

Überwindung semantischer Heterogenität bei multiplen Data-Warehouse-Systemen

von Stefan Hartmann



UNIVERSITY OF
BAMBERG
PRESS

Schriften aus der Fakultät
Wirtschaftsinformatik und Angewandte Informatik
der Otto-Friedrich-Universität Bamberg

Band 1

Schriften aus der Fakultät
Wirtschaftsinformatik und Angewandte Informatik
der Otto-Friedrich-Universität Bamberg

Band 1



University of Bamberg Press 2008

Überwindung semantischer Heterogenität bei multiplen Data-Warehouse-Systemen

von Stefan Hartmann



University of Bamberg Press 2008

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese
Publikation in der Deutschen Nationalbibliographie;
detaillierte bibliographische Informationen sind im Internet
über <http://dnb.ddb.de/> abrufbar.

Diese Arbeit hat der Fakultät Wirtschaftsinformatik und Angewandte Informatik
als Dissertation vorgelegen.

1. Gutachter: Prof. Dr. Elmar J. Sinz

2. Gutachter: Prof. Dr. Otto K. Ferstl

Tag der mündlichen Prüfung: 04. Dezember 2008

Dieses Werk ist als freie Onlineversion über den Hochschulschriften-Server
(OPUS; <http://www.opus-bayern.de/uni-bamberg/>) der Universitätsbibliothek
Bamberg erreichbar. Kopien und Ausdrücke dürfen nur zum privaten und
sonstigen eigenen Gebrauch angefertigt werden.

Herstellung und Druck: Digital Print Group, Erlangen

© University of Bamberg Press Bamberg 2008

<http://www.uni-bamberg.de/ubp/>

ISSN: 1867-7401

ISBN: 978-3-923507-41-2 (Druckausgabe)

URN: urn:nbn:de:bvb:473-opus-1601

Geleitwort

Data-Warehouse-Systeme (DWH-Systeme) stellen mittlerweile für viele Organisationen einen unverzichtbaren Bestandteil von Führungsinformationssystemen dar. Dabei existiert in den wenigsten Fällen ein umfassendes integriertes DWH-System für die Unterstützung des Managements. Vielmehr wurden meist in einzelnen Teil-Organisationen DWH-Systeme für unterschiedliche Zwecke und mit unterschiedlicher Reichweite isoliert voneinander entwickelt. Das Ergebnis sind DWH-Systeme zu ähnlichen oder verwandten Sachverhalten auf Basis individueller Begriffssysteme.

Sollen nun Berichte aus unterschiedlichen DWH-Systemen zusammengeführt oder wenigstens einheitlich interpretiert werden, so stellt sich das Problem der semantischen Homogenisierung dieser Systeme. Besonders gravierend wirkt sich die bestehende semantische Heterogenität bei verteilten Organisationen aus, wie z. B. Sparten-Organisationen oder Management-Holdings. Auch das Hochschulsystem eines Bundeslandes oder der Bundesrepublik Deutschland lässt sich in diesem Sinne als verteilte Organisation interpretieren. Eine inhaltliche Abstimmung der verschiedenen DWH-Systeme stellt bei diesen gewachsenen und dynamisch veränderlichen Organisationen mit ihren individuellen Bedürfnissen keine aussichtsreiche Strategie dar.

Ziel der Dissertation von STEFAN HARTMANN ist es, einen Beitrag zur Überwindung dieser semantischen Heterogenität zu leisten. Den allgemeinen Untersuchungsgegenstand der Arbeit bilden DWH-basierte Führungsinformationssysteme speziell in verteilten Organisationen. Im Mittelpunkt steht die semantische Heterogenität bei multiplen DWH-Systemen. Mit dem Begriff multiple DWH-Systeme werden eigenständig betriebene und zugleich fachlich nicht überschneidungsfreie, heterogene DWH-Systeme bezeichnet. Das Sachziel der Untersuchung ist die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen. Formalziele der Arbeit sind die Beibehaltung der Autonomie bestehender DWH-Systeme, die Flexibilität und Erweiterbarkeit des Lösungsansatzes zur Überwindung der semantischen Heterogenität sowie dessen Plattformunabhängigkeit. Zur Lösung des beschriebenen Problems schlägt STEFAN HARTMANN eine semantische Homogenisierungsschicht vor, die auf Technologien des Semantic Web aufbaut.

Die Arbeit entstand im Umfeld des Projekts CEUS^{HB} (Computerbasiertes Entscheidungsunterstützungssystem für die Hochschulen in Bayern) am Wissenschaftlichen Institut für Hochschulsoftware der Universität Bamberg (ihb).

Mit seiner Dissertation legt STEFAN HARTMANN einen wertvollen Beitrag zur Schließung einer Forschungslücke im Bereich der semantischen Homogenisierung vor. Die Arbeit ist gleichzeitig von hoher praktischer Relevanz, wenn man bedenkt, dass in vielen Organisationen „gewachsene“ DWH-Systeme existieren, die auf heterogenen Begriffssystemen beruhen und deren Berichte auf semantisch eindeutiger Basis zusammengeführt werden müssen. Die Arbeit sei allen zur Lektüre empfohlen, die in Wissenschaft und Praxis mit dem Problem der semantischen Homogenisierung von DWH-Systemen befasst sind.

Elmar J. Sinz

Vorwort

Data-Warehouse-Systeme (DWH-Systeme) sollen Entscheidungsträgern die für ihre aktuelle Aufgabe relevanten Informationen rechtzeitig und in geeigneter Form zur Verfügung stellen. Die Existenz multipler DWH-Systeme in betrieblichen Organisationen erschwert jedoch den Aufbau eines konsolidierten Berichtswesens. Eine korrekte Interpretation sowie direkte Vergleichbarkeit der abgeleiteten Berichtsergebnisse ist aus diesem Grund kaum bzw. nur mit hohem Aufwand möglich.

Während meiner Tätigkeit als Projektmitarbeiter von CEUS^{HB} am Wissenschaftlichen Institut für Hochschulsoftware der Universität Bamberg (ihb) habe ich meine Forschungsarbeit insbesondere der semantischen Beschreibung von Metadaten in DWH-Systemen gewidmet. In der vorliegenden Arbeit stelle ich meine Ergebnisse in Form eines neuartigen Ansatzes zur Abbildung von DWH-Metadaten vor. Dieser wird innerhalb einer semantischen Homogenisierungsschicht genutzt, um semantische Heterogenität bei multiplen DWH-Systemen zu überwinden.

Für die Betreuung und wertvolle Unterstützung bei der Erstellung der Arbeit gilt mein besonderer Dank Herrn Prof. Dr. Elmar J. Sinz. Für die Übernahme des Zweitgutachtens danke ich Herrn Prof. Dr. Otto K. Ferstl recht herzlich sowie Herrn Prof. Dr. Christoph Schlieder für die Begleitung der Arbeit als Mitglied der Promotionskommission.

Darüber hinaus möchte ich meinen Projektkollegen Achim Ulbrich-vom Ende, Markus Plaha, Julia Ringler, Tim-Oliver Förtsch und Carsten Jürck zum einen für die jahrelange, erfolgreiche und freundschaftliche Zusammenarbeit und zum anderen für die konstruktiven Beiträge zum Gelingen dieser Arbeit Danke sagen. Ein Dankeschön gilt in diesem Zusammenhang ebenfalls dem gesamten Team des Lehrstuhls für Wirtschaftsinformatik, insbesondere Systementwicklung und Datenbankanwendung. Für die eingehende Durchsicht meines Manuskripts bedanke ich mich bei Irene Berding und Ruth Hammerschmidt.

Vor allem möchte ich meinen Eltern und meiner Schwester danken, die mich in meinen Vorhaben stets ermuntert und gefördert haben. Sie gaben mir damit den erforderlichen Rückhalt für den erfolgreichen Abschluss dieser Arbeit.

Bamberg, Dezember 2008

Stefan Hartmann

Inhaltsverzeichnis

| | |
|-----------------------|------|
| Abbildungsverzeichnis | xv |
| Tabellenverzeichnis | xvii |
| Quellcodeverzeichnis | xix |
| Abkürzungsverzeichnis | xxi |

| | |
|---|-----------|
| 1 Einleitung | 1 |
| 1.1 Motivation | 2 |
| 1.2 Zielsetzung und Abgrenzung | 4 |
| 1.2.1 Untersuchungssituation | 4 |
| 1.2.2 Untersuchungsgegenstand | 5 |
| 1.2.3 Untersuchungsziel | 5 |
| 1.2.4 Untersuchungsverfahren | 6 |
| 1.2.5 Gegenwärtiger Forschungsstand | 8 |
| 1.3 Entstehung und Umfeld | 11 |
| 1.4 Aufbau | 16 |
| | |
| 2 Data-Warehouse-Systeme in betrieblichen Organisationen | 19 |
| 2.1 Betriebliche Organisationen | 19 |
| 2.1.1 Organisationsstrukturen | 19 |
| 2.1.2 Verteilte Organisationen | 22 |
| 2.1.3 Hochschulwesen als verteilte Organisation | 25 |
| 2.2 Führungsinformationssysteme | 27 |
| 2.2.1 Motivation und historische Entwicklung | 27 |
| 2.2.2 Begriffsbestimmung und Abgrenzung | 30 |
| 2.3 Data-Warehouse-Systeme | 32 |
| 2.3.1 Begriffsbestimmung und Abgrenzung | 32 |
| 2.3.2 Architektur eines Data-Warehouse-Systems | 34 |
| 2.3.3 Einordnung eines Data-Warehouse in eine betriebliche Organisation | 36 |
| 2.3.4 Aspekte einer Data-Warehouse-Einführung | 37 |
| 2.3.5 Data-Warehouse-Systeme in einer verteilten Organisation | 42 |
| 2.3.6 Spezifische Data-Warehouse-Konzepte | 43 |
| 2.4 Zusammenfassung | 44 |

| | | |
|----------|---|-----------|
| 3 | Metadaten in Data-Warehouse-Systemen | 45 |
| 3.1 | Begriffsbestimmung | 45 |
| 3.2 | Bedeutung von Metadaten für Data-Warehouse-Systeme | 45 |
| 3.3 | Klassifikation von Metadaten | 47 |
| 3.4 | Standardisierungen und Austauschformate für Metadaten | 50 |
| 3.4.1 | Repository-Standards | 50 |
| 3.4.2 | Metadaten-Standards und -Rahmenwerke | 51 |
| 3.4.3 | Metadaten-Austauschformate | 53 |
| 3.5 | Metadaten-Verwaltung | 54 |
| 3.6 | Zusammenfassung | 57 |
| 4 | Führungsinformationen in betrieblichen Organisationen | 59 |
| 4.1 | Daten, Information und Wissen | 59 |
| 4.1.1 | Begriffsbestimmung und Abgrenzung | 59 |
| 4.1.2 | Kodifizierbarkeit von Wissen | 62 |
| 4.1.3 | Wissensmanagement | 64 |
| 4.2 | Informationsversorgung von Entscheidungsträgern | 65 |
| 4.2.1 | Begriffsbestimmung und Abgrenzung | 66 |
| 4.2.2 | Entscheidungsaufgaben | 67 |
| 4.2.3 | Informationspathologien | 69 |
| 4.3 | Kontextinformationen bei Data-Warehouse-Systemen | 71 |
| 4.4 | Zusammenfassung | 77 |
| 5 | Heterogenität in Datenschemata von Data-Warehouse-Systemen | 79 |
| 5.1 | Grundlagen der Semiotik | 79 |
| 5.2 | Merkmale eines Begriffs | 82 |
| 5.2.1 | Repräsentationsebenen eines Begriffs | 82 |
| 5.2.2 | Sprachliche Defekte | 85 |
| 5.2.3 | Behandlung sprachlicher Defekte bei Data-Warehouse-Projekten | 86 |
| 5.2.4 | Methode zur Rekonstruktion von Fachbegriffen | 88 |

| | |
|--|------------|
| 5.3 Heterogenität in Datenschemata | 92 |
| 5.3.1 Arten von Heterogenität | 92 |
| 5.3.2 Semantische Heterogenität | 94 |
| 5.3.3 Ursachen für Heterogenität | 95 |
| 5.3.4 Überwindung von Heterogenität | 96 |
| 5.3.5 Semantische Ähnlichkeitsmaße | 97 |
| 5.4 Schema-Management | 99 |
| 5.4.1 Schema-Matching | 101 |
| 5.4.2 Schemaintegration | 103 |
| 5.4.3 Schema-Mapping | 103 |
| 5.4.4 Diskussion des Schema-Managements | 104 |
| 5.5 Anforderungen an eine semantisch reichhaltige Beschreibung von Data-Warehouse-Metadaten | 105 |
| 5.6 Zusammenfassung | 107 |
| 6 Exkurs: Konzepte des Semantic Web | 109 |
| 6.1 Motivation und Ziele des Semantic Web | 109 |
| 6.2 Technologien des Semantic Web | 110 |
| 6.2.1 Uniform Resource Identifier (URI) | 112 |
| 6.2.2 Extensible Markup Language (XML) | 112 |
| 6.2.3 Resource Description Framework (RDF) | 115 |
| 6.3 Ontologien | 125 |
| 6.3.1 Begriffsbestimmung und Abgrenzung | 125 |
| 6.3.2 Wissensrepräsentationssprachen | 131 |
| 6.3.3 Web Ontology Language (OWL) | 132 |
| 6.4 Zusammenfassung | 137 |

| | | |
|----------|---|------------|
| 7 | Realisierung einer semantisch reichhaltigen Beschreibung von Data-Warehouse-Metadaten | 139 |
| 7.1 | Potentiale von RDF und RDF-Schema zur semantischen Beschreibung von Data-Warehouse-Metadaten | 139 |
| 7.1.1 | Allgemeine Beschreibungskonzepte und -elemente | 140 |
| 7.1.2 | Data-Warehouse-Kernelemente | 144 |
| 7.1.3 | Data-Warehouse-Organisationselemente | 148 |
| 7.1.4 | Data-Warehouse-Kontextelemente | 150 |
| 7.1.5 | Data-Warehouse-Systemelemente | 151 |
| 7.1.6 | Zusammenfassung | 153 |
| 7.2 | Ein RDF-Schema zur semantischen Beschreibung von Data-Warehouse-Metadaten | 153 |
| 7.3 | Eine OWL-Ontologie zur semantisch reichhaltigen Beschreibung von Data-Warehouse-Metadaten | 166 |
| 7.3.1 | Allgemeine Entwicklungskriterien einer Ontologie für Data-Warehouse-Metadaten | 166 |
| 7.3.2 | Potentiale von OWL zur semantisch reichhaltigen Beschreibung von Data-Warehouse-Metadaten | 167 |
| 7.3.3 | Aufbau der OWL-Ontologie zur semantisch reichhaltigen Beschreibung von Data-Warehouse-Metadaten | 170 |
| 7.3.4 | Zusammenfassung | 179 |
| 7.4 | Zusammenfassung | 179 |
| 8 | Konzeption einer semantischen Homogenisierungsschicht zur Überwindung der semantischen Heterogenität bei multiplen Data-Warehouse-Systemen | 181 |
| 8.1 | Aufbau einer semantischen Homogenisierungsschicht für Data- Warehouse-Systeme | 181 |
| 8.1.1 | Grundidee einer semantischen Homogenisierungsschicht | 182 |
| 8.1.2 | Ontologiebasierte Ausgestaltung einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme | 183 |
| 8.1.3 | Zusammenfassung | 187 |

| | |
|--|------------|
| 8.2 Gesamt-Architektur einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme | 188 |
| 8.2.1 Vorstellung einer Gesamt-Architektur | 189 |
| 8.2.2 Kopplungskomponenten | 193 |
| 8.2.3 Versionierung | 196 |
| 8.2.4 Funktionen | 198 |
| 8.2.5 Zusammenfassung | 200 |
| 8.3 Kritische Würdigung einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme | 201 |
| 8.3.1 Modellierung von Data-Warehouse-Schemata | 201 |
| 8.3.2 Überwindung semantischer Heterogenität | 203 |
| 8.3.3 Erfüllung von Formalzielen | 205 |
| 8.4 Zusammenfassung | 206 |
| 9 Fallstudie zu einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme | 207 |
| 9.1 Einführung der Fallstudie | 207 |
| 9.2 Modellierung von Data-Warehouse-Schemata | 209 |
| 9.3 Überwindung der semantischen Heterogenität | 218 |
| 9.3.1 Kopplung der Data-Warehouse-Systeme der Fallstudie an eine semantische Homogenisierungsschicht | 218 |
| 9.3.2 Ermittlung semantischer Beziehungen zwischen den Data-Warehouse-Systemen der Fallstudie | 220 |
| 9.4 Zusammenfassung | 227 |
| 10 Zusammenfassung und Ausblick | 229 |
| Literaturverzeichnis | 235 |

