD* **42**(3-4), 1997, Seite 203–209. `http://www.ecdpm.org/pubs/ca_1.htm`.

[AF99] Bernd Amann und Irini Fundulaki. Integrating Ontologies and Thesauri to Build RDF Schemas. In *Research and Advanced Technology for Digital Libraries, Third European Conference, ECDL'99*. Springer, 1999. `ftp://sikkim.cnam.fr/pub/Reports/AMANN.html`.

[AHK+95] Yigal Arens, Richard Hull, Roger King, Michael Siegel, Hector Garcia-Molina, Michael Genesereth, Art Goldschmidt, Larry Kerschberg, Narinder Singh und Craig Thompson. Reference architecture for the Intelligent Integration of Information. Technischer Bericht, Program on Intelligent Integration of Information, ARPA, August 1995. Version 2.0 (Draft).

[Ait81] J. Aitchison. Integration of Thesauri in the Social Science. *International Classification* **8**(2), 1981, Seite 75–85.

[AK97] Naveen Ashish und Craig Knoblock. Semi-automatic Wrapper Generation for Internet Information Sources. In *CoopIS-97*, 1997.

[Alt02] AltaVista, Inc. AltaVista, 2002. http://www.altavista.com/.

[Amb92] Amba, Sanjeevi. *Computer-Based Linking of Thesauri*. Dissertation, The University of Alabama, 1992.

[AN72] H.A. Simon A. Newell. *Human Problem Solving*. Prentice Hall. 1972.

[ANOT96] S. Amba, N. Narasimhamurthi, K.C. O'Kane und P.M. Turner. Automatic Linking of Thesauri. In Hans-Peter Frei, Donna Harmann, Peter Schäuble und Ross Wilkinson (Hrsg.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM SIGIR, ACM Press, 1996, Seite 181–186.

[AP94] A. Aamodt und E. Plaza. Case-based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AICom- Artificial Intelligence Communications, IO Press* **7**(1), 1994, Seite 39–59.

[Bat94] W.-D. Batschi. Environmental Thesaurus and Classification of the Umweltbundesamt (Federal Environmental Agency), Berlin. In P. Stancikova und I. Dahlberg (Hrsg.), *Environmental Knowledge Organization and Information Management*, Seite 57–62. INDEKS Verlag, Frankfurt/Main, 1994.

[BK99] A. Borgida und R. Küsters. What's not in a name? Initial Explorations of a Structural Approach to Integrating Large Concept Knowledge-Bases. Technischer Bericht DCS-TR-391, Rutgers University, USA, 1999.

[BKK+98] D.C. Berrios, A. Kehler, D.K. Kim, V.L. Yu und L.M. Fagan. Automated Text Markup for Information Retrieval from an Electronic Textbook of Infectious Disease. In *Proc. of Amia Symp 1998*, 1998, Seite 975–986.

[BLN86] C. Batini, M. Lenzerini und S.B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys* **18**(4), 1986, Seite 323–364.

[BMR+96] Frank Buschmann, Regine Meunieur, Hans Rohnert, Peter Sommerlad und Michael Stal. *Pattern–oriented Software Architecture: A System of Patterns.* John Wiley and Sons. 1996.

[BMS98] C. Buckley, M. Mandra und A. Singhal. Improving Automatic Query Expansion. In *21rst ACM SIGIR International Conference on Research and Development in Information Retrieval*, 1998, Seite 206–214.

[Bor79] J. Bortz. *Lehrbuch der Statistik.* Springer. 1979.

[Bor99] Andreas Born. *Blackboardarchitekturen zur Entwicklung intelligenter tutorieller Systeme - Ein Prototyp am Beispiel der linearen Optimierung.* Dissertation, Philosophisch–Naturwissenschaftliche Fakultät, Universität Basel, Schweiz, 1999.

[BS85] R. J. Brachmann und J. G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science* **9**(2), 1985, Seite 171–216.

[Bußß90] Hadumod Bußmann. *Lexikon der Sprachwissenschaft.* Alfred Kröner Verlag. 1990.

[BYRN99] Ricardo Baeza-Yates und Berthier Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley. 1999.

[BZW98] Walter Brenner, Rüdiger Zarnekow und Hartmut Wittig. *Intelligente Softwareagenten: Grundlagen und Anwendungen.* Springer–Verlag, Berlin, Heidelberg. 1998.

[CAB99] CAB International. CAB Thesaurus, 1999. `http://www.cabi.org/catalog/dbmanual/thesaur.htm`.

[CCWB94] P.R. Cohen, A. Cheyer, M. Wang und S.C. Baeg. An open agent architecture. In O. Etzioni (Hrsg.), *Proc. of the AAAI Spring Symposium*, 1994, Seite 1–8. Stanford.

[Cen00] Center for Earth Obserevation. INFEO homepage, 2000. `http://infeo.ceo.org/`.

[CFF+98a] Vinay K. Chaudhri, Adam Farquhar, Richard Fikes, Peter D. Karp und James Rice. OKBC: A Programmatic Foundation for Knowledge Base Interoperability. In *AAAI/IAAI*, 1998, Seite 600–607.

[CFF+98b] Vinay K. Chaudhri, Adam Farquhar, Richard Fikes, Peter D. Karp und James Rice. Open Knowledge Base Connectivity 2.0.3, 1998.

[CG95] Cogis, O. und Guinaldo, O. A Linear Descriptor for Conceptual Graphs and a Class for Polynomial Isomorphism Test. In *Conceptual Structures: Applications, Implementation and Theory (LNAI 954)*. Springer-Verlag, 1995, Seite 263–291.

[Cha00] H. Chalupsky. OntoMorph: A translation system for symbolic knowledge. In *Proc. of KR 2000*, 2000, Seite 471–482.

[Cim98] J.J. Cimino. Desiderata for Controlled Medical Vocabularies. *Methods of Information in Medicine* **37**(4-5), 1998, Seite 394–403.

