

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

Zur Erlangung des akademischen Grades
eines Doktors der Ingenieurwissenschaften

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

von

Ralf Nikolai
aus Hilden

Tag der mündlichen Prüfung: 19.12.2002

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

Inhaltsverzeichnis

1	Einleitung und Motivation	1
1.1	Motivation	1
1.2	Zielbeschreibung	2
1.3	Abgrenzung	4
1.4	Aufbau der Arbeit	5
2	Problemanalyse	9
2.1	Thesauri im Information Retrieval	9
2.1.1	Motivation	9
2.1.2	Definition	11
2.2	Fallstudie	14
2.2.1	Ausgangssituation Umweltinformationssysteme . .	14
2.2.2	Auswahl der Thesauri	16
2.2.2.1	Auswahlkriterien	17
2.2.2.2	GEMET, AGROVOC und GCMD Parameter Validis	17
2.2.3	Zielsetzung	18
2.3	Analyse	19
2.3.1	Informationsmodell für integrierte Thesauri	19
2.3.1.1	Thesauri	19
2.3.1.2	Inter-Thesaurus-Relationen	21

2.3.1.3	Begriffe	23
2.3.1.4	Invarianten	25
2.3.1.5	Konflikte	26
2.3.2	Begriffsintegration	27
2.3.2.1	Analyse der Komponententheseauri	27
2.3.2.2	Auffinden und Klassifizieren von Inter- Thesaurus-Relationen	28
2.3.2.3	Einführen von Ergänzenden Begriffen . .	36
2.3.2.4	Konflikterkennung und -behandlung . . .	36
2.3.2.5	Vorgehensmodell zur Integration	37
2.3.2.6	Referenzarchitektur für die Begriffsinte- gration	38
2.3.2.7	Benutzeragent	38
2.3.3	Bewertung der Güte eines Multi-Thesaurus-Systems	39
2.3.4	Ausführungsmaschine	40
2.4	Fokus der Arbeit	41
3	Stand der Forschung	43
3.1	Informationsmodelle für integrierte Thesauri	43
3.1.1	Klassifikation von Modellen für Multi-Thesaurus- Systeme	44
3.1.1.1	Multi-Thesaurus-Umgebungen	44
3.1.1.2	Thesaurus-Wechsel-Systeme	44
3.1.1.3	Thesaurusverbünde	45
3.1.2	Modelle für Multi-Ontologie-Systeme	47
3.1.2.1	Die ONIONS-Methodologie	48
3.1.2.2	Der OBSERVER-Ansatz	48
3.1.2.3	Scalable Knowledge Composition (SKC) .	49
3.1.2.4	Chimera	49
3.1.2.5	PROMPT	50
3.1.3	Bewertung der Modelle	50

3.2	Begriffsintegration	53
3.2.1	Datengrundlagenorientierte Klassifikation der Ansätze	53
3.2.1.1	Dokumentenbestandsbasierte Ansätze	53
3.2.1.2	Thesaurusbasierte Ansätze	54
3.2.1.3	Anfragebasierte Ansätze	55
3.2.2	Verfahren zum Auffinden von Ergänzenden Begriffen	56
3.2.3	Verfahren zum Auffinden und Klassifizieren von Inter-Thesaurus-Relationen	56
3.2.3.1	Linguistische Verfahren	57
3.2.3.2	Strukturbasierte Verfahren	60
3.2.3.3	Attributbasierte Verfahren	63
3.2.3.4	Verwendung externer Wissensquellen	63
3.2.3.5	Bewertung	64
3.2.4	Konflikterkennung und -behandlung	65
3.2.5	Vorgehensmodelle	67
3.2.5.1	Mehrstufige Verfahren	67
3.2.5.2	Phasenmodelle	67
3.2.5.3	Transformation in ausdrucksstärkere Modelle	68
3.3	Bewertung der Güte eines Multi-Thesaurus-Systems	71
3.4	Resümee	72
4	Grundideen des Lösungsansatzes	75
4.1	Aufbau der Lösung	76
4.2	Bausteine der Lösung	76
4.2.1	Informationsmodelle	76
4.2.2	Begriffsintegration	78
4.2.2.1	Vorgehensmodell	78
4.2.2.2	Architektur	81
4.2.2.3	Benutzeragent	81