Abbildungsverzeichnis

| | |
|--|-----|
| Abbildung 1-1: Begriff der Untersuchungssituation (in Anlehnung an [Fers79, S. 43 ff.]..... | 5 |
| Abbildung 1-2: Logische Architektur multipler DWH-Systeme | 14 |
| Abbildung 1-3: Konzeptuelle Schemata zweier multipler DWH-Systeme in SDWM-Notation..... | 15 |
| Abbildung 2-1: Die fünf Komponenten einer Organisation [Mint92, S. 28]..... | 20 |
| Abbildung 2-2: Grundstruktur einer verteilten Organisation (in Anlehnung an [Bühn04, S. 142])..... | 23 |
| Abbildung 2-3: Führungsstruktur im Hochschulwesen [SiBU99, S. 3] | 26 |
| Abbildung 2-4: Entwicklungsphasen managementunterstützender Systeme (in Anlehnung an [WiDV04, S. 24], [BuFK92, S. 7]) | 29 |
| Abbildung 2-5: Komponenten von Managementunterstützungssystemen (in Anlehnung an [ChGI06, S. 9], [StHa05a, S. 383]) | 31 |
| Abbildung 2-6: Architektur eines DWH-Systems [BöU100, S. 17] | 34 |
| Abbildung 2-7: Regelkreisorientierte Einordnung eines DWH [Sinz02, S. 309] | 36 |
| Abbildung 2-8: Regelkreisorientierte Einordnung eines DWH bei einer verteilten Organisation | 42 |
| Abbildung 3-1: Architekturvarianten bei der Metadaten-Verwaltung [DoRa00, S. 8] | 55 |
| Abbildung 4-1: Daten-Information-Wissen-Modell (in Anlehnung an [AaNy95 S. 8], [FeSi06, S. 132]) | 60 |
| Abbildung 4-2: Kodifizierbarkeit von Wissen | 63 |
| Abbildung 4-3: Informationsbedarf und Informationsversorgung [PiRW03, S. 82] | 66 |
| Abbildung 4-4: Struktur einer Entscheidungsaufgabe [FeSi06, S. 33]..... | 67 |
| Abbildung 4-5: Kontexttypen des Informationsangebots eines DWH | 76 |
| Abbildung 5-1: Semiotische Ebenen der Informationsübertragung [PiRW03, S. 90] | 80 |
| Abbildung 5-2: Semiotisches Dreieck (in Anlehnung an [LeNa07a, S. 273])..... | 81 |
| Abbildung 5-3: Begriffsmodell (in Anlehnung an [Ortn97, S. 31], [Ortn00, S. 102], [Lehm01, S. 124])..... | 83 |
| Abbildung 5-4: Methode zur Rekonstruktion von Fachbegriffen [LeEl97, S. 177] | 89 |
| Abbildung 5-5: Organisation der Begriffsdefinition [LeEl97, S. 181] | 90 |
| Abbildung 5-6: Heterogenitätsarten | 93 |
| Abbildung 5-7: Aufgaben des Schema-Managements [LeNa07a, S. 115] | 100 |
| Abbildung 5-8: Klassifikation von Schema-Matching-Ansätzen [RaBe01, S. 338]..... | 101 |
| Abbildung 6-1: Semantic-Web-Schichten (in Anlehnung an [Brat07, S. 23], [Zieg07, S. 173])..... | 110 |
| Abbildung 6-2: RDF-Tripel [DaOS03, S. 88] | 117 |
| Abbildung 6-3: RDF-Graph | 118 |
| Abbildung 6-4: Beziehungen zwischen URI, XML und RDF [Jeck03, S. 23] | 122 |
| Abbildung 6-5: Typen von Ontologien (in Anlehnung an [SmWe01])..... | 127 |
| Abbildung 6-6: Sprachstufen der OWL | 134 |
| Abbildung 7-1: Integriertes RDF-Schema für DWH-Metadaten | 160 |
| Abbildung 7-2: RDF- und RDFS-Ebene | 161 |

| | |
|--|-----|
| Abbildung 7-3: OWL-Ontologie für DWH-Kernelemente | 172 |
| Abbildung 8-1: Grundstruktur einer semantischen Homogenisierungsschicht für DWH-Systeme | 186 |
| Abbildung 8-2: Gesamt-Architektur einer semantischen Homogenisierungsschicht für Data- Warehouse-Systeme | 189 |
| Abbildung 8-3: Semantische Ähnlichkeitsbeziehung zwischen Applikationsontologien | 191 |
| Abbildung 9-1: Darstellung der DWH-OWL mithilfe des Editors Protégé..... | 210 |
| Abbildung 9-2: Domain-Ontologie für Studentendaten (Ausschnitt) | 211 |
| Abbildung 9-3: Bildschirmkopie eines Modellierungsschritts innerhalb einer sHGS-DWH..... | 216 |
| Abbildung 9-4: Ontologiedarstellung in Protégé mithilfe der Erweiterung TGVizTab | 217 |
| Abbildung 9-5: Metadaten zu den Dimensionen eines DWH-Systems (JDW-Toolsuite) | 218 |
| Abbildung 9-6: Bildschirmkopie zum Werkzeug Convert To RDF | 219 |
| Abbildung 9-7: Applikationsontologie zum DWH-System der Hochschule A | 220 |
| Abbildung 9-8: Applikationsontologie zum DWH-System der Hochschule B | 221 |
| Abbildung 9-9: Gesamtsicht auf die zusammengeführten Applikationsontologien der Fallstudie in einer sHGS-DWH..... | 222 |

Tabellenverzeichnis

| | |
|--|-----|
| Tabelle 1: Anforderungen an eine IT-basierte Informationsversorgung [Back91] | 28 |
| Tabelle 2: Logistisches Prinzip für Informationen [Augu90, S. 23] | 30 |
| Tabelle 3: Einsatz- und Nutzenpotentiale von DWH-Metadaten (in Anlehnung an [Auth04, S. 36]) | 46 |
| Tabelle 4: Verwendungszweckgerichtete Metadaten-Kategorien [Auth04, S. 44 ff.] | 49 |
| Tabelle 5: RDFS-Elemente der Metadaten-Kategorie Terminologie | 142 |
| Tabelle 6: RDFS-Elemente der Metadaten-Kategorie Datenstruktur und -bedeutung | 143 |
| Tabelle 7: RDFS-Elemente der Metadaten-Kategorie Metadatenhistorie | 144 |
| Tabelle 8: RDFS-Elemente der Kategorie DWH-Kernelemente | 148 |
| Tabelle 9: RDFS-Elemente der Kategorie DWH-Organisationselemente | 149 |
| Tabelle 10: RDFS-Elemente der Kategorie DWH-Kontextelemente | 150 |
| Tabelle 11: RDFS-Elemente der Kategorie DWH-Systemelemente | 152 |
| Tabelle 12: Rollen in der Kategorie DWH-Kernelemente | 173 |
| Tabelle 13: Rollen in der Kategorie DWH-Organisationselemente | 174 |
| Tabelle 14: Rollen in der Kategorie DWH-Kontextelemente | 175 |
| Tabelle 15: Rollen in der Kategorie DWH-Systemelemente | 176 |

Quellcodeverzeichnis

| | |
|---|-----|
| Quellcode 1: RDF-Dokument in RDF/XML-Notation | 120 |
| Quellcode 2: RDF-Schema für DWH-Metadaten (Ausschnitt)..... | 163 |
| Quellcode 3: RDF-Dokument zu einem DWH-Schema (Ausschnitt)..... | 164 |
| Quellcode 4: OWL-Ontologie für DWH-Metadaten (Ausschnitt)..... | 178 |
| Quellcode 5: Domain-Ontologie für Studentendaten (Ausschnitt) | 214 |

Abkürzungsverzeichnis

| | |
|--------------------|---|
| AktG | Aktiengesetz |
| ANSI | American National Standards Institute |
| ASCII | American Standard Code for Information Interchange |
| AwS | Anwendungssystem |
| BayHSchG | Bayerisches Hochschulgesetz |
| bspw. | beispielsweise |
| bzw. | beziehungsweise |
| CDIF | Case Data Interchange Format |
| CEUS ^{HB} | Computerbasiertes Entscheidungsunterstützungssystem für die Hochschulen in Bayern |
| CORBA | Common Object Request Broker Architecture |
| CRM | Customer Relationship Management |
| CSV | Comma-Separated Values |
| CWA | Closed World Assumption |
| CWM | Common Warehouse Metamodel |
| d. h. | das heißt |
| DAML+OIL | DARPA Agent Markup Language + Ontology Inference Layer |
| DAML-ONT | DARPA Agent Markup Language - Ontology |
| DB | Datenbank |
| DBMS | Datenbankmanagementsystem |
| DC | Dublin Core |
| DL | Description Logic |
| DSS | Decision Support System |
| DTD | Document Type Definition |
| DWH | Data-Warehouse |
| DWH-OWL | OWL-Ontologie für DWH-Metadaten |
| DWH-RDFS | RDFS für DWH-Metadaten |
| EAI | Enterprise Application Integration |
| EBIS | Europe/Middle East/Africa Business Information System |
| ECA | Event Condition Action |
| EDW | Enterprise-Data-Warehouse |
| EII | Enterprise Information Integration |

| | |
|--------|---|
| EIS | Executive Information System |
| EMR | Enterprise-Metadaten-Repository |
| ERM | Entity-Relationship-Modell |
| etc. | et cetera |
| ETL | Extraktion, Transformation, Laden |
| EUS | Entscheidungsunterstützungssystem |
| f. | folgende [Seite] |
| ff. | fortfolgende [Seiten] |
| FIS | Führungsinformationssystem |
| FS | Fachsemester |
| GP | Geschäftsprozess |
| HOLAP | Hybrides OLAP |
| HS | Hochschulsemester |
| ID | Identifizier |
| ihb | Wissenschaftliches Institut für Hochschulsoftware der Universität Bamberg |
| IRD | Information Resource Dictionary |
| IRDS | Information Resource Dictionary System |
| ISBN | International Standard Book Number |
| ISO | International Organization for Standardization |
| IT | Informationstechnologie |
| JDW | Java-basiertes DWH |
| KIF | Knowledge Interchange Format |
| LfStad | Bayerisches Landesamt für Statistik und Datenverarbeitung |
| M&A | Mergers & Acquisitions |
| ME/RM | Multidimensionales ERM |
| MIS | Managementinformationssystem |
| MOF | MetaObject Facility |
| MOLAP | Multidimensionales OLAP |
| MUS | Managementunterstützungssystem |
| NB | Nebenbedingung |
| o. Ä. | oder Ähnliche[s] |
| OASYS | Offene Anwendungssystem-Architekturen in überbetrieblichen Wertschöpfungsketten |
| OIL | Ontology Inference Layer |
| OLAP | Online Analytical Processing |
| OLTP | Online Transaction Processing |

| | |
|----------|---|
| OMG | Object Management Group |
| OWA | Open World Assumption |
| OWL | Web Ontology Language |
| OWL-QL | OWL Query Language |
| PCTE | Portable Common Tool Environment |
| RDF | Resource Description Framework |
| RDFS | RDF-Schema |
| RDQL | RDF Data Query Language |
| ROLAP | Relationales OLAP |
| S. | Seite |
| SCM | Supply Chain Management |
| SDWM | Semantisches Data-Warehouse-Modell |
| SERM | Strukturiertes Entity-Relationship-Modell |
| SeRQL | Sesame Resource Query Language |
| SGML | Standard Generalized Markup Language |
| sHGS | Semantische Homogenisierungsschicht |
| sHGS-DWH | Semantische Homogenisierungsschicht für DWH-Systeme |
| SMORE | Semantic Markup, Ontology, and RDF Editor |
| SPARQL | SPARQL Protocol and RDF Query Language |
| SQL | Structured Query Language |
| SQL-MDi | SQL for Multi-Dimensional Integration |
| StMWFK | Bayerisches Staatsministerium für Wissenschaft, Forschung und Kunst |
| u. a. | unter anderem |
| URI | Uniform Resource Identifier |
| URL | Uniform Resource Locator |
| URN | Uniform Resource Name |
| vgl. | vergleiche |
| vs. | versus |
| W3C | World Wide Web Consortium |
| WWW | World Wide Web |
| XMI | XML Metadata Interchange |
| XML | Extensible Markup Language |
| z. B. | zum Beispiel |

1 Einleitung

Das heutige Wirtschaftsleben ist charakterisiert durch hohe Komplexität sowie permanenten und zügigen Wandel. Betriebliche Organisationen¹ müssen sich daher einerseits strukturell und andererseits prozessual kontinuierlich an System- und Umweltbedingungen anpassen, um langfristig ihr Überleben sicherzustellen. Vor diesem Hintergrund gilt es insbesondere, Trends frühzeitig zu erkennen und dieses Wissen gezielt gegenüber Wettbewerbern auszunutzen. Parallel hierzu sind kurz-, mittel- und langfristige Ziele der organisationalen Interessensträger zu verfolgen [BeBe96, S. 16], [Wint04, S. 317].

Eine fundierte Entscheidung – ob zur strategischen oder operativen Gestaltung und Lenkung einer betrieblichen Organisation – bedarf der Informationen aus allen involvierten und tangierten Bereichen. Die Qualität einer Entscheidung unterliegt dabei vor allem der schnellen Verfügbarkeit verlässlicher und bedarfsgerechter Informationen [Gabr99, S. 418]. Letztere gelten generell als unentbehrlich für jede Art der Aufgabenerfüllung in betrieblichen Organisationen [PiRW03, S. 80]. Hierdurch aktiviert und gefördert, besteht heute ein hohes Bewusstsein um den Wert der organisationseigenen Daten. Diese bilden eine elementare Grundlage der Informationsgewinnung. In betrieblichen Organisationen wird daher neben einer qualitätsgesicherten Datenerfassung insbesondere die Realisierung einer personen-, problem- und situationsgerechten Datenbereitstellung für beliebige Informationsnachfrager angestrebt [WiDV04, S. 24 ff.], [HoRS01, S. 14], [Gabr99, S. 419].

Häufig ist jedoch die Informationslage eines Führungsverantwortlichen nicht optimal auf die zu treffenden Entscheidungen ausgerichtet [BeBe96, S. 16 ff.], [Ditt99, S. 1], [Horv06, S. 3], [ChGl06, S. 4 f.]. Durch den Einsatz so genannter *Führungsinformationssysteme* (FIS; vgl. Abschnitt 2.2) wird versucht, diesem Defizit entgegen zu wirken. FIS dienen dazu, Führungskräften relevante Informationen rechtzeitig und in geeigneter Form zur Verfügung zu stellen [StHa05a, S. 382]. Eine wichtige Komponente in diesem Zusammenhang bilden Data-Warehouse-Systeme (DWH-Systeme; vgl. Abschnitt 2.3). Sie offerieren aufbereitete (konsolidierte, historisierte etc.) Daten aus operativen An-

¹ Eine *betriebliche Organisation* wird hier z. B. im Sinne einer Unternehmung verstanden. Sie stellt aus Sicht der entscheidungs- und systemorientierten Betriebswirtschaftslehre ein offenes, sozio-technisches und zielgerichtetes System dar (institutionelle Sicht) [FeSi06, S. 61].

wendungssystemen und externen Datenquellen in Form von multidimensionalen Datenstrukturen für beliebige Berichtsfragen [Sinz02].

1.1 Motivation

Projekte zur Einführung einer zentralen, holistischen DWH-Lösung konfrontieren betriebliche Organisationen nach wie vor mit umfassenden organisatorischen und technischen Herausforderungen (vgl. Abschnitt 2.3.4). Diese Gegebenheit, zusätzlich begünstigt durch die Autonomie der für einen Ausschnitt der organisationalen Gesamtaufgabe verantwortlichen Organisationsbereiche, führte in den letzten Jahren zu isoliert verwalteten, bereichsspezifischen Ausprägungen von DWH-Systemen. Von diesen wird im Allgemeinen eine kostengünstige und zeitnahe Erfüllung der individuellen Berichtsanforderungen² erwartet. Erweiterungen der häufig verteilt ausgerichteten Organisationsstruktur (z. B. durch Mergers&Acquisitions; vgl. Abschnitt 2.1.2) erhöhen zudem die Anzahl parallel betriebener DWH-Systeme [BaGü04, S. 357], [Toto06, S. 52], [ScDi06b, S. 72].