[CL92] Hsinchun Chen und Kevin J. Lynch. Automatic Construction of Networks of Concepts Characterizing Document Databases. *IEEE Transactions on Systems, Man, And Cybernetics* **22**(5), 1992, Seite 885ff.

[CLBD93] H. Chen, K. Lynch, K. Basu und T. Dorbin. Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval. *IEEE Expert*, April 1993, Seite 25–34.

[CLvRC98] F. Crestani, M. Lalmas, C.J. van Rijsbergen und I. Campbell. Is this Document Relevant? ... Probably - A Survey of Probabilistic Models in Information Retrieval. *ACM Computing Surveys* **30**(4), Dezember 1998, Seite 528–552.

[CM63] Cyril W. Cleverdon und J. Mills. The testing of indexing language devices. In *Aslib Proceedings*, Band 30, 1963, Seite 172–181.

[CNR97] CNR, Rome, Umweltbundesamt, Berlin. GEMET – General European Multilingual Environment Thesaurus. Technischer Bericht, European Environment Agency, September 1997. Version 1.0.

[Coe94] M. H. Coen. SodaBot: A software agent environment and construction system. Technischer Bericht A.I. Technical Report 1493, MIT, Artificial Intelligence Laboratory, 1994.

[Coh99] W. Cohen. Some practical observations from integration of Web integration. In *Workshop The Web and Databases (WebDB)*, 1999. `http://www-rocq.inria.fr/~cluet/WEBDB/procwebdb99.html`.

[Con97] Stefan Conrad. *Föderierte Datenbanksysteme – Konzepte der Datenintegration.* Springer. 1997.

[Con02] World Wide Web Consortium. The World Wide Web Consortium, 2002. `http://www.w3.org/`.

[Cor89] Daniel D. Corkill. *Design Alternatives for Parallel and Distributed Blackboard Systems*, Kapitel 6, Seite 99–136. Perspectives in Artificial Intelligence. Academic Press, Inc., Boston, San Diego, New York, Berkley, London, Sydney, Tokyo, Toronto. 1989.

[Cov01] Robin Cover. The XML Cover Pages: XML and Compression. Technischer Bericht, Organization for the Advancement of Structured Information Standards (OASIS), 2001. `http://www.oasis-open.org/cover/xmlAndCompression.html`.

[Cra95] Iain Craig. *Blackboard Systems.* Ablex Publishing Corporation, Norwood, New Jersey. 1995.

[CYFS95] Hsinchun Chen, Tak Yim, David Fye und Bruce R. Schatz. Automatic Thesaurus Generation for an Electronic Community System. *Journal of the American Society for Information Science (JASIS)* **46**(3), 1995, Seite 175–193.

[Dat95] C. J. Date. *An Introduction to Database Systems.* Addison-Wesley Systems Programming Series. Addison-Wesley Pub. Co., Inc., Reading, MA. 1995.

[Dav86] R. Davis. Knowledge-Based Systems. *Science* Band 231, Februar 1986, Seite 957–963.

[Der99] Michael L. Dertouzos. *What will be: die Zukunft des Informationszeitalters.* Springer-Verlag. 1999.

[DH90] Werner Digel und Gerhard Kwiatkoski (Hrsg.). *Meyers Großes Taschenlexikon in 24 Bänden, Band 16.* BI-Taschenbuchverlag, Mannheim, Wien, Zürich. 1990.

[DIN87] DIN 1463, Teil 1. DIN 1463: Richtlinien für die Erstellung und Weiterentwicklung von Thesauri, Teil 1: Einsprachige Thesauri. Technischer Bericht, Deutsches Institut für Normung, Berlin, 1987.

[DIN92] DIN 2342. DIN 2342, Teil 1: Begriffe der Terminologielehre: Grundbegriffe. Technischer Bericht, Deutsches Institut für Normung, Berlin, 1992.

[DIN93a] DIN 1463, Teil 2. DIN 1463: Erstellung und Weiterentwicklung von Thesauri, Teil 2: Mehrsprachige Thesauri. Technischer Bericht, Deutsches Institut für Normung, Berlin, 1993.

[DIN93b] DIN 2330. DIN 2330: Begriffe und Benennungen: Allgemeine Grundsätze. Technischer Bericht, Deutsches Institut für Normung, Berlin, 1993.

[DPA00] A.H. Doan, Domingos P. und Levy A.Y. Learning Source Description for Data Integration. In *Workshop The Web and Databases (WebDB)*, 2000. `http://www.research.att.com/conf/webdb2000/program.html`.

[Eft94] Efthimis N. Efthimiadis. End-users' Understanding of Thesaural Knowledge Structures and Interactive Query Expansion. In Hanne Albrechtsen und Susanne Oernager (Hrsg.), *Knowledge Organization and Quality Management.* International Society for Knowledge Organization, 1994.

[Eft96] Efthimis N. Efthimiadis. Query Expansion. In Martha E. Williams (Hrsg.), *Annual Review of Information Systems and Technology (ARIST)*, Band 31. 1996.

[EHRLR88] L.D. Ermann, F. Hayes-Roth, V.R. Lesser und D.R. Reddy. The HERSAY-II speech-understanding system: Integrating knowledge to resolve uncertainty. In Engelmore und Morgan [EM88], Seite 29–86.

[EM88] R.S. Engelmore und A.J. Morgan (Hrsg.). *Blackboard Systems.* Addison-Wesley. 1988.

[equ01] equero future net technologies AG. equero2001 homepage. `http://www.equero.de/`, Juni 2001.

[ETC98] ETC/CDS. Catalogue of Data Sources (CDS), 1998. `http://www.mu.uni-hannover.de/cds/`.

[Fel98] Christiane Fellbaum (Hrsg.). *WordNet: an electronical lexical database*. MIT Press. 1998.

[FFW91] R.H. Fowler, W.A.L. Fowler und B.A. Wilson. Integrating Query, Thesaurus, and Documents through a Common Visual Representation. In *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Interfaces, 1991, Seite 142–151.