4.2.3	Ausführungsmaschine	82
5	Informationsmodell für Thesauri	83
5.1	Formales Modell monolingualer Thesauri	84
5.1.1	Thesauri	84
5.1.2	Begriffe und Benennungen	85
5.1.3	Relationen	89
5.1.3.1	Äquivalenzrelation	90
5.1.3.2	Benutze-Kombination-Relation	90
5.1.3.3	Abstraktionsrelation	91
5.1.3.4	Bestandsrelation	92
5.1.3.5	Hierarchierelation	92
5.1.3.6	Assoziationsrelation	94
5.1.3.7	Paarweise Disjunktheit der Relationen	94
5.2	Beschreibung von Thesauri als Graphen	95
5.2.1	Knoten und Kanten	95
5.2.2	Pfade	99
5.2.3	Invarianten	100
5.2.3.1	Keine Selbstverweise	100
5.2.3.2	Einzigkeit einer Kante	100
5.2.3.3	Zyklenfreiheit der Abstraktionspfade	100
5.2.3.4	Zyklenfreiheit der Bestandspfade	101
5.2.3.5	Zyklenfreiheit der Hierarchiepfade	101
5.2.3.6	Redundanzfreiheit der Abstraktionspfade	101
5.2.3.7	Verbundenheit der Nicht-Deskriptoren	101
5.2.3.8	Einzigkeit einer Menge von BK-Kanten	102
5.3	Resümee	102
6	Informationsmodell für Thesaurusföderationen	103
6.1	Analyse	104

6.1.1	Thesauri	104
6.1.2	Relationen	105
6.1.3	Begriffe	106
6.1.4	Gruppen	107
6.1.5	Invarianten und Konflikte	108
6.2	Informationsmodell	112
6.2.1	Komponententhesauri	113
6.2.2	Integrationswissen	113
6.2.2.1	Metainformationen über Komponenten- thesauri	114
6.2.2.2	Relationen	114
6.2.2.3	Begriffe	115
6.2.2.4	Gruppen	116
6.2.2.5	Invarianten und Konflikte	116
6.2.2.6	Zusammenfassung	117
6.3	Formales Thesaurusföderations-Modell	117
6.3.1	Thesaurusföderation	117
6.3.2	Komponententhesauri und Metainformationen	119
6.3.3	Begriffe und Benennungen	119
6.3.4	Konfliktmarkierungen	127
6.3.5	Implizierte Intra-Thesaurus-Relationen	128
6.3.6	Inter-Thesaurus-Relationen	128
6.3.6.1	Inter-Thesaurus-Äquivalenzrelation	129
6.3.6.2	Inter-Thesaurus-Benutze-Kombination- Relation	129
6.3.6.3	Inter-Thesaurus-Abstraktionsrelation	129
6.3.6.4	Inter-Thesaurus-Bestandsrelation	135
6.3.6.5	Inter-Thesaurus-Hierarchierelation	136
6.3.6.6	Inter-Thesaurus-Assoziationsrelation	137
6.3.7	Relationsübergreifende Eigenschaften	137

6.3.7.1	Paarweise Disjunktheit der Inter- Thesaurus-Relationen	137
6.3.7.2	Keine Assoziationsbeziehungen zwischen Schwesterknoten	140
6.3.7.3	Beibehaltung des Hierarchierelationstyps	141
6.3.8	Konsistenz der Konfliktmarkierungen	143
6.4	Beschreibung von Thesaurusföderationen als Graphen . .	144
6.4.1	Knoten und Kanten	144
6.4.2	Pfade	149
6.4.3	Invarianten	150
6.4.3.1	Identität der Komponententhesauri . . .	150
6.4.3.2	Konsistenz der Komponententhesauri und des Thesaurus der Ergänzenden Be- griffe	150
6.4.3.3	Richtiger Einsatz der thesaurus- verbindenden Kanten	150
6.4.3.4	Verbundenheit der Ergänzenden Begriffe	151
6.4.3.5	Markierung bei Verstoß gegen Einzigar- tigkeit einer Kante	151
6.4.3.6	Markierung von Abstraktionszyklen . . .	153
6.4.3.7	Weitere Markierungsinvarianten	153
6.5	Resümee	153
7	Wissensakquisitionsarchitektur	155
7.1	Anforderungen	156
7.2	Blackboard-Architekturen	158
7.2.1	Einführung	158
7.2.1.1	Blackboard-Modell	158
7.2.1.2	Komponenten eines Blackboard-Systems	158
7.2.1.3	Entwurfsvfreiräume	159
7.2.1.4	Potentielle Probleme	160
7.2.2	Blackboard-Architekturen in der Praxis	161