Viele Organisationen verfügen somit oftmals ungeplant oder unwissentlich über zahlreiche eigenständige DWH-Anwendungen, um die bestehende Bandbreite an Berichts- und Analyseanforderungen annähernd abdecken zu können.³ Neben heterogenen Datenschemata⁴ kennzeichnet vor allem ein verschiedenartiges Verständnis der verwendeten Fachbegriffe eine historisch gewachsene DWH-Landschaft. Zudem sind Überschneidungsbereiche sowie Inkonsistenzen über die Datenbestände dieser DWH-Systeme festzustellen. Eigenständig betriebene und zugleich fachlich nicht überschneidungsfreie, heterogene DWH-Systeme werden in dieser Arbeit als *multiple DWH-Systeme* betitelt.

² Im betrieblichen Kontext wird unter einem *Bericht* ein Überblick betriebswirtschaftlicher Sachverhalte über einen abgegrenzten Verantwortungsbereich in aufbereiteter Form verstanden [KeMU06, S. 110].

³ Einen fundierten Beleg für die Existenz unabhängig voneinander verwalteter DWH-Systemen in der Praxis bietet z. B. die Studie [Ecke04b, S. 5 ff.]. Weitere Ursachen und Gründe einer unkoordinierten Datenverteilung in betrieblichen Organisationen werden u. a. in [Jung03, S.306] und [Wint04, S. 318] beleuchtet.

⁴ Zahlreiche Projekte zur Definition eines ganzheitlichen konzeptuellen Datenschemas in einer betrieblichen Organisation scheiterten letztlich auch an der Komplexität und Inflexibilität des konzeptuellen Datenschemas [Sinz04, S. 316].

The one word that best describes the state of analytic data in large organizations is „fragmented“ [Ecke04b, S. 5].

Ein dem DWH-Gedanken ursprünglich originäres Ziel – die horizontale sowie vertikale Datenintegration an einer zentralen Stelle – rückt hierdurch aus dem Fokus. Nur durch zeit- und ressourcenintensiven Abstimmungsaufwand können bei Existenz multipler DWH-Systeme aussagekräftige und vergleichbare Berichte generiert werden. An dieser Stelle ist vor allem die notwendige Überwindung der semantischen Heterogenität zwischen den jeweils zugrunde liegenden Begriffssystemen herauszustellen. Diese bedarf besonderer Beachtung, da keine Querverbindungen und somit keine semantischen Beziehungen zwischen multiplen DWH-Systemen definiert sind. Unstimmigkeiten oder Widersprüche in den Berichten sowie eine divergente Interpretation der dargestellten Sachverhalte sind die Folge einer fehlenden oder nicht vollständigen Harmonisierung der zugehörigen Terminologien. Dadurch bedingt nehmen die Aktualität, Qualität und Flexibilität der Datenanalyse und somit auch der Wert einer abgeleiteten Information für einen Anfrager ab.

Führungsverantwortliche einer betrieblichen Organisation benötigen jedoch eine auf einer einheitlichen Semantik basierende, aktuelle und vollständige Darstellung der organisationalen Gesamtsituation. Gestützt auf diese Fakten kann ein Informationsvorsprung, d. h. ein informationsbezogener Unterschied zu Wettbewerbern, erzielt und für den wirtschaftlichen Erfolg genutzt werden. Es besteht folglich ein nach wie vor hochgradiger Bedarf an einem organisationsweit sowie semantisch homogenen Berichtswesen⁵. Dies gilt insbesondere im Hinblick auf die Realisierung eines einheitlichen Lenkungssystems⁶ [LeE197, S. 171], [Horv06, S. 68, S. 315 ff.], [BeMu98, S.22 ff.], [Kell97, S. 191].

Aufschluss über die Definition sowie Angaben zur korrekten Interpretation von Begriffen in einem DWH-System liefern dessen Metadaten (vgl. Kapitel 3). Konsistente Metadaten sind die Voraussetzung für einen effizienten Betrieb sowie eine erfolgreiche und fachgerechte Nutzung eines DWH-Systems. Derzeit verfügbare Standards zur Beschrei-

⁵ Der Begriff des (betrieblichen) *Berichtswesens* umfasst u. a. die Erzeugung und Bereitstellung von Berichten [KeMU06, S. 110].

⁶ Nach dem Phasenprinzip werden die Teilsysteme *Lenkungs-* und *Leistungssystem* einer betrieblichen Organisation unterschieden. Dem Lenkungssystem sind die Phasen Planung, Steuerung und Kontrolle, dem Leistungssystem die Phase Durchführung zugeordnet [FeSi06, S. 6 ff.].

bung von Metadaten im DWH-Umfeld (z. B. das Common-Warehouse-Metamodel (CWM); vgl. Abschnitt 3.4.2) berücksichtigen semantische Aspekte jedoch nur am Rande. Zudem finden diese – wenn überhaupt – in der Praxis nur partiell Beachtung.

Exakte Aussagen über die Semantik der in einem DWH-System verwendeten Begriffe sind jedoch unerlässlich, um eine Abstimmung zwischen den Terminologien und somit eine Vergleichbarkeit zwischen den Berichten aus multiplen DWH-Systemen erreichen zu können. Eine entsprechende Möglichkeit, DWH-Metadaten semantisch sowohl reichhaltig als auch homogen auszeichnen zu können, besteht gegenwärtig jedoch nicht. Demnach existiert auch keine adäquate Unterstützung, um die semantische Heterogenität multipler DWH-Systeme überwinden zu können.

1.2 Zielsetzung und Abgrenzung

Zur genauen Abgrenzung der Problemstellung sowie zur konkreten Zielbeschreibung der vorliegenden Arbeit wird der Begriff der *Untersuchungssituation* nach FERSTL angewandt [Fers79, S. 43 ff.]. Dieser Terminus sowie dessen Komponenten werden nachstehend skizziert und anschließend in Bezug auf diese Arbeit entsprechend ausgestaltet. Der Abschnitt schließt mit einer Kommentierung des derzeitigen Forschungsstands im Kontext der gegebenen Untersuchungssituation.

1.2.1 Untersuchungssituation

Eine Untersuchungssituation wird durch das Tripel *Untersuchungsgegenstand*, *Untersuchungsziel* und *Untersuchungsverfahren* bestimmt. Ein Untersuchungsgegenstand bzw. ein Untersuchungsobjekt wird anhand seiner bekannten (System-)Eigenschaften abgegrenzt und beschrieben. Ein Untersuchungsziel bezieht sich auf unbekannte Eigenschaften des Untersuchungsgegenstands. Untersuchungsgegenstand und Untersuchungsziel bilden das *Untersuchungsproblem*. Zur Erreichung des Untersuchungsziels dienen so genannte *Untersuchungsverfahren*. Die Antwort eines Untersuchungsverfahrens in Bezug auf ein Untersuchungsziel heißt *Problemlösung*. Zustände, Verhaltens- oder Strukturmerkmale des Untersuchungsgegenstands als Resultat der Durchführung eines Untersuchungsverfahrens werden unter dem Begriff *Untersuchungsergebnisse* subsumiert. Nachfolgende Abbildung fasst die Erläuterungen zur Untersuchungssituation zusammen.

Als eine Vorbedingung hierfür wurde die semantisch reichhaltige Auszeichnung von DWH-Metadaten erkannt. Der zu erarbeitende Lösungsansatz ist daher so zu konzipieren, dass dessen Einsatz eine semantisch reichhaltige, maschinenlesbare Beschreibung von DWH-Metadaten unterstützt. Des Weiteren soll eine sHGS die Modellierung von DWH-Schemata mit hoher struktureller und semantischer Verwandtschaft ermöglichen.

FORMALZIELE

Dem Untersuchungsziel sind die nachfolgenden Formalziele zur Seite gestellt:

- Der Lösungsvorschlag zur Überwindung der semantischen Heterogenität multipler DWH-Systeme ist so zu gestalten, dass die *Autonomie bestehender DWH-Systeme* gewahrt bleibt. Die Erreichung des Untersuchungsziels darf keine Änderung an der Konfiguration eines DWH-Systems erfordern oder dessen (Produktiv-)Betrieb beeinflussen.
- Der Lösungsansatz ist des Weiteren so zu konzipieren, dass er *flexibel einsetzbar* und *erweiterbar* ist. Er soll demnach in beliebigen Anwendungsszenarios benutzt und ggf. um Merkmale zur Erfüllung zusätzlicher Anforderungen angereichert werden können.
- Das hier an letzter Stelle genannte Formalziel umfasst die Forderung nach *Plattform- und Datenbankmanagementsystem⁸-Unabhängigkeit* der zu konstruierenden Lösung. Hard- und softwareseitig dürfen keine Restriktionen zur Erreichung des Untersuchungsziels bestehen.

1.2.4 Untersuchungsverfahren

Die Lösungsstrategie zur Erreichung des Untersuchungsziels umfasst die Konzeption einer *semantischen Homogenisierungsschicht* (sHGS) als Hilfsmittel zur Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen. Diese soll generell in verteilten Organisationen Anwendung finden können.

Im Rahmen des Untersuchungsverfahrens sollen Technologien des *Semantic Web* auf ihre Portierbarkeit und Anwendbarkeit zur semantisch reichhaltigen Metadaten-Beschreibung im DWH-Umfeld untersucht und eingesetzt werden.

⁸ Erläuterungen zum Begriff *Datenbankmanagementsystem* (DBMS) bietet u. a. [Voss08, S. 10 ff.].

DIFFERENZIERUNG DES UNTERSUCHUNGSVERFAHRENS

Die Schwerpunkte des Untersuchungsverfahrens lassen sich anhand folgender Fragestellungen gruppieren:

- *Welche Faktoren führen zu multiplen Data-Warehouse-Systemen?*
Die Beantwortung dieser Fragestellung soll Gründe für das Entstehen multipler DWH-Systeme in betrieblichen Organisationen aufzeigen. In diesem Zusammenhang sind außerdem unterschiedliche Arten von Schema-Heterogenität zu erörtern sowie insbesondere Attribute semantischer Heterogenität zu beleuchten.
- *Welche Merkmale kennzeichnen eine semantisch reichhaltige Beschreibung von DWH-Metadaten?*
Ausgehend vom gegenwärtigen Beschreibungsumfang von DWH-Metadaten, ist dieser um semantische Merkmale anzureichern. Die Untersuchungen bezüglich semantischer Heterogenität sollen dabei als ein Anknüpfungspunkt dienen, um Anforderungen an eine semantisch reichhaltige Beschreibung von DWH-Metadaten zu definieren. Letztere stellt eine notwendige Bedingung für die Erreichung des gegebenen Untersuchungsziels dar.
- *Welche Potentiale besitzt die Semantic-Web-Technologie RDF im Hinblick auf eine semantisch reichhaltige Beschreibung von DWH-Metadaten?*
Das Resource Description Framework (RDF) stellt ein Metadaten-Rahmenwerk dar, um Ressourcen einheitlich zu annotieren und in einer maschinenlesbaren und -interpretierbaren Form zu speichern. Die Potentiale von RDF sollen hinsichtlich der Erfüllung obiger Anforderungen an die Auszeichnung von DWH-Metadaten beleuchtet werden. Diese Erkenntnisse sind anschließend zur Ausarbeitung eines entsprechenden RDF-Schemas einzusetzen.
- *Welchen Mehrwert bietet eine ontologiebasierte Beschreibung von DWH-Metadaten im Hinblick auf deren semantische Auszeichnung?*
An dieser Stelle soll die Anwendbarkeit von Ontologien zur semantisch reichhaltigen Beschreibung von DWH-Metadaten aufgezeigt sowie der im Vergleich zu einer RDF-basierten Beschreibung erzielbare Mehrwert herausgestellt werden. Gemäß der Vorlage des zuvor entwickelten RDF-Schemas ist mithilfe der Ontologiesprache Web Ontology Language (OWL) eine adäquate OWL-Ontologie auszuarbeiten.

- *Wie ist die Architektur⁹ einer semantischen Homogenisierungsschicht (sHGS) zur Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen zu gestalten?*

Unter Einbeziehung der entwickelten OWL-Ontologie für DWH-Metadaten soll eine sHGS für DWH-Systeme konzipiert werden. Für diese ist zudem eine Gesamt-Architektur auszuarbeiten, mit deren Hilfe das vorliegende Untersuchungsziel erreicht und zugleich die gegebenen Formalziele beachtet werden können.

FALLSTUDIE

Anhand einer Fallstudie aus dem Hochschulumfeld wird abschließend dargelegt, wie mithilfe einer sHGS die semantische Heterogenität bei multiplen DWH-Systemen überwunden werden kann (Proof-of-Concept). Repräsentative Ausschnitte von DWH-Systemen aus dem Kontext des Projekts CEUS^{HB} (vgl. Abschnitt 1.3) dienen hierbei als Vorlage.

1.2.5 Gegenwärtiger Forschungsstand

Ansätze, um Heterogenität auf DWH-Ebene zu begegnen, werden in der Literatur – oftmals im Zusammenhang mit Heterogenität zwischen Datenbanken – eingehend diskutiert. Einen wichtigen Ursprung finden sie in den Bestrebungen um das Thema *Schema-Management*, welches in den letzten Jahren unter dem Namen *Model-Management* wieder verstärkt Aufmerksamkeit findet (vgl. Abschnitt 5.4). Schema-Management konzentriert sich insbesondere auf die Überwindung struktureller Heterogenität.

Neben den nachfolgend hervorgehobenen Lösungsvorschlägen wird zusätzlich auch im Verlauf dieser Arbeit auf etablierte Forschungsergebnisse hingewiesen.

- *Föderierte DWH-Systeme* stellen eine logische Integration¹⁰ mehrerer DWH-Systeme bzw. Data Marts dar. Eine zugehörige Architektur wird in [BeSc08] vorgestellt. Deren Grundstruktur weist in wesentlichen Punkten Übereinstimmung mit der Architektur föderierter Datenbanken auf (z. B. [Conr97]). Für die Definition der Abbildungen zwischen den lokalen Schemata und dem globalen –

⁹ Der Begriff *Architektur* umfasst sowohl den Bauplan eines Objektsystems (im Sinne einer Spezifikation der Komponenten und Beziehungen aus allen relevanten Blickwinkeln) als auch die Konstruktionsregeln für die Erstellung des Bauplans [FeSi06, S. 185].

¹⁰ Unter *Integration* wird die (Wieder)Herstellung eines Ganzen verstanden (z. B. [Mert01, S. 244]).

von einem kanonischen Modell abgeleiteten – Schema werden eine so genannte *Dimension Algebra* sowie eine *Fact Algebra* vorgeschlagen. Die Überwindung der Heterogenität erfolgt mithilfe spezieller Operatoren dieser Algebren (z. B. rename, change, convert). Im Gegensatz zum Untersuchungsziel der vorliegenden Arbeit werden bei dem hier kommentierten Ansatz eines föderierten DWH-Systems die Dimensionen und Fakten der lokalen DWH-Systeme in ein föderiertes Dimension- bzw. Fact-Schema integriert, um auf diese Weise z. B. Berichtsanfragen an die Föderation zu ermöglichen. Die vollständige und maschinenlesbare Erfassung der Semantik der Schemaelemente aus den zu integrierenden DWH-Systemen wird dabei nicht explizit betrachtet.

- Eine *Automatisierung des Schema-Matching-Prozesses* bei heterogenen DWH-Systemen, die neben strukturellen auch linguistische Kriterien berücksichtigt, ist in [BVTS07] beschrieben. Die Ähnlichkeit zwischen multidimensionalen Strukturen wird nur in Bezug auf deren Namen, Datentyp und Unterstrukturen ermittelt. Für das erstgenannte Merkmal greifen die Autoren auf *WordNet*¹¹ zurück.
- Einen kombinierten Ansatz, um Daten aus unterschiedlichen DWH-Systemen zusammenzuführen, bietet das so genannte *kollaborative Data-Warehousing*. Dieses gründet auf dem Prinzip der zusicherungs-basierten Integration (Correspondance Assertion [SpPD92]) sowie dem Grundgedanken föderierter Datenbanken (z. B. [Conr97]). Durch Anwendung vorgegebener Integrationsregeln und Einbeziehung der Korrespondenz-Zusicherungen wird ein integriertes, virtuelles Schema über die zu verknüpfenden DWH-Systeme konstruiert. Im Vordergrund dieses Ansatzes steht die flexible Integration und De-Integration autonomer DWH-Systeme in eine bzw. aus einer Kollaboration. Gemäß dem in der Literatur dargelegten Entwicklungsstand werden jedoch nur Datenwürfel unterstützt, die durch Fakten und Dimensionen mit einfachen Hierarchien modelliert sind [MaWL07], [MaWe05].