[FG96] S. Franklin und A. Graesser. Is it an agent, or just a program? A taxonomy for autonomous agents. In *Proc. of Third International Workshop on Agent Theories, Architectures, and Languages*. Springer-Verlag, 1996.

[Fid91] Raya Fidel. Searchers' Selection of Search Keys: II. Controled Vocabulary or Free-Text Searching. *Journal of The American Society for Information Science* **42**(7), 1991, Seite 501–514.

[Fis00] Florian Fischer. Visualisierung von Thesauri mittels Fischaugensichten. Studienarbeit, Universität Karlsruhe, Karlsruhe, April 2000. Advisors: Peter C. Lockemann, R. Nikolai.

[FIZ00] FIZ Karlsruhe. FIZ Karlsruhe and STN International, 2000. `http://www.fiz-karlsruhe.de/`.

[FJ00] Bruna Felluga und Stefan Jensen, 2000. Persönliches Gespräch am 18.01.2000 auf dem Open Forum on Metadata Registries in Santa Fe, New Mexico, USA.

[FP93] C. Francalanci und B. Pernici. View Integration: a survey of current developments. Technischer Bericht, Politecnico Milano, 1993. Internal Report 93-053.

[Fri00] Stefan Fricke. *Werkzeuggestützte Entwicklung kooperativer Agenten im Dienstkontext*. Dissertation, TU-Berlin, 2000.

[FSW00] Mary Fernandez, Jérôme Siméon und Philip Wadler. XML Query Languages: Experiences and Exemplars. Technischer Bericht, Bell Labs, 2000. `http://www-db.research.bell-labs.com/user/simeon/xquery.html`.

[Gan90] Jochen Ganzmann. Criteria for the Evaluation of Thesaurus Software. *International Classification* **17**(3/4), 1990, Seite 148–157.

[Gau95] Wilhelm Gaus. *Dokumentations- und Ordnungslehre - Theorie und Praxis des Information Retrieval*. Springer, Berlin, Heidelberg, New York. 1995.

[GCM99] GCMD. GCMD Parameter Valids, 1999. `http://gcmd.nasa.gov/cgi-bin/md/valids_display.pl`.

[Geb00] Markus Gebhard. Fischaugen-basierter Browser für Thesaurusföderationen. Studienarbeit, Universität Karlsruhe, Karlsruhe, September 2000. Advisors: Peter C. Lockemann, Ralf Nikolai.

[GES01] GESIS. Datenbank FORIS, 2001. `http://www.gesis.org/Information/FORIS/index.htm/`.

[GK94] M. R. Genesereth und S. P. Ketchpel. Software agent. *Communications of the ACM* **37**(7), 1994, Seite 48–53.

[GKD97] Michael R. Genesereth, Arthur M. Keller und Oliver Duschka. Infomaster: An Information Integration System. In *Proceedings of 1997 ACM SIGMOD Conference*, Mai 1997.

[GLN92] W. Gotthard, P.C. Lockemann und A. Neufeld. System-Guided View Integration for Object-Oriented Databases. *IEEE Transactions on Knowledge and Data Engineering* **4**(1), 1992, Seite 1–22.

[Goo02] Google, Inc. Google, 2002. `http://www.google.com/`.

[GPG99] A. Gangemi, D.M. Pisanelli und G.Steve. An Overview of the ONIONS project: Applying Ontologies to the Integration of Medical Terminologies. *Data and Knowledge Engineering* Band 31, September 1999, Seite 183–220. `http://saussure.irmkant.rm.cnr.it/onto/dke/dke.pdf`.

[GPS98] Aldo Gangemi, M. Pisanelli, Domenico und Geri Steve. Ontology Integration: Experiences with Medical Terminologies. In N. Guarino (Hrsg.), *Formal Ontology in Information Systems*, IOS Press, 1998.

[Gru92] Thomas Gruber. A Translation Approach fo Portable Ontology Specifications. In *Proc. of the Second Japanese Knowledge Acquisition for Knowledge-Based Systems Workshop, Kobe, Japan*, 1992, Seite 199–221.

[GRW97] W. Geiger, M. Reissfelder und R. Weidemann. Distribution of expert information on contaminated soil via Internet, Intranet, CD-ROM and print media. In *Information and Communication in Environmental and Health Issues: Proc. of Eco-Informa '97*, Band 12, Neuherberg, Oktober 1997. Eco-Informa Press, Seite 72–77.

[Gut00] Michael Gutbier. Flexible Architektur zur Akquisition und Aktualisierung von Integrationswissen in Thesaurusföderationen. Diplomarbeit, Universität Karlsruhe, Karlsruhe, September 2000. Advisors: Peter C. Lockemann, Ralf Nikolai.

[Hab96] Corinna Habeck. Integration of Thesuari-Databases for WWW-based Dataretrieval. Diplomarbeit, Forschungszentrum Informatik (FZI), Karlsruhe, August 1996. Advisors: Peter C. Lockemann, R. Kramer, R. Nikolai, Y. Al-Salqan.

[Ham89] Rolf Hammerl. Untersuchung struktureller Eigenschaften in Begriffsnetzen. In Rolf Hammerl (Hrsg.), *Glottometrika 10*, Band 38 der *Quantitative Linguistics*. Studienverlag Dr. N. Brockmeyer, 1989.

[Hea92] Marti. A Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proc. of 14th International Conference on Computational Linguistics*, 1992.

[HK85] P. Harmon und D. King. *Expert Systems: Artificial Intelligence in Business*. Wiley, New York. 1985.

[HL77] Harry Halfar und Horst Langendörfer. Automatische Erstellung und Pflege eines Thesaurus für ein Dokumentationssystem. *Sprache und Datenverarbeitung*, 1977, Seite 265ff.

[Hol95] David Hollingsworth. The Workflow Reference Model. Technischer Bericht, Workflow Management Coalition (WfMC), 1995.