7.2.3	Anforderungsabgleich	162
7.3	Blackboardbasierte Wissensakquisitionsarchitektur FA ² ITH165	
7.3.1	Überblick	165
7.3.2	Blackboard und Blackboard-Einträge	167
7.3.2.1	Hypothesen	169
7.3.2.2	Fakten	171
7.3.2.3	Registrierungen	171
7.3.2.4	Bewertungen	173
7.3.2.5	Kontexte	173
7.3.3	Planungsagent	174
7.3.4	Experten	175
7.3.5	Benutzeragent	177
7.3.5.1	Anforderungen	177
7.3.5.2	Lösungsansatz	178
7.3.6	Steueragent	180
7.3.7	Moderator	181
7.3.8	Ausführungsagent	182
7.4	Bewertungsmodell	183
7.4.1	Bewertung der Hypothesen	184
7.4.2	Bewertung der Experten	187
7.4.3	Berechnungsmodell für eine aggregierte Gesamtbewertung	189
7.5	Resümee	191
8	Vorbereitungsphase	193
8.1	Herstellung der Konformität der Informationsmodelle	194
8.1.1	Begriffe und Benennungen	194
8.1.1.1	Keine Ausweisung von Vorzugsbenennungen	194
8.1.1.2	Keine Trennung Benennung – Annotationen	196

8.1.1.3	Identische BK-Verweismengen von verschiedenen Nicht-Deskriptorknoten	197
8.1.1.4	Weitere Informationsmodellabweichungen	198
8.1.2	Semantische Relationen	198
8.1.2.1	Keine Differenzierung der Hierarchierelation	198
8.1.3	Gruppen	206
8.1.3.1	Gruppenzuordnung ohne Gruppen in den Komponententhesauri	207
8.1.3.2	Gruppenzuordnung mit einer Menge an Gruppen	207
8.1.3.3	Gruppenzuordnung mit verschiedenen Mengen an Gruppen	208
8.2	Herstellen normierter Benennungen	208
8.2.1	Allgemeine Benennungsnormierung	209
8.2.2	Normierung von Eigennamen	211
8.3	Herstellen normierter Definitionen	211
8.4	Herstellen von Zugriffsschnittstellen	212
8.5	Resümee	212
9	Analyse von Thesauri	217
9.1	Teilbereiche einer Analyse	218
9.2	Kriterien zur Komponententhesaurusanalyse	220
9.2.1	Qualitative Analysen	221
9.2.2	Quantitative Analyse der Benennungen	222
9.2.3	Quantitative Analyse der Relationen	227
9.2.4	Quantitative Analyse der Struktur	229
9.2.5	Klassifikation der Thesauri	232
9.3	Evaluierung ausgewählter Thesauri	233
9.3.1	Analyse der Benennungen	233
9.3.2	Analyse der Relationen	235

9.3.3	Analyse der Struktur	237
9.3.4	Zusammenfassung der Ergebnisse	239
9.4	Resümee	241
10	Integrationsstrategie	243
10.1	Ziele der Integrationsstrategie	243
10.2	Spezifikation einer allgemeingültigen Integrationsstrategie	245
10.2.1	Strategie-Ebene 1: Top-Level-Integrationsstrategie	246
10.2.2	Strategie-Ebene 2: Teilphasen der Integration . . .	249
10.2.3	Strategie-Ebene 3: Ablauf innerhalb der Teilphasen	251
10.2.3.1	Initiale Integration	251
10.2.3.2	Zwischenergebnisbasierte Optimierung . .	252
10.2.3.3	Bewertungsbasierte Optimierung	253
10.3	Strategiemodifikationen und Anpassungen der Aufgabena-	
	genda	254
10.4	Resümee	257
11	Realisierungsphase	261
11.1	Einbringen der Problemlösungsverfahren	262
11.1.1	Ordnungskriterien zur Einbringung der Verfahren .	263
11.1.2	Spezielle Verfahren und deren Einordnung	265
11.1.3	Konfigurieren von Verfahren	280
11.2	Faktenerzeugung und Konfliktmarkierung	282
11.2.1	Berechnung einer aggregierten Hypothesenbewer-	
	tung	282
11.2.2	Qualitative Überprüfung von Hypothesen und	
	Fakten	283
11.2.2.1	Überprüfung einzelner Hypothesen	284
11.2.2.2	Überprüfung der Menge der Hypothesen	285
11.2.2.3	Überprüfung von Faktenmenge und vor-	
	handenem Integrationswissen	288