¹¹ *WordNet* ist eine lexikalische Datenbank zur englischen Sprache. Diese offeriert zu einem Term z. B. eine Liste mit Synonymen, Homonymen oder Antonymen. Weiterführende Informationen stehen unter <http://wordnet.princeton.edu/> (Abruf am 10.06.2008) zur Verfügung.

- Ein weiterer Lösungsvorschlag in diesem Zusammenhang sieht die *Vereinheitlichung der Dimensionen* heterogener DWH-Systeme vor. Zu diesem Zweck wird eine zentralisierte Pflege der Dimensionsdaten in fachlicher, technischer und organisatorischer Hinsicht angestrebt. An diesen Perspektiven ausgerichtete Handlungsempfehlungen sowie die Vorstellung einer Zielarchitektur zeigen eine mögliche Realisierung der zentralisierten Dimensionswartung auf [ScDi06a].
- Der Ansatz der *verteilten Anfrage* (distributed query) ist nicht direkt auf die Problematik heterogener DWH-Systeme ausgerichtet. Eine Anwendbarkeit ist jedoch auch in diesem Kontext denkbar. Eine verteilte Anfrage bildet ein komplexes SQL (Structured Query Language) über die zu befragenden Datenquellen. Die Stärken dieses Vorgehens liegen vor allem in der Anreicherung bestehender DWH-Systeme mit externen oder mit nahezu Echtzeitdaten zur Anfragezeit. Jedoch sind entsprechende Kenntnisse der Datenschemata erforderlich, um Daten abfragen und semantisch korrekt zusammenführen zu können. Werkzeuge in diesem Umfeld sind heute häufig auch unter dem Begriff *Enterprise Information Integration* (EII)¹² subsumiert [Ecke04b, S. 26]. Ein verwandter Lösungsvorschlag, der speziell auf die *Analyse multidimensional aufbereiteter Daten über autonome DWH-Systeme* ausgerichtet ist, findet sich in [BeSc06]. Dieser beruht auf einer Erweiterung der Anfragesprache SQL für die Integration multidimensionaler Daten (SQL-MDi: SQL for multi-dimensional integration).
- Für ein homogenes Verständnis der Objekte in einem DWH wurde des Weiteren ein so genannter *Informationsnavigator* vorgeschlagen, der als zusätzliche Komponente zu einem DWH implementiert ist. Dieser fungiert als Hilfe- und Suchsystem über eine angeschlossene Lexikonkomponente, die ein konsolidiertes Begriffssystem beinhaltet [LeJa99]. Ein Informationsnavigator kann zugleich als eine konsistente Dokumentation über die Elemente eines spezifischen DWH-Systems angesehen werden. Für die Behandlung semantischer Heterogenität bei multiplen DWH-Systemen ist dieser jedoch nicht ausgelegt.

Bei den bis dato bekannten Untersuchungen und Forschungsergebnissen in Bezug auf die Heterogenität multipler DWH-Systeme wurden semantische Aspekte entweder nicht

¹² Eine kritische Auseinandersetzung mit dem Begriff *Enterprise Information Integration* enthält u. a. folgende Veröffentlichung <http://www.dmreview.com/news/1009669-1.html> (Abruf am 10.06.2008).

beachtet oder nur peripher beleuchtet. Demzufolge existieren für das Untersuchungsziel dieser Arbeit, die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen, kaum oder nur wenig praktikable Lösungsansätze. Darüber hinaus ist im Allgemeinen zu konstatieren, dass ein Großteil der Architekturansätze und Methoden, die sich im DWH-Umfeld etabliert haben, auf die Neuentwicklung eines DWH-Systems ausgerichtet ist (z. B. [ScDi06b, S. 76]).

Abschließend ist hervorzuheben, dass der Schwerpunkt der hier vorliegenden Untersuchungssituation sich in wesentlichen Punkten vom gegenwärtig in der Praxis lancierten *Stammdaten-Management* (Master Data Management; z. B. [Russ06]) abgrenzt. Systeme zum Stammdaten-Management dienen u. a. der Bereinigung (Konsolidierung, Dubletten-Eliminierung etc.) der Stammdaten aus unterschiedlichen oder verteilt verwalteten Anwendungssystemen einer betrieblichen Organisation. Hierbei werden auch die fachlichen Definitionen von Entitäten einbezogen.

Die vorliegende Arbeit konzentriert sich zum einen ausschließlich auf die Betrachtung der Schemaebene von DWH-Systemen. Eine Harmonisierung auf Ausprägungsebene wird zunächst nicht verfolgt. Zum anderen wird die korrekte semantische Erfassung der Schemaelemente fokussiert. Dies erfolgt vor allem im Hinblick auf die Überwindung der semantischen Heterogenität multipler DWH-Systeme. Im Gegensatz zum Stammdaten-Management wird die Heterogenität jedoch nicht aufgelöst, sondern ein Weg zu deren Überwindung erarbeitet.

1.3 Entstehung und Umfeld

Die vorliegende Arbeit entstand im Rahmen des Projekts CEUS^{HB} (Computerbasiertes Entscheidungsunterstützungssystem für die Hochschulen in Bayern)¹³ am Wissenschaftlichen Institut für Hochschulsoftware der Universität Bamberg (ihb)¹⁴.

Das Projekt CEUS^{HB} wurde mit Beginn der Hochschulreform 1999 vom Bayerischen Staatsministerium für Wissenschaft, Forschung und Kunst (StMWFK) initiiert. Die Motivation für CEUS^{HB} liegt im Aufbau und der Gewährleistung einer einheitlichen und fundierten Informationsversorgung der Entscheidungsträger im bayerischen Hoch-

¹³ Vgl. <http://www.ceushb.de> (Abruf am 10.06.2008).

¹⁴ Vgl. <http://www.ihb.uni-bamberg.de> (Abruf am 10.06.2008).

schulwesen. Basierend auf einer vom Bayerischen Staatsinstitut für Hochschulforschung und Hochschulplanung¹⁵ durchgeführten Informationsbedarfsanalyse galt es, ein abgestimmtes DWH-System für das bayerische Hochschulwesen zu entwerfen. Die zuverlässige Informationsversorgung aller Entscheidungsbereiche, von den Lehreinheiten bzw. Lehrstühlen (z. B. zur Planung der Räumlichkeiten für Vorlesungen auf der Basis der Studienanfängerzahlen) über das interne Hochschulmanagement (z. B. zur intra-universitären Mittelverteilung, bei der Auswertungen über die finanzielle Situation der Lehreinheiten und die Anzahl der zu betreuenden Studenten erforderlich sind) bis hin zur übergreifenden Planung des Hochschulwesens durch das StMWFK (z. B. zur Unterstützung der Hochschulen bei der Profilbildung), stand dabei im Fokus der Forschungsarbeit.

Aus den Ergebnissen der Evaluierung des CEUS^{HB}-Prototyps wurde eine Referenzarchitektur abgeleitet, die eine hierarchisch verteilte Architektur für DWH-Systeme im Hochschulwesen beschreibt. Diese dient als Vorlage für die Realisierung lokaler DWH-Systeme an den Hochschulen in Bayern sowie am StMWFK (z. B. [SUPH07a]).

Das Wissenschaftliche Institut für Hochschulsoftware der Universität Bamberg (ihb) ist seit dem Jahr 2003 mit der landesweiten Einführung von CEUS^{HB} beauftragt. Die Abteilung 2 des ihb hat sich auf FIS im Hochschulwesen spezialisiert und mittlerweile zu einem etablierten sowie bundesweit bekannten Kompetenzzentrum entwickelt. Das Leistungsspektrum des ihb reicht von der Konzeption und Entwicklung der DWH-Schemata über die Einführung und den Betrieb der DWH-Systeme bis hin zur Unterstützung und Schulung der Endanwender. In Bezug auf das Projekt CEUS^{HB} ist das ihb zudem für die kontinuierliche Anpassung und Erweiterung der DWH-Systeme des StMWFK und der bayerischen Hochschulen verantwortlich.

In einer ersten Ausbaustufe stellt CEUS^{HB} einen Grunddatenbestand an den Hochschulen sicher. Hierzu werden anonymisierte und konsolidierte Daten der Domänen *Studenten/Prüfungen*, *Mittelbewirtschaftung* und *Personalverwaltung* über die lokalen DWH-Systeme der Hochschulen bereitgestellt. Auf Landesebene werden zudem Daten des Bayerischen Landesamts für Statistik und Datenverarbeitung (LfStaD) und des Landesamts für Finanzen berücksichtigt. Derzeit ist CEUS^{HB} für sieben Universitäten und fünf Fachhochschulen in Bayern sowie das StMWFK realisiert.

¹⁵ Vgl. <http://www.ihf.bayern.de/> (Abruf am 10.06.2008).

Den vom ihb implementierten DWH-Teilsystemen liegen die Konzepte gemäß der CEUS^{HB}-Referenzarchitektur zugrunde. Die Hochschulen besitzen jedoch bei der Einführung von CEUS^{HB} Freiheitsgrade in der Wahl der Plattform und Software (aktuell existieren Realisierungen auf Basis von MicroStrategy¹⁶ und SAP BW¹⁷). Zudem können die detaillierte Ausgestaltung der DWH-Schemata sowie die Laderoutinen für die Daten gemäß hochschulspezifischer Vorgaben modifiziert werden. Somit kann jede Hochschule CEUS^{HB} an ihre individuellen Informationsbedarfe und Berichtsstrukturen anpassen und entsprechend weiterentwickeln [HaU105, S. 20].

BEZUG ZUR UNTERSUCHUNGSSITUATION

Die Motivation für die vorliegende Arbeit entstand u. a. aus der Erkenntnis, dass trotz ähnlicher bzw. identischer Quellsysteme und homogener Realisierungskonzepte an den Hochschulen unterschiedliche Schwerpunkte auf und in die einzelnen Domänen gesetzt werden. Dies spiegelt sich insbesondere in der Anzahl, dem Detaillierungsgrad und dem (Fach-)Verständnis der definierten Dimensionen – folglich den Attributen/Merkmalen und Hierarchien – sowie der erfassten Kennzahlen wider. Würde ein landesweiter Vergleich der Hochschulen aus den jeweils lokalen DWH-Systemen generiert, so würde bei der Zusammenführung der einzelnen Berichte deren – gegenwärtig bestehende – semantische Heterogenität ersichtlich. Die Eigenschaft multipel trifft somit zu einem bestimmten Grad auch auf die DWH-Systeme des Projekts CEUS^{HB} zu. Das Ergebnis dieser Arbeit soll daher auch einen Beitrag zur Überwindung der semantischen Heterogenität im Kontext des Projekts CEUS^{HB} leisten.

VERANSCHAULICHUNG DER PROBLEMSTELLUNG

Abstimmungsbedarf hinsichtlich der Semantik entsteht im Projekt CEUS^{HB} z. B. zwischen Berichten aus den lokalen DWH-Systemen der Hochschulen und denen aus dem landesweiten DWH-System des StMWFK. Darüber hinaus können auch Berichte mit anscheinend gleichem Inhalt aus den lokalen DWH-Systemen zweier Hochschulen auf unterschiedlichen Begriffsdefinitionen basieren.

Nachstehende Abbildung skizziert – in Anlehnung an eine mögliche DWH-Implementierung im Projekt CEUS^{HB} – eine logische Architektur der DWH-Systeme zweier

¹⁶ Vgl. <http://www.microstrategy.com> (Abruf am 10.06.2008).

¹⁷ Vgl. <http://www.sap.com/bi> (Abruf am 10.06.2008).

Hochschulen. *Hochschule A* verfügt über zwei parallel betriebene DWH-Systeme (*DWH HL* u. a. für die Hochschulleitung und *DWH SPR* z. B. für Fakultätsangehörige), an *Hochschule B* wird dagegen ein zentrales DWH-System betrieben. Entscheidendes Merkmal ist, dass jedes der dargestellten DWH-Systeme eine eigenständige Metadaten-Verwaltung ohne Querbezüge zu den anderen DWH-Systemen besitzt. Zu erkennen ist ebenfalls, dass die dargestellten DWH-Systeme nicht überschneidungsfrei sind. Sie umfassen alle – zumindest partiell – Daten über Studenten.

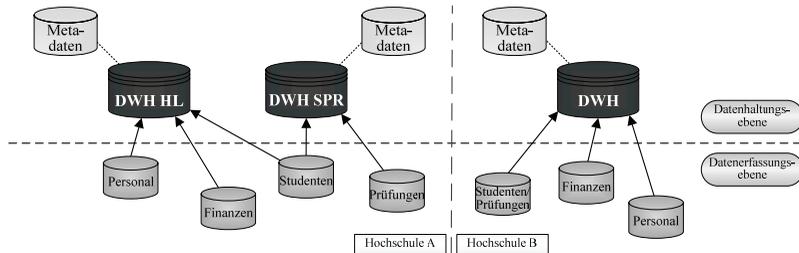


Abbildung 1-2: Logische Architektur multipler DWH-Systeme

Um Berichte zu einem (vermeintlich) gleichen Sachverhalt aus den als multipel identifizierten DWH-Systemen zweier Hochschulen miteinander vergleichen zu können, ist in jedem Fall eine manuelle Überprüfung der zugrunde liegenden Begriffsdefinitionen und davon abhängig eine entsprechende Abstimmung oder Angleichung notwendig.

Exemplarisch soll diese Problematik anhand der Berichte über Studienanfängerzahlen einer Hochschule herausgestellt werden. Studienanfängerzahlen können z. B. über den *Status eines Studenten* (Erst- und Neueinschreibung), über dessen *Fachsemesterzahl* (Fachsemester = 1) oder *Anzahl seiner Hochschulsemester* (Hochschulsemester = 1) ermittelt werden. Darüber hinaus bestehen Variationen, die bspw. beurlaubte Studenten im ersten Fachsemester bei der Zählung (nicht) beachten oder die Studienanfänger in einem höheren Fachsemester (nicht) mit einbeziehen. Definitive Vorgaben, wie sie z. B. vom Statistischen Bundesamt oder LfStaD existieren,¹⁸ werden dabei häufig nicht vollständig bzw. nicht durchgängig berücksichtigt oder gemäß eigener Spezifikationen angepasst und erweitert.

¹⁸ Der Definitionenkatalog für die Studenten- und Prüfungsstatistik kann bspw. unter folgender Adresse eingesehen werden: http://www.statistik.bayern.de/imperia/md/content/lfstad/definition_teil_1.pdf (Abruf am 10.06.2008)

Abbildung 1-3 visualisiert repräsentative Ausschnitte der konzeptuellen Schemata zu den multiplen DWH-Systemen von Hochschule A und B. Diese weisen auch Kennzahlen zur Bestimmung der Studienanfängerzahlen auf. Zur Veranschaulichung der Problemstellung genügt es, nur jeweils ein DWH-System pro Hochschule zu untersuchen. Die gewählte Darstellung folgt der Notation des *Semantischen Data-Warehouse-Modells* (SDWM) [Böhn01]. Rechtecke symbolisieren so genannte Dimensionshierarchiestufen (hier im Weiteren als Dimensionselemente benannt) oder Dimensionen, Elemente mit abgerundeten Ecken repräsentieren Kennzahlen bzw. abgeleitete Kennzahlen (gestrichelte Umrandung) [Böhn01, S. 287].