[HR95] B. Hayes-Roth. An architecture for adaptive intelligent systems. *Artificial Intelligence* (72), 1995, Seite 329–365.

[Int96] International Atomic Energy Agency (IAEA), European Commission (EC), World Health Organization (WHO). *International Conference: One Decade after Chernobyl*, Wien, April 1996. `http://www.iaea.or.at/worldatom/thisweek/preview/chernobyl/`.

[ISO90] ISO. Iso 1087: Terminology - Vocabulary. Technischer Bericht, International Standardisation Organisation (ISO), 1990.

[Jan99] Jan Jannink. Thesaurus Entry Extraction from an On-line Dictionary. In *Proc. of Fusion '99*, Sunnyvale CA, Juli 1999.

[JBS97] Stefan Jablonski, Markus Böhm und Wolfgang Schulze (Hrsg.). *Workflow Management – Entwicklung von Anwendungen und Systemen*. dpunkt.verlag. 1997.

[JC95] E.H. Johnson und P.A. Cochrane. A Hypertextual Interface for a Searcher's Thesaurus. In *Proceedings of the Digital Libraries Conference 1995*, 1995. http://www.csdl.tamu.edu/DL95/papers/johncoch/johncoch.html, 3.4.97.

[JMN+99] Jan Jannink, Prasenjit Mitra, Erich Neuhold, Srinivasan Pichai, Rudi Studer und Gio Wiederhold. An Algebra for Semantic Interoperation of Semistructured Data. In *Proc. of 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, Chicago, November 1999.

[JW99] Jan Jannink und Gio Wiederhold. Ontology Maintenance with an Algebraic Methodology: a Case Study. In *Proc. of 1999 AAAI workshop on Ontology Management*, Orlando FL, Juli 1999.

[JWZ95] Jiang, Tao, Wang, Lusheng und Zhang, Kaizhong. Alignment of trees - an alternative to tree edit. *Theoretical Computer Science* Band 143, 1995, Seite 137–148.

[KAM+01] Craig A. Knoblock, Jose Luis Ambite, Steven Minton, Cyrus Shahabi, Mohammad Kolahdouzan, Maria Muslea, Jean Oh und Snehal Thakkar. Integrating the World: The WorldInfo Assistant. In *International Conference on Artificial Intelligence (IC-AI)*, 2001. `http://www.isi.edu/cgi-bin/ariadne/info-agents/papers.cgi`.

[Ken99] Kennedy Institute of Ethics. Bioethics Thesaurus, 1999. `http://www.georgetown.edu/research/nrcbl/ir/thes99.htm`.

[Kin96] Hauke Kindler. *Eine Architektur für medizinische wissensbasierte Assistenzsysteme und ihre Realisierung für das akute Strahlensyndrom*, Band 455 der *Fortschritt–Berichte: Informatik/Kommunikation, Reihe 10*. VDI Verlag GmbH, Düsseldorf. 1996.

[Kir96] S. Kirn. Kooperativ–Intelligente Softwareagenten. *Information Management* **11**(1), 1996.

[KKN+96] Arne Koschel, Ralf Kramer, Ralf Nikolai, Wilhelm Hagg und Joachim Wiesel. A Federation Architecture for an Environmental Informaton System incorporating GIS, the World-Wide Web, and CORBA. In *Third International Conference/Workshop Integrating GIS and Environmental Modeling*, San-

ta Fe, New Mexico, USA, Januar 1996. National Center for Geographic Information and Analysis (NCGIA). http://ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/sf_papers/nikolai_ralf/fedarch.html.

[KKN+97] Arne Koschel, Ralf Kramer, Ralf Nikolai, Gergely Lukacs und Thomas Heinemeier. Data and Metadata Management in Distributed Environmental Information Systems. In Ralf Denzer, David A. Swayne und Gerald Schimak (Hrsg.), *Environmental Software Systems, Volume 2*, Band 2, Whistler, Canada, April 1997. Chapman and Hall, Seite 144–151.

[KNK+97] Ralf Kramer, Ralf Nikolai, Arne Koschel, Claudia Rolker, Peter Lockemann, Andree Keitel, Rudolf Legat und Konrad Zirm. WWW-UDK: A Web-based Environmental Metainformation System. *ACM SIGMOD Record* **26**(1), März 1997, Seite http://www.cs.umd.edu/areas/db/record/issues/9703/index.html.

[KNR+97] Ralf Kramer, Ralf Nikolai, Claudia Rolker, Sigfus Bjarnason und Stefan Jensen. Interoperability Issues of the European Catalogue of Data Sources (CDS). In *Proceedings of the Second IEEE Metadata Conference*, Silver Spring, Maryland, USA, September 1997. `http://www.llnl.gov/liv_comp/metadata/md97.html`.

[Koc74] Manfred Kochen. *Integrative Mechanisms in Literature Growth*, Band 9 der *Contributions in Librarianship and Information Science*, Kapitel A Cost-Effectiveness Analysis of See-Reference Structure. Greenwood Press. 1974.

[Kos99] Arne Koschel. *Ereignisgetriebene CORBA-Dienste für heterogene, verteilte Informationssysteme*. Dissertation, Universität Karlsruhe, Germany, July 1999.

[KR98] R. Kramer und C. Rolker. European Environmental Information Services (EEIS): Facilitating Common Catalogue Access for EEA and CEO Users. 5th European Earth Observation Strategy (EEOS) Worshop, Oktober 1998.

[Kri93] J. Kristensen. Expanding end-user's query statements for free text searching with a search-aid thesaurus. *Information Processing and Management* **29**(6), 1993, Seite 733–744.

[KT68] Manfred Kochen und Renata Tagliacozzo. A Studie of Cross-Referencing. *Journal of Documentation* **24**(3), September 1968.

[KWD97] N. Kushmerick, D. Weld und B. Doorenbos. Wrapper Induction for Information Extraction. In *IJCAI-97*, 1997.