11.2.2.4	Überprüfung bei zusätzlicher Betrachtung der durch die Hypothesen/Fakten implizierten Beziehungen	291
11.3	Erweiterungen und Veränderungen am Integrationswissen	298
11.3.1	Einfügen von Inter-Thesaurus-Beziehungen	299
11.3.2	Einfügen von Ergänzenden Begriffen	302
11.3.3	Entfernen von Inter-Thesaurus-Beziehungen	303
11.3.3.1	Entfernen von Kanten	303
11.3.3.2	Aktualisieren der Konfliktmenge	305
11.3.4	Entfernen von Ergänzenden Begriffen	307
11.4	Resümee	308
12	Analyse und Bewertung von Thesaurusföderationen	309
12.1	Steuerung der Begriffsintegration durch Hypothesenziele	310
12.2	Quantitative Analyse einer Thesaurusföderation	312
12.2.1	Quantitative Analyse der Benennungen	312
12.2.2	Quantitative Analyse der Relationen	315
12.2.3	Quantitative Analyse der Struktur	318
12.2.4	Quantitative Analyse der Konflikte	319
12.3	Qualitative Analyse einer Thesaurusföderation	322
12.3.1	Korrektheit	322
12.3.2	Vollständigkeit	325
12.3.3	Berücksichtigung von Ergänzenden Begriffen	326
12.4	Exemplarische Evaluierung eines Zwischenergebnisses	327
12.4.1	Anzahl Äquivalenzbeziehungen	328
12.4.2	Benutze-Kombinations-Beziehungs-Anteil	328
12.4.3	Inter-Thesaurus-Interkonnektivität	329
12.4.4	Zugänglichkeit und polyhierarchische Begriffe	330
12.5	Resümee	331
13	Ausführungsmaschine	333

13.1	Analyse	333
13.1.1	Voraussetzungen und Annahmen	335
13.1.2	Anfragebearbeitung	336
13.1.3	Einheitliche Zugriffsschnittstelle	336
13.2	Architektur	337
13.2.1	Übersicht	337
13.2.2	Kommunikationsschnittstellen und -formate	339
13.2.3	Protokolle	341
13.3	Thesaurusföderationsmediator	342
13.3.1	Schnittstellen	343
13.3.2	Anfragebearbeitung	344
13.3.2.1	Detailanfragen	345
13.3.2.2	Navigationsanfragen	346
13.3.2.3	Abbildungsanfragen	355
13.3.3	Anfragerereformulierung und -erweiterung	361
13.4	Kapseln	363
13.4.1	Kapseln für Thesauri mit Anfrageschnittstellen	364
13.4.2	Kapseln für Thesauri mit HTML/HTTP-Anfrageschnittstellen	364
13.5	Facilitator und Informationssystemmediator	365
13.6	Resümee	367
14	Zusammenfassung und Ausblick	369
14.1	Zusammenfassung	369
14.1.1	Ausgangssituation	369
14.1.2	Lösungsansatz	370
14.1.3	Realisierung des Ansatzes	374
14.2	Ausblick	376
14.2.1	Weiterentwicklung	376
14.2.2	Übertragbarkeit	378

A Aufgaben-Agenda-Definitions-Sprache AADS	381
B Spezifikation der Expertenein-/ausgaben	385
B.1 Informationen über erforderliche Eingaben	385
B.2 Informationen über mögliche Ausgaben	386
C SOAP-Repräsentation einer Anfrage an den Mediator	387
D Glossar	391
D.1 Allgemeine Begriffe	391
D.2 Begriffe aus den Bereichen Terminologielehre und Infor- mation Retrieval	392
D.3 Begriffe aus der Linguistik	398
D.4 Begriffe aus dem Bereich der Systemintegration	400