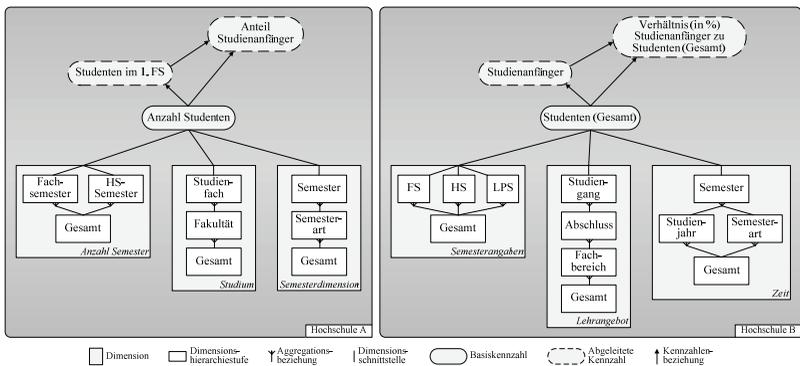


Abbildung 1-3: Konzeptuelle Schemata zweier multipler DWH-Systeme in SDWM-Notation

Anhand dieser Abbildung sind exemplarisch einige Problemkategorien zu identifizieren.

- Es existieren grundsätzlich gleiche Dimensionselemente, die jedoch unterschiedlich benannt sind. Exemplarisch sind hier die Dimensionselemente *Fachsemester* (Hochschule A) und *FS* (Hochschule B) zu nennen (Problemkategorie: *unterschiedliche Benennung bei gleicher Bedeutung*).
- Die Hierarchien fachlich zusammengehöriger Dimensionen sind verschiedenartig ausgestaltet. Dies trifft bspw. auf die Dimensionen *Studium* und *Lehrangebot* zu (Problemkategorie: *verschiedenartige Ausgestaltung der Dimensionen*).

- Die den Schemaelementen zugrunde liegenden Definitionen können aus dieser Repräsentation nur rudimentär erkannt werden. Es ist z. B. unklar, ob die Kennzahl *Studienanfänger* (Hochschule B) ebenso wie bei Hochschule A über das Attribut *Fachsemester* ermittelt wird oder auf einer der anderen, oben genannten Berechnungsvorgaben basiert (Problemkategorie: *vermeintlich gleiche Kennzahldefinitionen*).

Die letztgenannte Problematik besteht in gleicher Form bei der Verwendung eines Dimensionselements oder einer Kennzahl zur Definition eines Berichts. Gängige DWH-Werkzeuge offerieren einem Endanwender in erster Linie nur den durch das Schemaelement vertretenen Begriff, jedoch nicht dessen exakte (semantische) Definition. Veröffentlichten Berichten fehlt darüber hinaus zumeist eine die Semantik der Berichtsbestandteile eindeutig erläuternde Legende.

Die beiden eingeführten DWH-Schemata sollen die Vorlage für eine Fallstudie zur Erprobung des erarbeiteten Lösungsansatzes bilden (vgl. Kapitel 9). In diesem Zusammenhang ist zu belegen, dass durch den Einsatz einer sHGS zum einen die Semantik der hier skizzierten Schemaelemente eindeutig erfasst werden kann und zum anderen semantische Beziehungen zwischen diesen ermittelt werden können. Letztere dienen dazu, die semantische Heterogenität der multiplen DWH-Systeme von Hochschule A und B zu überwinden.

1.4 Aufbau

Nachstehend ist der Aufbau der vorliegenden Arbeit dargelegt. Zudem werden Hinweise zu den Gestaltungsrichtlinien gegeben.

In *Kapitel 1* werden die Untersuchungssituation abgegrenzt sowie Entstehung und Umfeld der Arbeit beleuchtet. Dem schließen sich in *Kapitel 2* Grundlageninformationen zu DWH-Systemen in betrieblichen Organisationen an. Parallel dazu werden Ursachen für das Auftreten multipler DWH-Lösungen identifiziert.

Kapitel 3 beinhaltet Untersuchungen über Metadaten in DWH-Systemen. Insbesondere die Bedeutung von Metadaten für DWH-Systeme sowie der gegenwärtige Praxis- und Forschungsstand bei der Standardisierung und Verwaltung von Metadaten werden herausgearbeitet und im Hinblick auf das gegebene Untersuchungsziel beleuchtet.

Gegenstand von *Kapitel 4* sind Führungsinformationen in betrieblichen Organisationen. Neben allgemeinen Ausführungen zu Daten, Information und Wissen zählt dazu eine Diskussion über die Informationsversorgung von Entscheidungsträgern sowie über die Auswirkungen so genannter Informationspathologien. Die Erarbeitung einer Definition für den Begriff Kontext bei DWH-Systemen ist ebenfalls Bestandteil dieses Kapitels.

Die Erörterung der Heterogenität in Datenschemata von DWH-Systemen bildet den Schwerpunkt von *Kapitel 5*. In diesem Zusammenhang werden charakteristische Merkmale einer vollständigen Begriffsbeschreibung herausgestellt und sprachliche Defekte untersucht. Zudem werden unterschiedliche Heterogenitätsarten sowie der prominente Forschungsbereich des Schema-Managements diskutiert. Die Erläuterungen münden in eine Anforderungsaufstellung zur semantisch reichhaltigen Beschreibung von DWH-Metadaten.

Ziel von *Kapitel 6* ist die Einführung etablierter Semantic-Web-Technologien. Zu diesem Zweck werden insbesondere das Resource Description Framework (RDF) und die Web Ontology Language (OWL) vorgestellt und hinsichtlich ihrer semantischen Ausdrucksstärke analysiert.

In *Kapitel 7* fließen die in den vorausgegangenen Kapiteln erarbeiteten Erkenntnisse zur Realisierung einer semantisch reichhaltigen Metadaten-Beschreibung zusammen. Für dieses Ziel werden zunächst die Potentiale von RDF zur semantischen Beschreibung von DWH-Metadaten dargelegt und ein – diesen Ergebnissen entsprechendes – RDF-Schema erarbeitet. Dieses dient als Vorlage für die Erstellung einer OWL-Ontologie zur semantisch reichhaltigen Metadaten-Beschreibung eines DWH-Systems.

Kapitel 8 umfasst die Konzeption einer semantischen Homogenisierungsschicht (sHGS) zur Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen. Den Kernpunkt verkörpert die Ausarbeitung einer möglichen Gesamt-Architektur für eine sHGS. Diesbezüglich wird auch auf die ontologiebasierte Beschreibung von DWH-Metadaten des vorherigen Kapitels zurückgegriffen. Das Kapitel schließt mit einer kritischen Würdigung der erarbeiteten Ergebnisse.

Gegenstand von *Kapitel 9* ist es, die Anwendbarkeit des vorgestellten Lösungsvorschlags sowie die Erreichung des Untersuchungsziels zu belegen. Ausschnitte von DWH-Schemata multipler DWH-Systeme aus dem Umfeld des Projekts CEUS^{HB} dienen hierbei als Untersuchungsgegenstand.

In *Kapitel 10* wird eine Zusammenfassung zentraler Ergebnisse dieser Arbeit sowie ein Ausblick auf weitere bzw. neu eröffnete Forschungsfelder gegeben.

ALLGEMEINE GESTALTUNGSRICHTLINIEN

Die vorliegende Arbeit besitzt zu Beginn jedes Kapitels einen einleitenden Überblick sowie am Kapitelende eine Zusammenfassung zentraler Erkenntnisse. Zusätzlich erlaubt die Bereitstellung verschiedenartiger Verzeichnisse dem Leser einen schnellen Zugang zu spezifischen Sachverhalten.

Weitere Orientierungshilfen innerhalb der Ausführungen werden durch unterschiedliche Formatierungen bestimmter Textstellen gegeben.

- Neu eingeführte oder für einen bestimmten Sachverhalt wesentliche Begriffe werden durch eine *kursive* Schrift hervorgehoben.
- Sofern Codebeispiele angeführt werden, unterscheiden sich diese durch eine *nicht proportionale* Schrift vom fortlaufenden Text.
- Autorennamen sind anhand der Schreibweise in **VERSALIEN** zu erkennen.
- Zitate sind zum Beleg bzw. zur Ergänzung der Ausführungen als eigener, vom linken Rand zusätzlich eingerückter Absatz mit kursiver Schrift eingefügt.
- Querverweise zu anderen Abschnitten oder Abbildungen sind in runden Klammern und durch Angabe der jeweiligen Abschnitts- bzw. Abbildungsnummer gekennzeichnet.
- Auf zugrunde liegende Literaturquellen wird durch eine Signatur bestehend aus [Autorenkürzel Jahr, Seite] hingewiesen. Die zugehörigen vollständigen bibliographischen Angaben finden sich im Literaturverzeichnis.

2 Data-Warehouse-Systeme in betrieblichen Organisationen

Das vorliegende Kapitel dient der Erläuterung wesentlicher Grundlagen der Informationsversorgung in betrieblichen Organisationen. Hierzu werden zunächst Strukturtypen betrieblicher Organisationen beschrieben. Dem schließt sich eine Erörterung zu Führungsinformationssystemen an, bei der DWH-Systeme im Fokus stehen. In diesem Zusammenhang werden auch Ursachen für das Entstehen multipler DWH-Lösungen herausgearbeitet.

2.1 Betriebliche Organisationen

Betriebliche Organisationen verkörpern aus systemtheoretischer Sicht ein *offenes, sozio-technisches* und *zielgerichtetes System* (institutionelle Sicht). Sie stehen in vielfältigen Interaktionen (Lenkungs- und Leistungsflüssen) mit ihrer Umwelt (offenes System). Die betrieblichen Aufgaben¹⁹ einer Organisation werden in Kooperation von Mensch und Maschine durchgeführt (sozio-technisches System). Eine Organisation richtet ihr Verhalten auf ein Zielsystem aus, welches durch Sach- und Formalziele determiniert ist (zielgerichtetes System). Sachziele legen Art und Zweck der Leistungserstellung fest, Formalziele bestimmen Art und Umfang der Sachzielerreichung [Stae99, S. 416 f.], [FeSi06, S. 61].

Nachstehend werden Komponenten einer Organisation sowie charakteristische Organisationsstrukturen vorgestellt. Dem folgt eine vertiefende Erläuterung der divisionalen (verteilten) Organisationsstruktur. Deren Überschneidungsbereiche mit der Führungsstruktur im Hochschulwesen werden abschließend aufgezeigt.

2.1.1 Organisationsstrukturen

Charakteristika von Organisationen lassen sich in bestimmte Organisationsstrukturen gruppieren. Um eine treffende Beschreibung und Unterscheidung von Organisationsstrukturen zu ermöglichen, werden zunächst die von MINTZBERG identifizierten fünf Grundbausteine einer Organisation eingeführt [Mint79, S. 18 ff.], [Stae99, S. 482 f.]:

¹⁹ Merkmale einer *Aufgabe*, die als Zielsetzung für zweckbezogenes menschliches Handeln definiert werden kann, sind in [FeSi06, S. 91 f.] nachzulesen.

- Die *strategische Spitze* (strategic apex), auch (Top-)Management oder strategisches Management genannt, hat dafür Sorge zu tragen, dass eine Organisation ihren Auftrag in effektiver Weise erfüllt (Sachzielerfüllung) und den Anforderungen der Stakeholder (z. B. Eigentümer, Behörden, Gewerkschaften) gerecht wird (Formalzielerfüllung). Die strategische Spitze bildet das Lenkungssystem.
- Mitarbeiter, die dem *betrieblichen Kern* (operating core) angehören, verrichten so genannte Basisarbeiten, d. h. sie sind direkt mit der Fertigung von Produkten oder der Bereitstellung von Dienstleistungen betraut (Leistungserstellung).
- Die *Mittellinie* (middle line) wird auch als *Mittleres Management* bezeichnet und verbindet die strategische Spitze mit dem betrieblichen Kern über eine formale Autoritätskette von Führungskräften (operative Lenkungebene).
- Analysten, die Systeme zur Planung und Kontrolle der Arbeit entwerfen und somit eine Form der Standardisierung in der Organisation bewirken, bilden die *Technostruktur* (technostructure).
- Mitarbeiter des *Hilfsstabs* (support staff) stellen indirekte Dienste für die Organisation bereit, d. h. sie unterstützen die Organisation außerhalb ihres betrieblichen Ablaufs.

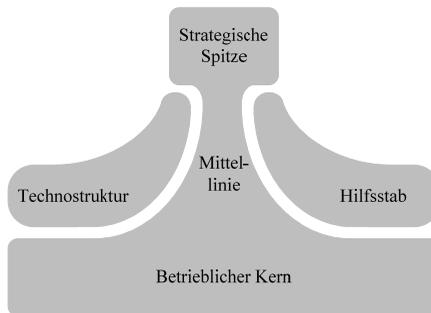


Abbildung 2-1: Die fünf Komponenten einer Organisation [Mint92, S. 28]

Reale Organisationsstrukturen setzen sich aus einer unterschiedlichen Anzahl und in ihrer Komplexität variierenden Kombination dieser Komponenten zusammen. Abhängig von der Dominanz einer Komponente in der Gesamtorganisation, bildet sich eine spezifische Konfiguration von Strukturvariablen heraus. MINTZBERG unterscheidet folgende fünf Strukturtypen [Mint79, S. 299 ff.], [Mint92], [Stae99, S. 483 ff.]:

- Die *Einfachstruktur* (simple structure) beschreibt eine Einheit, die vor allem durch die strategische Spitze sowie den betrieblichen Kern gebildet wird. Technostruktur und Hilfsstab sind kaum bzw. nicht ausgebaut. Die Koordination erfolgt in erster Linie durch persönliche Weisung. Die Einfachstruktur eignet sich aufgrund der zentralisierten Kontrolle durch die strategische Spitze gut für schnelle und flexible Innovationsvorhaben oder bei einer dynamischen Umwelt.
- Für spezialisierte, betriebliche Routineaufgaben (z. B. Massenproduktion) ist die Konfiguration *Maschinenbürokratie* (machine bureaucracy) prädestiniert. Besondere Bedeutung kommt dabei der Prozess-Standardisierung und Verhaltensformalisierung zu, für deren Ausgestaltung sich die Technostruktur verantwortlich zeigt. Die Mittellinie oberhalb des betrieblichen Kerns ist voll ausgebaut und in funktionale Einheiten gegliedert. Die Strategieentwicklung liegt bei den Führungskräften der strategischen Spitze.
- Bei der *Profibürokratie* (professional bureaucracy) führen professionelle Mitarbeiter eigenverantwortlich Aufgaben im betrieblichen Kern aus. Der betriebliche Kern stellt somit den bedeutsamsten Organisationsteil dar und wird vom Hilfsstab vornehmlich unterstützt. Die organisationale Strategie der Profibürokratie erscheint als kumulative Wirkung der Projekte der Organisationsmitglieder. Profibürokratie ist für Organisationen in komplexer und stabiler Umwelt effektiv.
- Die *divisionale* (im Folgenden auch als *verteilt* bezeichnete) *Organisationsform* (divisionalized form) setzt sich aus einer Menge unabhängiger Einheiten (Sparten, Geschäftsbereichen etc.) zusammen, die von einer losen Administration überlagert sind. Im Gegensatz zur Profibürokratie handelt es sich hierbei um Einheiten der Mittellinie, der somit eine tragende Rolle bei dieser Organisationskonfiguration zukommt. Jede Einheit funktioniert als ein in sich geschlossenes, integriertes System mit abgegrenzten Zielen. Die Kommunikation zwischen zentralem Leitungsgremium und Sparten beschränkt sich auf die Vorgabe von Leistungsstandards für die Sparten sowie die Berichterstattung der Sparten über ihre Leistungsergebnisse. Das schmal ausgelegte oberste Management bestimmt die strategische Ausrichtung. Für alle Sparten werden zentral diverse Unterstützungsdienste bereitgestellt (Technostruktur, Hilfsstab). Die verteilte Organisationsform besitzt besonders bei diversifizierten Märkten sowie verschiedenartigen Produktlinien Vorteile.

- Die Konfigurationsform *Adhokratie* (adhocracy) erlaubt es, Experten aus verschiedenen Disziplinen zu interagierenden (Ad-hoc-)Projektteams zusammenzuführen. Sie stellt eine komplexe und nicht den klassischen Managementprinzipien folgende Struktur dar. Der vorrangige Koordinationsmechanismus ist die gegenseitige Abstimmung. Die Strategie wird dabei nicht von einer übergeordneten Instanz vorgegeben, sondern ergibt sich aus der Menge der Entscheidungen, die für die Projekte getroffen werden. Dem Hilfsstab fällt in dieser Organisationsform eine hohe Bedeutung zu. Die Adhokratie eignet sich besonders für Innovationsvorhaben sowie bei komplexen und dynamischen Umweltbedingungen.