[Lan72] F.W. Lancaster. *Vocabulary Control for Information Retrieval*, Kapitel Some Cost-Effectiveness Aspects of Vocabulary Control. Washington D.C. 1972.

[Lan77] F.W. Lancaster. *Evaluation and Scientific Management of Libraries and Information Centres*. C.W. Cleverdon. 1977.

[Lan86] F.W. Lancaster. *Vocabulary Control for Information Retrieval*, Kapitel Evaluation of Thesauri, Seite 155–157. Information Ressources Press. 1986.

[LE88] V.R. Lesser und L.D. Ermann. A retrospective view of the HEARSAY-II architecture. In Engelmore und Morgan [EM88], Seite 87–111.

[Lin02] Alexander Linden. Vieles läßt sich auch ohne Semantic Web lösen. *Computer Zeitung* (20), Mai 2002, Seite 15. Alexander Linden ist Research Director der Gartner Group für Trends und Technologien.

[LKK+97] Peter C. Lockemann, Ulrike Kölsch, Arne Koschel, Ralf Kramer, Ralf Nikolai, Mechtild Wallrath und Hans-Dirk Walter. The Network as a Global Database: Challenges of Interoperability, Proactivity, Interactiveness, Legacy. In M. Jarke, M. Carey, K.R. Dittrich, F. Lochovsky, P. Loucopoulos und M.A. Jeusfeld (Hrsg.), *Proceedings of the Twenty-third International Conference on Very Large Data Bases*, Athens, Greece, August 1997. Seite 567–574.

[LSM95] X. Li, S. Szpakowicz und S. Matwin. A WordNet-based algorithm for word sense disambiguation, 1995.

[MBF90] George A. Miller, R. Beckwith und C. Fellbaum. Introduction to WordNet: An On-Line Lexical Database. *Journal of Lexicography* **3**(5), 1990, Seite 235–244.

[McG91] R. McGregor. Inside the LOOM classifier. *SIGART Bulletin* **2**(3), 1991, Seite 88–92.

[Men98] E. Mena. *OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies.* Dissertation, Universidad de Zaragoz, 1998. `http://siul02.si.ehu.es/~jirgbdat/PUBLICATIONS/thesis.ps.gz`.

[MFRW00] D.L. McGuiness, R. Fikes, J. Rice und S. Wilder. An environment for merging and testing large ontologies. In *Proc. of KR 2000*, 2000, Seite 483–493.

[MIKS00] E. Mena, A. Illarramendi, V. Kashyap und A. Sheth. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. *Distributed and Parallel Databases (DAPD)* **8**(2), April 2000. `http://siul02.si.ehu.es/~jirgbdat/PUBLICATIONS/dapd00.ps.gz`.

[Mil91] Jessica L. Milstead. Specifications for Thesaurus Software. *Information Processing and Management* **27**(2/3), 1991, Seite 165–175.

[Mil92] Jessica Milstead. Methodologies for Subject Analysis in Bibliographic Databases. *Information Processing and Management* **28**(3), 1992, Seite 407–431.

[Mil97] Uri Miller. Thesaurus Construction: Problems and their Roots. *Information Processing and Management* **33**(4), 1997, Seite 481–493.

[Mil98] G.A. Miller. WordNet: A lexical database for English. *Communications of the ACM* **38**(11), 1998, Seite 39–41.

[Min85] M. Minsky. *Society of mind.* New York. 1985.

[Mor98] Thilo Morgenstern. Halbautomatische Erzeugung von Inter-Thesaurus-Beziehungen. Diplomarbeit, Forschungszentrum Informatik (FZI), Karlsruhe, Dezember 1998. Advisors: Peter C. Lockemann, R. Nikolai.

[MR88] Hafedh Mili und Roy Rada. Merging Thesauri: Principles and Evaluation. *IEEE Transactions on pattern analysis and machine intelligence* **10**(2), March 1988.

[MWJ99] Prasenjit Mitra, Gio Wiederhold und Jan Jannink. Semi-automatic Integration of Knowledge Sources. In *Proc. of Fusion '99*, Sunnyvale CA, Juli 1999.

[MWK99] Prasenjit Mitra, Gio Wiederhold und Martin Kersten. A Graph-Oriented Model for Articulation of Ontology Interdependencies. Technischer Bericht CSL-TN-99-411, Stanford University, August 1999. `http://www-db.stanford.edu/SKC/publications.html`.

[NAS98] NASA. NASA Thesaurus, 1998. `http://www.sti.nasa.gov/thesfrm1.htm`.

[Nat98] M. Natlacen. AGRIS: Guide to Indexing. Technischer Bericht, FAO - Food and Agriculture Organization of the UN, 1998. ftp://ftp.iaea.org/dist/agris/outgoing/indguide.

[ND98] Ralf Nikolai und Henning Dudat. A SGML Parser for Off-Line Transfer of Thesaurus Data. Technischer Bericht, FZI Forschungszentrum Informatik, September 1998.

[Nev70] H. H. Neville. Feasibility study of a scheme for reconciling thesauri covering a common subject. *Journal of Documentation* **26**(4), 1970, Seite 313–336.

[New62] A. Newell. Some problems of basic organisation in problem solving programs. In M.C. Yovits, G.T. Jacobi und G.D. Goldstein (Hrsg.), *Conf. of Self-Organizing Systems*. Spartan Books, 1962, Seite 393–423.

[NK99] Ralf Nikolai und Ralf Kramer. Technische und semantische Aspekte der losen Integration heterogener und autonomer Thesauri. In *Heterogene, aktive Umweltdatenbanken*. Metropolis, Marburg, 1999, Seite 181–210.

[NKK+99] Ralf Nikolai, Wassili Kazakos, R. Kramer, Sven Behrens, Walter Swoboda und Fred Kruse. WWW-UDK 4.0: Die neue Generation eines Web-Portals zu deutschen und österreichischen Umweltdaten. In C. Rautenstrauch und M. Schenk (Hrsg.), *Umweltinformatik 99: Umweltinformatik zwischen Theorie und Industrienwendung, 13. Internationales Symposium Informatik für den Umweltschutz, Magdeburg*, Nr. 23 der Umwelt-Informatik Aktuell. Metropolis, 1999, Seite 347–361.

[NKS+99] Ralf Nikolai, Ralf Kramer, Marc Steinhaus, Bruno Felluga und Paolo Plini. GenThes: A General Thesaurus Browser for Web-based Catalogue Systems. In *Proc. IEEE Meta-Data'99*, Bethesda, Maryland, USA, April 1999.

[NM85] Robert Niehoff und Greg Mack. The vocabulary switching system. Description of Evaluation Studies. *Internation Classification* **12**(1), 1985, Seite 2–6.

[NM00] Natalya Fridman Noy und Mark A. Musen. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, 2000.

[NM02] Natalya Fridman Noy und Mark A. Musen. PromptDiff: A Fixed-Point Algorithm for Comparing Ontology Versions. In *Proc. of the Eighteenth National Conference on Artificial Intelligence (AAAI-02)*, 2002.

[NTK98] Ralf Nikolai, Andreas Traupe und Ralf Kramer. Thesaurus Federations: A Framework for the Flexible Integration of Heterogeneous, Autonomous Thesauri. In *Proc. Conference on Research and Technology Advances in Digital Libraries (ADL'98)*, Santa Barbara, USA, April 1998. Seite 46–55.

[PAG99] D.M. Pisanelli, A.Gangemi und G.Steve. A Medical Ontology Library that Integrates the UMLS Metathesaurus. In *Proc. of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM 99)*, 1999. `http://saussure.irmkant.rm.cnr.it/onto/aime99/aime99.pdf`.

[Pai91] C. Paice. A Thesaural Model of Information Retrieval. *Information Processing and Management* **27**(5), 1991, Seite 433–447.

[PB98] Lawrence Page und Sergey Brin. The anatomy of a large-scale hypertextual search engine. In *Proc. of the 7th Annual World Wide Web Conference*, 1998.

[PCN00] Martin Purvi, Stephen Cranefield und Mariusz Nowostawski. A Distributed Architecture for Environmental Information Systems. In David A. Swayne, Ralf Denzer und Gerald Schimak (Hrsg.), *Environmental Software Systems, Volume 5*, Zell am See, Austria, 2000.

[PE94] Steven A. Pollitt und Geoffrey P. Ellis. Improving Search Quality using Thesauri for Query Specification and the Presentation of Search Results. In Hanne Albrechtsen und Susanne Oernager (Hrsg.), *Knowledge Organization and Quality Management*. International Society for Knowledge Organization, 1994.

[Pol93] Richard Pollard. A Hypertext-Based Thesaurus as a Subject Browsing Aid for Bibliographic Databases. *Information Processing & Management* **29**(03), 1993, Seite 345–358.

[Por80] M.F. Porter. An Algorithm for Suffix Stripping. *Program* **14**(3), 1980, Seite 130–137.

[PP69] A.W. Pratt und M. Pacak. Identification and transformation of terminal morphems in medical english. In *Methodik der Information in der Medizin*, Nr. 8, 1969.

[Pra93] L. Y. Pratt. Discriminability-Based Transfer between Neural Networks. In Stephen José Hanson, Jack D. Cowan und C. Lee Giles (Hrsg.), *Advances in Neural Information Processing Systems*, Band 5. Morgan Kaufmann, San Mateo, CA, 1993, Seite 204–211.

[PSBT96] A.S. Pollitt, M.P. Smith, P. A. J. Braekevelt und M. Treglown. HIBROWSE for EMBRASE - the application of view-based searching techniques to Europe's most important biomedical data base. In *presented at INFO 96 - Annual Meeting, Tel Aviv*, 1996.

[Pup88] Frank Puppe. *Einführung in Expertensysteme*. Studienreihe Informatik. Springer, Berlin. 1988.

[Qui68] M. Quillian. *Semantic Information Processing (M. Minsky)*, Kapitel Semantic memory, Seite 216–270. MIT Press, Cambridge, Mass. 1968.

[Rad87] Roy Rada. Connecting and Evaluating Thesauri: Issues and Cases. *International Classification* **14**(2), 1987, Seite 63–69.

[Rad90] Roy Rada. Maintaining Thesauri and Metathesauri. *International Classification* **17**(3/4), 1990, Seite 158–164. Presented at the thesaurus software seminar, Darmstadt, 1990.

[Rec99] Alan L. Rector. Clinical Terminology: Why is it so hard? Technischer Bericht, Medical Informatics Group, Department of Computer Science, University of Manchester, Mai 1999.

[Rie99] Birgitta König Ries. *Ein Verfahren zur semi-automatischen Generierung von Mediatorspezifikationen.* Dissertation, Universität Karlsruhe, 1999.

[RKK99] Claudia Rolker, Ralf Kramer und Wassili Kazakos. Interoperability among Earth Observation and General Environmental Data Catalogues via CIP. In *Proceedings of the Earth Observation and Geo-Spatial Web and Internet Workshop '99*, Washington D.C., USA, Februar 1999.

[RM87] Roy Rada und Brian K. Martin. Augmenting Thesauri for Information Systems. *ACM Transactions on Office Information Systems* Band 5, 1987, Seite 378–392.

[RMC97] A. Rossi Mori und F. Consorti. Exploiting terminological standards from CENTC251 to support interoperability of health record systems. Band 48, 1997, Seite 111–124.

[RMCEM97] A. Rossi Mori, F. Consorti, Galeazzi E. und P. Merialdo. A second generation of terminological systems is coming. In *Medical Informatics Europe 97*, 1997, Seite 111–124.

[Rol01] Claudia Rolker. *Ein iteratives Information Retrieval Verfahren mit automatischerSuchmechanismenauswahl.* Dissertation, Universität Fridericiana zu Karlsruhe (TH), 2001.