Grundsätzlich ist davon auszugehen, dass keine universell effiziente Organisationsstruktur existiert (*Situativer Ansatz*, z. B. [KiWa07, S. 43 ff.]). Eine Organisation muss ihre Struktur an die jeweils vorherrschende (Umwelt-)Situation anpassen. Situative Faktoren sind u. a. die Stabilität der Umwelt sowie die von außen einwirkenden Kräfte und Kontrollinstanzen (z. B. Umweltbedingungen, Machtfaktoren), das Stadium der strukturellen Entwicklung (Alter und Größe) einer Organisation und das technische System des betrieblichen Kerns [Mint79, S. 215 ff.], [Mint92, S. 167 ff.]. Daraus ergeben sich Kombinationen bzw. Mischstrukturen der angeführten Strukturtypen. Bei der Ausgestaltung einer Organisationskonfiguration ist die Wahrung struktureller Harmonie unerlässlich, da sie einen Eckpfeiler der Effektivität einer Organisation markiert [Mint91], [Bühn04, S. 107 ff.].

2.1.2 Verteilte Organisationen

Als Ergänzung zu den Ausführungen des vorangegangenen Abschnitts werden im Folgenden weitere Charakteristika einer verteilten Organisation herausgestellt.

KOORDINATION UND KOMMUNIKATION

Die Effektivität einer verteilten Organisation ist durch das Zusammenwirken der Sparten mit den Zentralbereichen (Technostruktur und Hilfsstab) sowie der strategischen Spitze bedingt. Die strategische Spitze trägt Verantwortung, die Sparten und Zentralbereiche zu einer handlungsfähigen Einheit zusammenzuführen. Um das Verhalten einer

(verteilten) Organisation regulieren zu können, sind Pläne oder Zielvorgaben²⁰ sowie die komplementäre Überwachung der Planerfüllung oder Zielerreichung notwendig. Die dafür erforderliche Kommunikation zwischen einer Sparte und der strategischen Spitze, bspw. in Form von Berichten über den Grad der Zielerreichung, kann durch FIS entscheidend unterstützt werden. Ohne adäquate Berichterstattung über die Leistungsergebnisse der Sparten sind Führungskräfte kaum in der Lage, Entwicklungen auf Ebene der Geschäftsbereiche sowie der Gesamtorganisation rechtzeitig zu erkennen und angemessen (re-)agieren zu können. Die nachfolgende Abbildung visualisiert die Grundstruktur sowie die Kommunikationsflüsse in einer verteilten Organisation. Weiterführende Informationen zu dieser Organisationsform bieten z. B. [KiWa07, S. 242 ff.], [Bühn04, S. 141 ff.], [Fres05, S. 458 ff.], [Mint92, S. 108 ff., S. 287 ff.].

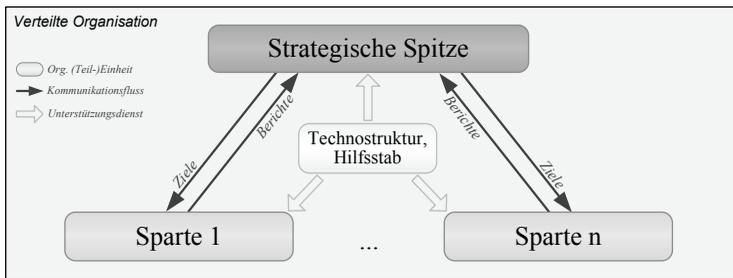


Abbildung 2-2: Grundstruktur einer verteilten Organisation (in Anlehnung an [Bühn04, S. 142])

REORGANISATION

Die verteilte Konfiguration bietet für Neu- und Umstrukturierungsmaßnahmen einer betrieblichen Organisation gute Rahmenbedingungen. Sparten können mit verhältnismäßig geringem Aufwand an- bzw. ausgegliedert werden. Eine verteilt aufgestellte Organisation ist somit in hohem Maße strategisch und strukturell anpassungsfähig. In diesem Zusammenhang sind insbesondere *Mergers & Acquisitions* (M&A)²¹ zu nennen. Der M&A-Ansatz gilt als eine entscheidende Vorgehensweise für nachhaltig profitables Wachstum einer betrieblichen Organisation und gehört inzwischen zu den grundlegen-

²⁰ Zielvorgaben parametrisieren die Formalziele einer Aufgabe [FeSi06, S. 204 ff.].

²¹ Unter dem Begriffspaar *Mergers & Acquisitions* werden die Fusion sowie die Übernahme oder Veräußerung von Unternehmen bzw. Unternehmensteilen subsumiert [Jans01, S. 43 f.], [Hark04, S. 4].

den Bestandteilen der strategischen Planung (externes Wachstum) [Jans01, S. 105 f.], [Bühn04, S. 147 ff.], [KiWa07, S. 244], [Hark04].

In der *Post-Merger-Phase* eines M&A-Projekts geht es u. a. darum, zeitnah ein Kommunikations- und Koordinationsverfahren zwischen der neu eingegliederten Sparte und der strategischen Spitze zu etablieren oder bestehende Verfahren adäquat anzupassen. Auf diese Weise können Zielvorgaben auf die neue organisationale Gesamtsituation abgestimmt und kommuniziert werden [KrSt02, S. 523 ff.], [BuMi06, S. 1065]. Abhängig von der gewählten IT-Integrationsstrategie (Absorption, Best-of-Breed, Koexistenz oder Neuentwicklung [BuMi06, S. 1066 ff.]) hat dies entscheidenden Einfluss auf die Gestaltung und somit auch auf die Homogenität der FIS-Landschaft.

RECHTLICHE AUSGESTALTUNG DER GESCHÄFTSBEREICHE

Die Gestaltung der strategischen Spitze wird von der Rechtsform beeinflusst, in der eine verteilte Organisation geführt wird. Diesbezüglich grundlegende, rechtliche Alternativen sind nachstehend aufgezeigt [Bühn04, S. 143 ff.]:

- Die Sparten werden als *rechtlich unselbständige* Unternehmensteile mit einer eigenständigen *wirtschaftlichen Selbständigkeit* in einem Einheitsunternehmen gebildet. Die Rechtsform des Gesamtunternehmens kann eine Personen- oder Kapitalgesellschaft sein. Mithilfe adäquater Planungs- und Kontrollsysteme stellt die strategische Spitze eine optimale Ressourcenallokation sowie eine einheitliche Ausrichtung der Sparten auf das Gesamtziel der Unternehmung sicher.
- Die Ausgestaltung der Sparten erfolgt als *rechtlich selbständige* Unternehmensteile im Rahmen der Rechtsstruktur eines Konzerns. Gemäß §18 Abs. 1 Aktiengesetz (AktG)²² zeichnet sich ein Konzern durch seine einheitliche Leitung gegenüber seinem bzw. seinen abhängigen Unternehmen aus. Eine weit verbreitete Variante der Geschäftsbereichsbildung ist die Management-Holding.

Die Rechtsform kann die Parameter zur Gestaltung der organisationsweiten IT-Infrastruktur entscheidend beeinflussen. Insbesondere Aspekte hinsichtlich der Granularität der Einsichtnahme in sparteneigene Daten durch die strategische Spitze oder in andere, der Organisation angehörige Sparten, sind hier zu nennen. Die Etablierung eines übergreifenden FIS kann dadurch eingeschränkt sein.

²² Das Aktiengesetz (AktG) kann unter <http://bundesrecht.juris.de/aktg> eingesehen werden (Abruf am 10.06.2008).

2.1.3 Hochschulwesen als verteilte Organisation

Die im vorherigen Abschnitt erörterte Konfiguration einer verteilten Organisation lässt sich auch der Führungsstruktur im Hochschulwesen zugrunde legen.

Die Hochschulen eines Landes unterstehen dem Management²³ der Landesregierung in Form des zuständigen *Wissenschaftsministeriums* (Strategische Spitze; vgl. Abschnitt 2.1.1). Dieses vereinbart mit dem Management einer Hochschule Zielvereinbarungen und überwacht deren Erfüllung anhand von Rechenschaftsberichten zur Hochschulsituation, die vom *Hochschulmanagement* übermittelt werden. Die Operationalisierung der Zielvorgaben obliegt dem Management einer Hochschule (Mittellinie).

Eine Hochschule umfasst im Allgemeinen zwei Hauptprozesse, die Leistungen erzeugen und an Nachfrager in der Umwelt der Hochschule übergeben:

- Der Hauptprozess *Studium und Lehre* erzeugt Ausbildungs- und Prüfungsleistungen für Studierende.
- Der Hauptprozess *Forschung* erzeugt Forschungsleistungen, die Forschungspartnern oder der interessierten Öffentlichkeit bereitgestellt werden.

Die Ausgestaltung und Bereitstellung des Leistungsspektrums von *Studium und Lehre* sowie von *Forschung* erfolgt durch die Fakultäten oder Fachbereiche einer Hochschule (Betrieblicher Kern). Das *Fakultätsmanagement*, das dem Hochschulmanagement untersteht, koordiniert und begutachtet die Leistungserbringung einer Fakultät.

Das Hochschulmanagement ist zudem für das Management der Serviceprozesse einer Hochschule zuständig (z. B. *Personalwirtschaft*, *Mittelbewirtschaftung*). Serviceprozesse stellen ihre Leistungen Hauptprozessen und weiteren Serviceprozessen zur Verfügung. Abbildung 2-3 visualisiert die strukturellen Beziehungen im Hochschulwesen [Sinz95], [SiBU99], [SBPU01].

²³ *Management* wird hier im Sinne von Führen und Gestalten sowie unabhängig von der konkreten hochschulspezifischen Implementierung der Entscheidungsträger verstanden [SiBU99, S. 2].

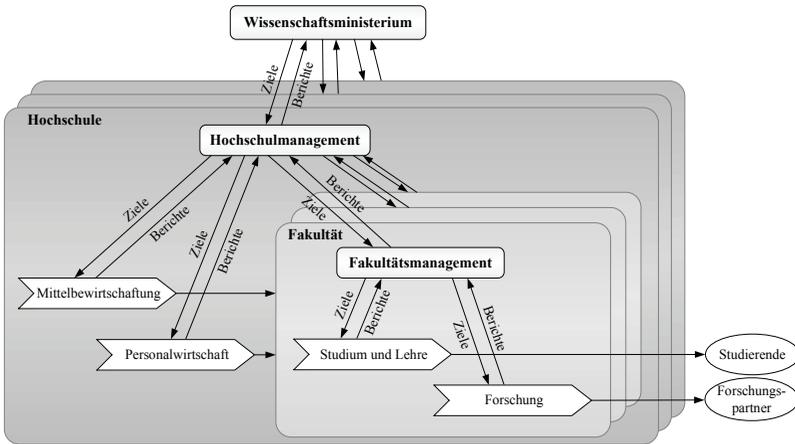


Abbildung 2-3: Führungsstruktur im Hochschulwesen [SiBU99, S. 3]

Das Hochschulwesen eines Landes weist zudem folgende Kennzeichen einer verteilten Organisation auf:

- *Autonomie der Hochschulen:* Eine Hochschule agiert bei der Realisierung ihrer Zielvorgaben weitgehend autonom und steht mit anderen Hochschulen im Wettbewerb um Studierende und Forschungspartner sowie um ihre Positionierung.
- *Strategische Ausrichtung:* Strukturpolitische Richtlinien und die Verteilung von Personal- und Sachmitteln erfolgen durch zentrale Vorgaben auf Landesebene. Innerhalb einer Hochschule gestalten und erbringen die Fakultäten ihre Leistungsprozesse zumeist eigenständig. Sie kooperieren im Rahmen gemeinsamer Studienangebote und Forschungsprogramme. Zentrale Entscheidungen durch das Hochschulmanagement betreffen auch hier strukturpolitische Vorgaben und die intra-universitäre Mittelverteilung [Sinz95, S. 66 f.], [SBPU01].

Es konnte dargelegt werden, dass wesentliche Charakteristika einer verteilten Organisation sich auch in der Führungsstruktur des Hochschulwesens wieder finden. Auch der Kommunikationsfluss in Form von Zielvorgaben und entsprechenden Rechenschaftsberichten zwischen Wissenschaftsministerium und Hochschulmanagement ist äquivalent zur Abstimmung der organisationalen Ausrichtung zwischen strategischer Spitze und Sparten. Die genannten Rahmenbedingungen zur Implementierung eines ganzheitlichen FIS gelten aus diesen Gründen analog.

2.2 Führungsinformationssysteme

Die Ressource *Information* (vgl. Abschnitt 4.1) gilt heute als bedeutender Produktionsfaktor einer betrieblichen Organisation [BaGü04, S. 5], [PiFr88]. Insbesondere der gestiegene (internationale) Wettbewerb, die Dynamik der Märkte sowie deutlich kürzere Produktlebenszyklen sind dafür verantwortlich. Fundierte, organisationsintern und -extern erhobene Informationen sind zu einem unverzichtbaren Baustein einer gesamtorganisationalen Lagebewertung geworden. Die Integration und bedarfsgerechte Aufbereitung der Informationen kann durch *Führungsinformationssysteme* (FIS) unterstützt werden. Die folgenden Abschnitte beleuchten die Motivation für die Entwicklung von FIS sowie eine begriffliche Abgrenzung.

2.2.1 Motivation und historische Entwicklung

Das Treffen von Entscheidungen im Rahmen des Führungsprozesses besitzt Tragweite für die Fortentwicklung einer Organisation. Die für die Vorbereitung von Entscheidungen erforderliche Zeitspanne wird oftmals – z. B. aufgrund zunehmender Umweltkomplexität – größer, die zeitgleich steigende Dynamik der Umwelt verlangt jedoch nach immer kürzeren Reaktionszeiten (*Zeitschere*) [GrZa92, S. 10 ff.]. Einer zielgerichteten Informationsversorgung des Entscheidungsträgers kommt somit ein hoher Stellenwert zu. Die Verfügbarkeit fundierter und bedarfsorientierter Informationen beeinflusst transitiv die Qualität einer Entscheidung [Back91]. Eine auf die Informationsversorgung von Entscheidungsträgern ausgerichtete IT-Infrastruktur kann hier einen Unterstützungsbeitrag leisten [Horv06, S. 3], [Holt97, S. 3 ff.], [Öste95, S. 1 ff.].

ANFORDERUNGEN AN EINE IT-BASIERTE INFORMATIONSVERSORGUNG

Entscheidungsträger sollen im Moment des Informationsbedarfs einen adäquaten Zugang zu einer validen sowie homogenen Datengrundlage als Ausgangspunkt für eine bedarfsorientierte Informationsversorgung besitzen. Die Aufbereitung und Integration sowohl organisationsinterner als auch -externer Daten sind dafür eine notwendige Voraussetzung. Zudem ist das Formalziel Bereitstellung personenbezogener, problem- und situationsgerechter Informationen und damit die Vermeidung einer Informationsüberflutung (*Information Overload*) des Nachfragers zu erfüllen. Eine bloße Bereitstellung von Informationen durch ein IT-System ist unzureichend [HoRS01, S. 14], [Gabr99, S. 420 f.], [BeSc93], [BuFK92], [Acko67].

Nachstehende Tabelle gibt einen Überblick über grundlegende Anforderungen an eine IT-basierte Informationsversorgung.

| Anforderung | Erläuterung |
|--|---|
| (Automatisierbares) Zusammenführen von Daten aus unterschiedlichen Quellen | Über Schnittstellen zu Datenbeständen interner (Fach-)Abteilungen oder externen Quellen können Daten nach Bedarf oder in regelmäßigen Abständen extrahiert werden. |
| Datenspeicherung in gemeinsamer Datenbasis mit einheitlichem Zugang | Die aus unterschiedlichen Quellen stammenden Daten werden in einer zentralen Datenbasis zusammengeführt und verwaltet. |
| Benutzerfreundliche Bedienung | Berichte sollen ohne Kenntnis einer Programmiersprache, möglichst graphisch unterstützt, erstellt werden können. Eine intuitive Bedienung fördert die Akzeptanz des Systems bei Endanwendern. |
| Sichtenspezifischer und stufenweise aggregierbarer Informationsabruf | Die Informationen können in unterschiedlichem Detail- bzw. Verdichtungsgrad sowie in der jeweils bevorzugten Form (Tabelle, Graphik etc.) kompakt und anschaulich bereit gestellt werden. |
| Ausnahmeberichtswesen | Das Über- oder Unterschreiten bestimmter Grenzen für Kenngrößen soll in Berichten hervorgehoben werden können. |
| Weiterverarbeitungsfunktionen | Berichte sollen automatisch an ihre Empfänger weitergeleitet und von diesen kommentiert oder bearbeitet werden können. |
| Werkzeugcharakter | Der individuelle Aufbau eines Systems zur Informationsbeschaffung und -versorgung des Managements soll durch bedienerfreundliche Entwicklungsumgebungen unterstützt werden. |

Tabelle 1: Anforderungen an eine IT-basierte Informationsversorgung [Back91]

REALISIERUNGSKONZEPTE

Die Vision einer IT-gestützten Informationsaufbereitung und -bereitstellung für Entscheidungs- und Planungsaufgaben kam erstmals in den 1950er Jahren mit der Diskussion um *Managementinformationssysteme* (MIS) auf [LeWh58], [Holt99a, S. 30]. Forderten zunächst Controlling-Abteilungen und später die obere Führungsebene eine fundierte, bedarfsorientierte und schnelle Informationsversorgung, so postulieren heute Mitarbeiter aus allen Unternehmensbereichen Zugang zum Produktionsfaktor Information [ChGl06, S. 4], [BeMu98, S. 12]. Ein Informationsvorsprung kann für den Erfolg auf einzel- und gesamtwirtschaftlicher Ebene entscheidend sein [PiRW03, S. 60 ff.], [PiFr88].