[Rou92] Corentin Roulin. Sub-thesauri as Part of a Metathesaurus. In *International Study Conference on Classification Research (5th: Toronto, 1991), Classification research for knowledge representation and organization.* Elsevier, 1992, Seite 329–336.

[RPR+98] J.E. Rogers, C. Price, A.L. Rector, W.D. Solomon und N. Smejko. Validating Clinical Terminology Structures: Integration and Cross-Validation of Read Thesaurus and GALEN. In *Annual Fall Symposium of American Medical Informatics Association*, 1998, Seite 845–849.

[RRMCZ98] A.L. Rector, A. Rossi Mori, F. Consorti und P. Zanstra. Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN Organisation. *International Journal on Medical Informatics* **48**(1–3), Februar 1998, Seite 71–84.

[Rug92] Gerda Ruge. Experiments on Linguistically-Based Terms Association. *Information Processing & Management* **28**(3), 1992, Seite 317ff.

[RWP98] David Robinson, Karen Wanger und Colin Price. The Clinical Terms and ICPC: Identifying Equivalence and Enabling Compatibility for „Non-medical" Terms. In *Proc. of Primary Health Care Specialist Group Annual Conference*, Cambridge, 1998. `http://www.schin.ncl.ac.uk/phcsg/conferences/cambridge1998/robinson.htm%`.

[Sal73] G. Salton. Experiments in Multi-Lingual Information Retrieval. *Information Processing Letters* **2**(1), 1973, Seite 6–11. TR 72-154 at `http://cs-tr.cs.cornell.edu,16.4.97`.

[Sal87] G. Salten. *Information Retrieval - Grundlegendes für Informationswissenschaftler.* McGraw-Hill, Hamburg, New York u.a. 1987.

[SB85] E.H. Shortliffe und B.G. Buchanan. A Model for Inexact Reasoning in Medicine. In *Rule-Based Expert Systems*, Seite 233–262. Addison-Wesley,, 1985.

[SB92] Charles A. Sneiderman und Ellen J. Bicknell. Computer-Assisted Dynamic Integration of Multiple Medical Thesauruses. *Comp. Biol. Med.* **22**(1/2), 1992, Seite 135–145.

[SC96] G. Spanoudakis und P. Constantopoulos. Elaborating Analogies from Conceptual Models. *International Journal of Intelligent Systems* **11**(11), 1996, Seite 917–974.

[SC97] Marios Sintichakis und Panos Constantopoulos. A Method for Monolingual Thesauri Merging. In Nicholas J. Belkin, A. Desai Narasimhalu und Peter Willet (Hrsg.), *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM SIGIR, ACM press, 1997, Seite 129–138.

[Sch88] Peter Schöndorf. Nicht-Konventionelle Thesaurusrelationen als Orientierungshilfen für Indexierung und Recherche - Analyse ausgewählter Beispiele. *Nachrichten für Dokumentation* (39), 1988, Seite 231–244.

[Sch93] T.J. Schult. Tafelgeschäfte - Blackboard-Systeme meistern Komplexität. *ct* (3), 1993, Seite 46–50.

[Sch94] Rene Schönfeldt. Mathematische Eigenschaften für Thesaurusrelationen. *Nachrichten für Dokumentation* (54), 1994, Seite 203–212.

[Sch99a] Dan Schiller. *Digital Capitalism: Networking the Global Market System.* MIT Press. 1999.

[Sch99b] Andreas Schmeiler. Mehrsprachiges Information Retrieval innerhalb der verteilten Dokumentensammlung des European Environment Information and Observation Network. Studienarbeit, Forschungszentrum Informatik (FZI), Karlsruhe, Februar 1999. Advisors: Peter C. Lockemann, R. Nikolai.

[Sea69] J. R. Searle. *Speech Acts.* Cambridge University Press. 1969.

[SFDB99] Rudi Studer, Dieter Fensel, Stefan Decker und V. Richard Benjamins. Knowledge Engineering: Survey and Future Directions. In Frank Puppe (Hrsg.), *XPS-99: Knowledge Based Systems. Survey and Future Directions. Proceeding of the 5th German Conf. on Knowledge-based Systems*, Lecture Notes in Artificial Intelligence. Springer, 1999, Seite 1–23.

[Sim77] H.A. Simon. Scientific discovery and the psychology of problem solving. In *Models of Discovery.* D. Reidel Publishing Company, Boston, 1977.

[SL90] A. P. Sheth und J. A. Larson. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys* **22**(3), March 1990, Seite 183–236.

[Sly76] George Van Slype. Definition of Thesauri Essential Characteristics. Technischer Bericht 2, Commission of the European Communities, July 1976.

[Sly77] George Van Slype. Qualitative und quantitative Merkmale ein- und mehrsprachiger Thesauri. In *Die Überwindung der Sprachbarrieren: Dritter Europäischer Kongress über Dokumentationssysteme und -netze (Third European Congress on Information Systems and Networks Overcoming the Language Barrier)*, Band 1. Verlag Dokumentation München, 3-6 Mai 1977.

[SM01] G. Stumme und A. Maedche. FCA-Merge: A Bottom-Up Approach for Merging Ontologies. In *Proc. of the 17th International Joint Conference on Artificial Intelligence JCAI*, 2001.

[Sow84] John F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley. 1984.

[Sow00] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co. 2000.

[Squ93] Steven J. Squires. Access to Biomedical Information: The Unified Medical Language System. *Library Trends* **42**(1), 1993, Seite 127–151.

[SR90] A. Stern und N. Rischette. On the Construction of a Super Thesaurus Based on Existing Thesauri. In *Tools for Knowledge Organization and the Human Interface*, Band 2, Seite 133–144. Indeks Verlag, Frankfurt/Main, 1990.

[Sta00] Stanford University Database Group. Scalable Knowledge Composition (SKC), zuletzt besucht: 08.06.2000. `http://www-db.stanford.edu/SKC/`.