[...] a new direction and strategy must be one that promotes and encourages the use of data and makes the required information available to every person in a particular interest group who has need of it [DeMu88, S. 60].

Um den oben aufgeführten Anforderungen einer IT-basierten Informationsversorgung für Entscheidungsträger gerecht zu werden, wurden unterschiedliche Konzepte vorgeschlagen. Hier sind neben MIS vor allem *Decision Support Systems* (DSS)/*Entscheidungsunterstützungssysteme* (EUS) und *Executive Information Systems* (EIS)/*Führungsinformationssysteme* (FIS) zu nennen. Beginnend bei zeitdiskret generierten Standardberichten ohne Verdichtung operativer Daten (MIS) führte die Entwicklung über zielgruppenbezogene Dialogsysteme und adaptive Entwicklung von Entscheidungsmodellen (EUS) hin zur interaktiv und subjektbezogenen Informationsbereitstellung (FIS)²⁴. Mit jeder Generation eines so genannten *Managementunterstützungssystems* (MUS) stieg auch die Unterstützung der Entscheidungsträger. Nachstehende Abbildung veranschaulicht diesen Zusammenhang anhand der chronologischen Entwicklung. Dabei gilt nicht, dass die Techniken eines bestimmten Konzepts grundsätzlich auch in den zeitlich später aufgetretenen Konzepten enthalten sind.

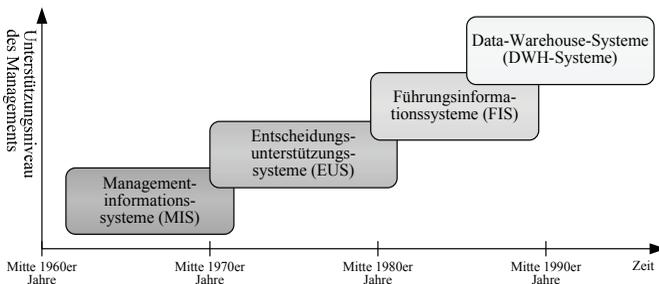


Abbildung 2-4: Entwicklungsphasen managementunterstützender Systeme
(in Anlehnung an [WiDV04, S. 24], [BuFK92, S. 7])

²⁴ Eine ausführliche Beschreibung der genannten Konzepte findet sich in [GIGC97, S. 149 ff.], [BeMu98, S. 15 ff.] und [Holt99a, S. 29 ff.].

Fehlende technische Voraussetzungen wie bspw. schnelle Kommunikationstechnologien, graphische Benutzeroberflächen, ausreichende sowie kostengünstige Datenspeicher und leistungsfähige Prozessoren ließen die frühen Ansätze scheitern. Die Behebung technischer und infrastruktureller Mängel ermöglichte es schließlich, mit dem Ansatz des *Data-Warehouse* (DWH; vgl. Abschnitt 2.3) das angestrebte Ziel einer umfassenden, IT-basierten Informationsversorgung der Entscheidungsträger annähernd zu erreichen [WiDV04, S. 24 ff.], [GlKe06], [ChGl06, S. 6 ff.], [StHa05a, S. 382 f.], [BaGü04, S. 11 f.], [RoTr82], [Dev197, S. 7 ff.].

2.2.2 Begriffsbestimmung und Abgrenzung

Führungsinformationssysteme (FIS) dienen dazu, Führungskräften relevante Informationen rechtzeitig und in geeigneter Form zur Verfügung zu stellen. Es gilt somit die pragmatische Forderung nach der richtigen Information zum richtigen Zeitpunkt, in der richtigen Menge, am richtigen Ort und in der erforderlichen Qualität zu erfüllen. Die Informationsversorgung eines Entscheidungsträgers (vgl. Abschnitt 4.2) gilt dann als optimal. In der Literatur wird in diesem Zusammenhang auch vom logistischen Prinzip für Informationen gesprochen [StHa05a, S. 382], [Holt99a, S. 29 ff.], [BeSc93], [Augu90, S. 23 f.], [RoTr82]. Folgende Tabelle bietet dazu eine Erläuterung:

| Merkmal | Erläuterung |
|---------------------------------|--|
| Die richtige Information ... | vom Empfänger benötigt und verstanden |
| zum richtigen Zeitpunkt ... | für die Fällung von Entscheidungen rechtzeitig |
| in der richtigen Menge ... | „so viel wie nötig, so wenig wie möglich“ |
| am richtigen Ort ... | beim Empfänger verfügbar |
| in der erforderlichen Qualität. | ausreichend detailliert und wahr, unmittelbar verwendbar |

Tabelle 2: Logistisches Prinzip für Informationen [Augu90, S. 23]

FIS unterstützen vornehmlich Entscheidungsprozesse bei Führungsaufgaben. Eine mehrheitlich anerkannte Strukturierung der im Kontext von FIS genannten Begriffe und Systeme offeriert folgende Abbildung. Es wird von dem jeweils aktuellen technologischen Entwicklungsstand der aufgeführten Systeme ausgegangen.

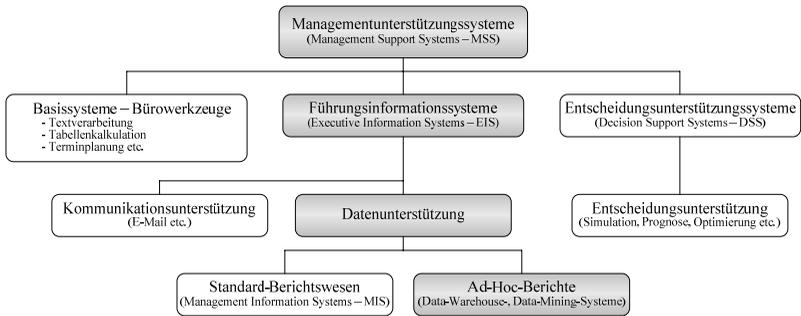


Abbildung 2-5: Komponenten von Managementunterstützungssystemen (in Anlehnung an [ChGl06, S. 9], [StHa05a, S. 383])

Nachstehend sind die Elemente der Abbildung 2-5 erläutert:

- Unter dem Begriff *Managementunterstützungssysteme* werden alle Formen von Systemen zusammengefasst, die Entscheidungsträger bei ihrer Aufgabendurchführung unterstützen.
- *Führungsinformationssysteme* dienen der Bereitstellung interner und externer Informationen für Manager. Neben der Kommunikationsunterstützung ist hier die Unterstützung bei der Informationsgewinnung (Datenunterstützung) der zentrale Aspekt. *Managementinformationssysteme* offerieren Informationen in Form von Standardberichten, *Data-Warehouse-* und *Data-Mining-Systeme*²⁵ bieten Möglichkeiten des Ad-Hoc-Reportings.
- Modelle und Methoden für Planungs- und Kontrollprozesse werden Führungsverantwortlichen durch *Entscheidungsunterstützungssysteme* angeboten.
- *Basissysteme* umfassen Werkzeuge für allgemeine Büroarbeiten (Textverarbeitung, Tabellenkalkulation etc.).

Detaillierte Ausführungen zu den genannten Varianten von MUS sowie weiterführende Informationen zu FIS finden sich bspw. bei [ChGl06, S. 9], [StHa05a, S. 382 ff.], [Holt99a, S. 29 ff.], [BeSc93], [RoTr82]. Die in obiger Abbildung grau unterlegten Objekte markieren den Fokus der vorliegenden Arbeit: *DWH-basierte FIS*.

²⁵ *Data-Mining-Systeme* dienen der Mustererkennung in Datenbeständen. Eine Einführung in grundlegende Konzepte und Techniken des Data-Mining bietet z. B. [HaKa06]. Synergieeffekte zwischen Data-Warehouse- und Data-Mining-Systemen werden u. a. in [BöKU03] diskutiert.

2.3 Data-Warehouse-Systeme

Daten können operative Belange sehr effizient unterstützen. Sie führen jedoch nicht zwangsläufig zu Informationen, die Wissen verändern und Prozesse optimieren. Eine große Herausforderung besteht darin, verdeckte Informationen aus den operativen Systemen für Entscheidungsprozesse offen zu legen. Zudem bestehen operative Daten in einer Form, die für Führungsverantwortliche weder leicht zugänglich noch intuitiv verständlich ist. Speziell zur Erfüllung dieser Anforderungen wurde das Konzept des *Data-Warehouse* (DWH) vorgeschlagen.

*DWH-Systeme*²⁶ stellen eine zentrale Komponente eines FIS dar [Sinz02, S. 311]. Sie ermöglichen es, Endanwendern qualitativ hochwertige Daten bedarfs- und zeitgerecht bereitzustellen. In betrieblichen Organisationen und Behörden sind DWH-Systeme zu einem festen Bestandteil der operativen und strategischen Planung geworden [LeJa99, S. 7-1], [Inmo05, S. 457 ff.], [BaGü04, S. 13 ff., S. 483], [WiDV04, S. 69].

Nach der Abgrenzung des DWH-Begriffs wird im Folgenden die Architektur eines DWH-Systems vorgestellt. Daran schließt sich die Einordnung eines DWH-Systems in eine betriebliche Organisation an. Die weiteren Ausführungen beinhalten eine Diskussion grundlegender Aspekte bei der DWH-Einführung sowie Besonderheiten von DWH-Systemen in verteilten Organisationen. Abschließend wird ein Überblick über Varianten und Erweiterungen des klassischen DWH-Konzepts gegeben.

2.3.1 Begriffsbestimmung und Abgrenzung

Der Ursprung des heute bekannten DWH-Konzepts findet sich in dem von IBM initiierten Projekt *Europe/Middle East/Africa Business Information System* (EBIS) [DeMu88]. Mit dem EBIS-Projekt wurde erstmals eine zentrale Bereitstellung aller entscheidungsrelevanten Informationen, die in einer Organisation elektronisch gespeichert sind, angestrebt. Als wesentliche Neuerung zu früheren Ansätzen managementunterstützender Anwendungssysteme galt dabei die, von den operativen Systemen losgelöste, redundante Datenhaltung. Diesem Paradigma folgend, wurde ein DWH als eine von den operativen Systemen getrennte, integrierte und fachspezifische Datenbasis zur Entscheidungsunterstützung für beliebige Nachfrager konzipiert [Inmo93, S. 29 ff.], [Holt99a, S. 39 ff.], [Sinz02, S. 309 f.], [WiDV04, S. 26].

²⁶ Eine Differenzierung zwischen DWH und DWH-System erfolgt in Abschnitt 2.3.2.

Eine anerkannte DWH-Definition stammt von INMON:

The data warehouse [...] is the foundation of all DSS processing. [It] is a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management's decisions [Inmo93, S. 29].

Folgende vier Eigenschaften sind demnach charakteristisch für ein DWH [Inmo93, S. 29 ff.], [BöU100, S. 16 f.], [MuBe00, S. 9 ff.]:

- *Subject-oriented*: Der Aufbau und Inhalt eines DWH ist auf die für die Unternehmensführung relevanten Entscheidungsfelder ausgerichtet.
- *Integrated*: In einem DWH werden Daten aus unterschiedlichen, sowohl internen als auch externen Datenquellen integriert und konsolidiert.
- *Non-volatile*: In das DWH geladene Daten werden nicht mehr verändert oder gelöscht. Zu definierten Zeitpunkten werden die Inhalte eines DWH sukzessive erweitert.
- *Time-variant*: Die Daten eines DWH werden historisiert gespeichert, sodass zeitraumbezogene Analysen durchgeführt werden können.

ONLINE-ANALYTICAL-PROCESSING

Die Nutzung eines DWH folgt dem Konzept des *Online Analytical Processing* (OLAP)²⁷, welches dem Entscheidungsträger eine interaktive Datenanalyse ermöglicht [Sinz02, S. 311]. Die spezifischen Anfrageoperatoren und -techniken, die den OLAP-Begriff charakterisieren, werden dazu in multidimensionalen Datenstrukturen abgebildet. Die Grundidee multidimensionaler Datenstrukturen beruht auf der Unterscheidung von qualitativen und quantitativen Daten [Shos82, S. 208 ff.]. Quantitative Größen (z. B. Fakten, Kennzahlen) werden nach unterschiedlichen qualitativen Gesichtspunkten (Dimensionen) aufgeschlüsselt. Eine fundierte Einführung in das multidimensionale Datenmodell geben u. a. [Böhn01, S. 126 ff.] und [BaGü04, S. 102 ff.].

²⁷ Demgegenüber unterliegen operative Systeme dem Verarbeitungskonzept *Online Transaction Processing* (OLTP). Zur Abgrenzung der Verarbeitungskonzepte OLTP und OLAP siehe z. B. [BöU100, S. 2 f.].

Die OLAP-Technologien sind zudem nach den zugrunde liegenden Speicherstrukturmodellen der Datenablage im DWH (z. B. [BöU100, S. 26 f.]) zu differenzieren:

- Beim *multidimensionalen OLAP* (MOLAP) werden multidimensionale Datenbanksysteme eingesetzt.
- *Relationales OLAP* (ROLAP) transformiert die multidimensionalen Anfragen und Operationen in Konstrukte relationaler Anfragesprachen. Somit kann die relationale Datenbanktechnologie²⁸ als Speichermedium genutzt werden.
- *Hybrides OLAP* (HOLAP) verknüpft die jeweiligen Vorteile von MOLAP und ROLAP.

Für eine ausführliche Erläuterung des OLAP-Konzepts wird insbesondere auf folgende Literaturquellen hingewiesen: [GlCh06], [BöU100, S. 1 ff.] und [ChDa97].

2.3.2 Architektur eines Data-Warehouse-Systems

Die Architekturbeschreibung eines DWH-Systems umfasst insbesondere die Erläuterung der einzelnen Systemkomponenten hinsichtlich ihrer Art, ihrer funktionalen Eigenschaften und ihres Zusammenwirkens [WiDV04, S. 27], [KeMU06, S. 19]. Die logische Architektur eines DWH-Systems spiegelt Abbildung 2-6 wider.

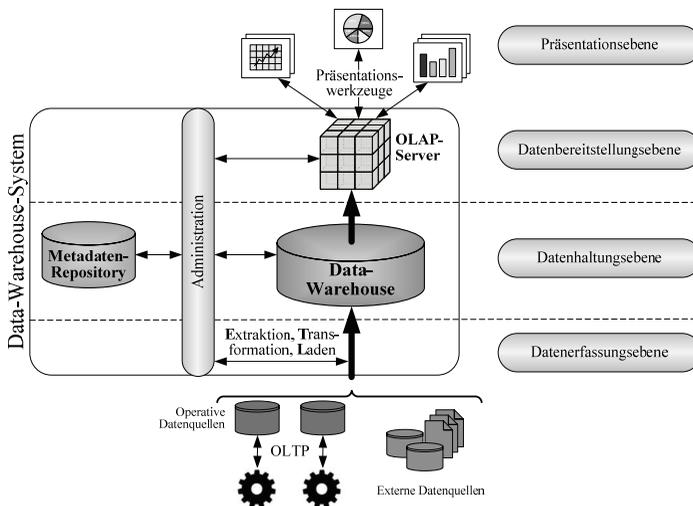


Abbildung 2-6: Architektur eines DWH-Systems [BöU100, S. 17]

²⁸ Eine eingehende Einführung in das relationale Datenmodell bietet z. B. [Voss08, S. 95 ff.].