[Str01] Udo Straub. Konzeption einer informationstechnischen Unterstützung von Geschäftsprozessen in Virtuellen Unternehmen. Diplomarbeit, Universität Karlsruhe, Karlsruhe, Juli 2001. Advisors: Wolfried Stucky, Peter Weiß, Ralf Nikolai.

[Sun01] Hua Sun. Entwurf und Realisierung von Integrationsagenten für Thesaurusföderationen. Studienarbeit, Universität Karlsruhe, Karlsruhe, Februar 2001. Advisors: Peter C. Lockemann, R. Nikolai.

[The01] The J. Paul Getty Trust. Getty Research Library Catalog, 2001. `http://library.getty.edu/`.

[Tol00] Robert Tolksdorf. Coordination Technology for Workflows on the Web: Workspaces. In *Proc. of the Fourth International Conference on Coordination Models and Languages COORDINATION 2000*. Springer-Verlag, 2000.

[Tra97] Andreas Traupe. Flexible, multilinguale Thesaurusföderation im Internet. Diplomarbeit, Forschungszentrum Informatik (FZI), Karlsruhe, August 1997. Advisors: Peter C. Lockemann, R. Nikolai.

[Tve77] A. Tversky. Features of Similarity. *Psychological Review* **84**(4), 1977.

[Vas99] Nerantzoula Vassiliadis. Bewertung der Qualität von Föderationen. Diplomarbeit, Forschungszentrum Informatik (FZI), Karlsruhe, Juli 1999. Advisors: Peter C. Lockemann, Ralf Nikolai.

[Vie97] Johannes Viegener. *Inkrementelle, domänenunabhängige Thesauruserstellung in dokumentbasierten Informationssystemen durch Kombination von Konstruktionsverfahren*. Dissertation, Universität Karlsruhe, 1997. Infix.

[VJBCS98] Pepijn R.S. Visser, Dean M. Jones, T.J.M Bench-Capon und M.J.R Shave. Assessing Heterogeneity by Classifying Ontology Mismatches. In N. Guarino (Hrsg.), *Formal Ontology in Information Systems*. IOS Press, 1998, Seite 148–162.

[Wei01] Tim Young Weißschädel. Entwurf und Realisierung eines Benutzeragenten zur Einbringung des Expertenwissens in de Thesaurus-Integrationsprozeß. Studienarbeit, Universität Karlsruhe, Karlsruhe, Juli 2001. Advisors: Peter C. Lockemann, R. Nikolai.

[Wer85] Gernot Wersig. *Thesaurus-Leitfaden*. K.G.Saur Verlag KG, München. 1985.

[Wie94] Gio Wiederhold. Interoperation, Mediation and Ontologies. In *Proceedings of the International Symposium on Fifth Generation Computer Systems (FGCS94), Workshop on Heterogenous Cooperative Knowledge-Bases (ICOT), Tokyo, Japan, Dec. 1994*, Band W3, http://db.stanford.edu/pub/gio/1994/medont.ps, 1994. Seite 33–48.

[Wie96] Gio Wiederhold. *Intelligent Integration of Information*. Kluwer Academic. Mai 1996.

[Wil97] Wille, Rudolf. Conceptual Graphs and Formal Concept Analysis. In *Conceptual Structures: Fullfilling Peirces Dream*, LNAI, Seattle, Washington, USA, August 1997. Springer-Verlag, Seite 290–303.

[Wil98] Wille, Rudolf. Triadic Concept Graphs. In *Conceptual Structures: Theory, Tools and Applications*, LNAI, Montpellier, Frankreich, August 1998. Springer-Verlag, Seite 194–221.

[WJ95] M. Wooldridge und N. R. Jennings. Intelligent agents: theory and practice. *The Knowledge Engineering Review* **10**(2), 1995, Seite 115–152.

[WK96] G. Widmer und M. Kubat. Learning in the presence of context drift and hidden contexts, 1996.

[Wor98] World Wide Web Consortium. Extensible Markup Language (XML) 1.0. W3c recommendation, W3C, 1998. `http://www.w3.org/TR/1998/REC-xml-19980210`.

[Wor99] Work Group 1. Interface 1: Process Definition Interchange - Process Model. Technischer Bericht, Workflow Management Coalition (WfMC), 1999. Version 1.1.

[Wor00a] Workflow Management Coalition. Homepage of the Workflow Management Coalition, 2000. `http://www.aiim.org/wfmc/`.

[Wor00b] World Wide Web Consortium. Simple Object Access Protocol (SOAP) 1.1. W3c note, W3C, 2000. `http://www.w3.org/TR/SOAP/`.

[Wor01] World Wide Web Consortium. XQuery 1.0: An XML Query Language. W3c working draft, W3C, 2001. `http://www.w3.org/TR/xquery/`.

[WSN98] C.H. Wu, S. Srinivasan und S.J. Nelson. An Experience in Merging Thesauri: Using Terms From Other Thesauri to Enhance MeSH. In *American Medical Informatics Association Annual Symposium (AMIA)*, 1998.

[WV99] M.J. Wooldridge und M. Veloso (Hrsg.). *Artificial Intelligence Today, Recent Trends and Developments*, Kapitel Logic-Based Knowledge Representation (F. Baader), Seite 13–41. number 1600 in Lecture Notes in Computer Science. Springer. 1999.

[Yon95] Qu. Yonggang. Automatic Query Expansion Based on a Similarity Thesaurus. PhD Thesis Swiss Federal Institute of Technology ETH Zürich, 1995.

[ZRSJ+92] Edith Ziffels-Röhr, Eberhard Stage, Gudrun Johannsen, Elke Ketelaer, Marianne Andres und Helmut Michels. *AGROVOC - Mehrsprachiger Thesaurus für die Agrarwissenschaften.* ZADI, Bonn. Zweite Auflage, deutsche Version, 1992.