Ein DWH-System lässt sich in die Ebenen *Datenerfassung*, *Datenhaltung*, *Datenbereitstellung* sowie *Präsentation* gliedern. Auf der Datenerfassungsebene werden die relevanten Daten aus den Vorsystemen extrahiert, transformiert und in eine zentrale Datenbasis – das eigentliche DWH – geladen (*ETL-Prozess*; z. B. [WiDV04, S. 38 ff.], [KeMU06, S. 23 ff.]). Dort erfolgt die permanente Speicherung der Daten (Datenhaltungsebene). Die Datenhaltung eines DWH wird in unterschiedliche Organisationsformen differenziert (z. B. zentrales oder verteiltes DWH, Data Marts²⁹). Der auf der Datenbereitstellungsebene durch einen OLAP-Server unterstützte Analyseprozess operiert auf den Daten des DWH und stellt diese über Präsentationswerkzeuge den Endanwendern zur Verfügung (Präsentationsebene).

Allen Ebenen des DWH sind *Administrationsfunktionen* zugeordnet, welche durch ein *Metadaten-Repository* (vgl. Abschnitt 3.2) unterstützt werden. Zu einem DWH-System zählen somit das DWH-Managementsystem mit Extraktions-, Bereinigungs-, Datenbereitstellungs- und Administrationsfunktionen als auch das eigentliche DWH [ChDa97], [BöU100, S. 17 ff.], [Sinz02, S. 315]. Detaillierte Erläuterungen zur Architektur eines DWH-Systems offerieren auch [BaGü04, S. 34 ff.] und [KeMU06, S. 19 ff.].

Die konkrete Ausgestaltung der Architektur eines DWH-Systems wird durch die individuellen Anforderungen der beteiligten Organisationsbereiche sowie bestehende IT-Rahmenbedingungen beeinflusst. Das DWH-System ist zudem in die Führungs-, Arbeits-, Informations- und Kommunikationsprozesse der zukünftigen Systembenutzer zu integrieren [WiDV04, S. 27], [BuFK92, S. 10]. Diese Sachverhalte tragen zum Entstehen isolierter und sowohl auf systemtechnischer als auch auf fachlicher Ebene heterogener DWH-Systeme bei.

²⁹ *Data Marts* werden eingesetzt, um bereichs- oder problemspezifische Analyseszenarien zu realisieren. Data Marts können als Ausschnitt eines konsolidierten DWH (abhängige Data Marts) oder als eigenständige (Teil-)DWH-Lösung (unabhängige Data Marts) realisiert werden. Besonders bei unabhängigen Data Marts kann sich deren Unterhalt aufwändig gestalten und aufgrund unterschiedlicher Datenschemata oder Aktualisierungszyklen zu Inkonsistenzen führen. Zudem gestaltet sich eine Konsolidierung inependenter Data Marts und deren Zusammenführung in ein zentrales DWH schwierig [BaGü04, S. 59 ff.].

2.3.3 Einordnung eines Data-Warehouse in eine betriebliche Organisation

Die Positionierung eines DWH-Systems innerhalb einer betrieblichen Organisation wird nachfolgend anhand einer regelkreisorientierten Betrachtung eines betrieblichen Systems (z. B. Geschäftseinheit, Unternehmen, Holding) beleuchtet (vgl. Abbildung 2-7).

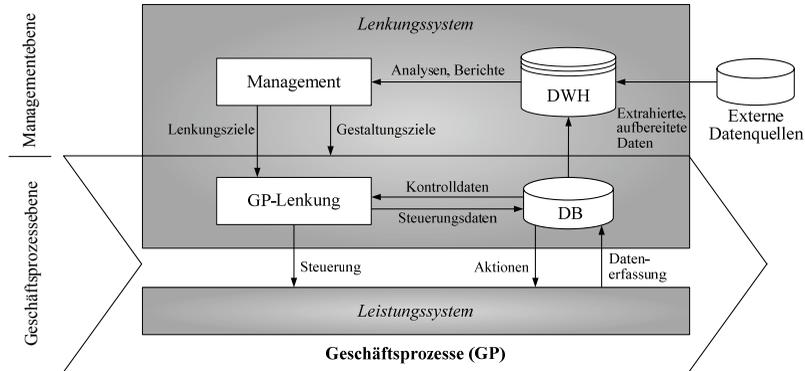


Abbildung 2-7: Regelkreisorientierte Einordnung eines DWH [Sinz02, S. 309]

Gemäß dem Phasenprinzip besteht ein betriebliches System aus den beiden Teilsystemen Leistungserstellung (*Leistungssystem*) und Lenkung der Leistungserstellung (*Lenkungssystem*) [FeSi06, S. 6]. Das Lenkungssystem (Regler) kann wiederum in zwei Ebenen von Reglern unterteilt werden:

- Im Rahmen der laufenden Geschäftsprozesse (GP) hat die GP-Lenkung (Regler) die Aufgabe der direkten Planung, Steuerung und Kontrolle des Leistungssystems (Regelstrecke). GP-Lenkung und Leistungssystem sind der GP-Ebene eines betrieblichen Systems zugeordnet.
- Dem Management³⁰ (Regler) obliegt die Aufgabe der Lenkung und Gestaltung des gesamten betrieblichen Systems. Über Zielvorgaben erfolgen die Lenkung der GP und die Gestaltung der Gesamtheit der GP (Regelstrecke).

Die jeweiligen Regler (GP-Lenkung und Management) erfassen die Zustände und Strukturen der zugehörigen Regelstrecke indirekt über eine Hilfsregelstrecke in Form

³⁰ Die verwendete Bezeichnung *Management* ist hier mit dem Ausdruck *strategische Spitze*, der Terminus *GP-Lenkung* mit dem Begriff *Mittellinie* äquivalent (vgl. Abschnitt 2.1.1).

eines Modells³¹ der jeweiligen Regelstrecke. Auf GP-Ebene stellt die Datenbank (DB), auf Managementebene das DWH die Hilfsregelstrecke dar. Das DWH enthält aus den operativen Datenbanken extrahierte und aufbereitete Daten. Darüber hinaus kann es mit Daten aus externen Quellen angereichert werden [Sinz02, S. 310 f.].

2.3.4 Aspekte einer Data-Warehouse-Einführung

Zentrales Erfolgskriterium bei der Einführung eines DWH-Systems ist der Nutzen für den Anwender [ChGI06, S. 12]. Im Folgenden werden Gründe für die Implementierung sowie das Nutzenpotential eines DWH-Systems diskutiert. Abschließend werden kritische Faktoren eines DWH-Einführungsprojekts erörtert.

MOTIVATION EINER DATA-WAREHOUSE-IMPLEMENTIERUNG

Betriebliche Organisationen sehen sich einer steigenden Datenflut gegenüber und beklagen gleichzeitig ein hohes Informationsdefizit, da es ihnen nicht gelingt, die Vielzahl ihrer Daten in geeigneter Weise zu nutzen [BeMu98, S. 9], [Lehm01, S. 4]. Ein DWH kann bei der Aufbereitung und Konsolidierung großer Datenmengen und somit bei der Erzielung einer organisationsweiten Datenintegrität beitragen. Die wachsende Bedeutung der Daten für eine betriebliche Organisation, das hohe Datenvolumen und die Vielzahl unterschiedlicher Datenquellen sowie die Forderung nach einer effizienten Entscheidungsunterstützung zählen daher mit zu den wichtigsten Faktoren, die für eine DWH-Investition sprechen [Pott98, S. 92 ff.], [FrWe00, S. 32 ff.], [BaGü04, S. 13 f.].

Diese Aspekte belegt auch eine Studie, bei der Unternehmen hinsichtlich der Erzielung von Wettbewerbsvorteilen durch den Einsatz von DWH-Systemen befragt wurden. Insbesondere die nachstehenden Argumente wurden dabei als Gründe für die Implementierung eines DWH-Systems herausgestellt [WiDV04, S. 79 ff.]:

³¹ Ein *Modell* ist ein System, welches ein anderes System zielorientiert abbildet. Ein Modell M besteht aus dem 3-Tupel $M = (S_O, S_M, f)$, wobei S_O ein (zu modellierendes) Objektsystem über der Systemträgermenge V_O , S_M ein Modellsystem über der Systemträgermenge V_M und $f: V_O \rightarrow V_M$ eine Modellabbildung ist [FeSi06, S. 20].

- Integration heterogener Datenbestände in ein unternehmensweites DWH
- Realisierung eines Berichtswesens mit zentralem Charakter
- Mangelnde Reporting- und Analysefähigkeit bestehender IT-Systeme, die den aktuellen Anforderungen nicht mehr gerecht werden (Ablösung des Altsystems)
- Zugriff auf bisher nicht verfügbare Daten
- Verbesserung der Transparenz
- Realisierung anwendungsübergreifender Auswertungen bzw. eines modulübergreifenden Berichtswesens

Insbesondere die beiden erstgenannten Gesichtspunkte zielen auf eine einheitliche Informationsversorgung hin. Aufgrund der Tatsache, dass in der Praxis multiple DWH-Lösungen existieren, ist zu konstatieren, dass diese Forderung bis dato nur in geringem Maße realisiert werden konnte. Auch Untersuchungen zum Reifegrad von DWH-Initiativen (*Data Warehousing Maturity Model*) belegen dieses [Ecke04a]. Die Vollständigkeit einer Datenbasis in einem (zentralen) DWH gilt jedoch als ein notwendiges Kriterium zur Steigerung der Entscheidungsqualität [Pott98, S. 96]. Die Überwindung der semantischen Heterogenität zwischen DWH-Systemen kann folglich einen wesentlichen Beitrag zur Realisierung einer einheitlichen Informationsversorgung leisten.

NUTZENPOTENTIALE

Eine DWH-Implementierung kann vielfältige Chancen- und Nutzenpotentiale für eine betriebliche Organisation mit sich bringen. Eine umfassende Aufstellung und Diskussion hierzu bieten u. a. [Pott98, S. 95 ff.], [WiDV04, S. 10 ff.], [BaGü04, S. 433 ff.]. Ohne Anspruch auf Vollständigkeit sollen an dieser Stelle einige im Einklang mit dieser Arbeit stehende Nutzenaspekte vorgestellt werden:

- Die Integration von Daten aus unterschiedlichen Datenquellen ermöglicht neue Kombinations- und Auswertungsmöglichkeiten des bestehenden Zahlenmaterials. Zudem trägt die Bereitstellung konsolidierter Daten in einem zentralen DWH zur Verbesserung der Informationsqualität bei. Ein direkter und flexibler Zugriff über mächtige Analysewerkzeuge auf die Datenbasis eines DWH erlaubt die Informationsversorgung der Endanwender effektiv und effizient zu gestalten. Der Wert eines DWH kann zudem durch Anreicherung mit externen Daten gesteigert werden [Jone98, S. 46, S. 50], [Pott98, S. 95 ff.], [FrWe00, S. 33].

- Ein DWH kann sowohl die ihm vor- als auch nachgelagerten Systeme positiv beeinflussen. Durch den Rückfluss der Ergebnisse an die daten-produzierenden Organisationseinheiten kann eine Verbesserung der Datenqualität in den operativen Systemen erzielt werden. Ein DWH dient zudem als konsolidierte Datenbasis für die Weiterverarbeitung der Daten durch andere Anwendungen, bspw. aus den Bereichen *Customer Relationship Management* (CRM) oder *Supply Chain Management* (SCM). Abhängig von der Ausgestaltung und Qualität eines DWH-Systems können die angeschlossenen Systeme entsprechend profitieren [WiDV04, S. 116, S. 127].
- Die Daten eines DWH werden in Form von anwendungs- und fachbezogenen Informationsobjekten repräsentiert. Die Komplexität der Datenstrukturen eines DWH bleibt für den Endanwender somit durch die semantische Schicht der eingesetzten Terminologie verborgen [LeJa99, S. 7-1 f.]. Die Einführung eines DWH kann daher auch die Homogenisierung von betriebswirtschaftlichen Begriffen initiieren.

Um die hier im ersten Punkt angesprochene Verbesserung der Informationsversorgung und -qualität in umfassender Form erreichen zu können, ist die Zusammenführung der Daten in einem zentralen DWH erforderlich. Ebenso kann der Nutzen, den dem DWH nachgeordnete Anwendungen bieten (Punkt zwei), nur dann vollständig realisiert werden, wenn die benötigten Daten nicht auf mehrere, heterogene DWH-Systeme verteilt sind. Zudem sollte einem DWH-System ein homogenes Begriffssystem zugrunde liegen (Punkt drei). Analog gelten hier die Vorteile einer zentralen DWH-Lösung. Viele dem DWH originäre Nutzenpotentiale lassen sich folglich nur bei Implementierung eines zentralen, holistischen DWH in vollem Umfang verwirklichen.

NUTZENBEWERTUNG

Zu ergänzen ist, dass es im Vergleich zu konventionellen IT-Investitionsprojekten bei einer DWH-Implementierung noch schwieriger ist, den Nutzen quantifiziert darzulegen. Besonders der qualitative Charakter der mit der Zielformulierung in Verbindung stehenden Nutzenpotentiale unterbindet eine Wirtschaftlichkeitsanalyse nach quantitativen Kostengrößen. Darüber hinaus handelt es sich bei einem DWH-Projekt im Allgemeinen um ein Kooperationsvorhaben verschiedener (Fach-)Abteilungen und des IT-Bereichs, sodass aufgrund vorherrschender Leistungsverflechtungen keine exakte Kosten-Nutzen-

Bestimmung durchgeführt werden kann. Auch die Tatsache der fehlenden Nutzenquantifizierung trägt dazu bei, dass Vorhaben zur Einführung eines zentralen DWH-Systems unterlassen werden [Wint04, S. 318], [FrWe00, S. 33 f.], [WiDV04, S. 115], [BaGü04, S. 433 f.].

KRITISCHE FAKTOREN EINER DATA-WAREHOUSE-EINFÜHRUNG

Projektrisiken sowie Erfolgsfaktoren einer DWH-Einführung wurden in diversen Studien und Veröffentlichungen umfassend analysiert (z. B. [Ditt99], [Kurz99, S. 306 ff.], [AdMo00, S. 77 ff.], [WiDV04, S. 109 ff.], [BuKD06, S. 42 ff.]). Eine Zusammenfassung erfolgskritischer Faktoren der DWH-Einführung ist nachstehend gegeben:

- *Organisatorische Auswirkungen:* Die Einführung eines DWH-Systems löst sowohl organisatorische als auch infrastrukturelle Veränderungen innerhalb einer betrieblichen Organisation aus. Zudem ist festzuhalten, dass der Aufbau eines DWH meist nicht an technischen, sondern an organisatorischen Herausforderungen scheitert. Das Projektmanagement muss daher nach einem geeigneten Vorgehensmodell eine stringente Organisation und Steuerung der am Projekt beteiligten (Fach-)Abteilungen und (System-)Verantwortlichen durchsetzen. Die Zusammenarbeit zwischen den Fachteilungen aber auch die Unterstützung aus der Führungsebene ist für die erfolgreiche DWH-Einführung von besonderer Bedeutung [FrWe00, S. 32], [Ditt99, S. 97 ff.], [Kurz99, S. 309], [HoKB01, S. 41], [JLVV03, S. 13 f.], [Bess04], [WiDV04, S. 90].
- *Datenzugriff:* Insbesondere die Autonomie und die Datenhoheit einzelner Organisationsbereiche oder Geschäftseinheiten können ein Hindernis für die Realisierung einer ganzheitlichen DWH-Lösung darstellen. Die Weitergabe der eigenen Daten ist selbst organisationsintern nicht als selbstverständlich anzunehmen. Komplexe oder verteilte Organisationsstrukturen verkomplizieren diesen Sachverhalt zusätzlich. Das konzeptionelle Idealbild eines organisationsweit einzigen Datenschemas ist somit kaum zu realisieren. Darüber hinaus sind bestehende Rahmenbedingungen, wie z. B. rechtliche Vorgaben oder die Einhaltung der Datenschutzgesetze bei der Verarbeitung und Speicherung von Daten in einem DWH (z. B. [SiPU02, S. 4 ff.], [Hafn02, S. 25 ff.]), stets zu berücksichtigen [HoKB01, S. 41], [Lehn03, S. 7].

- *Projektrisiken*: Die Einführung eines DWH-Systems ist aufgrund der Komplexität und Dauer eines solchen Projekts mit hohen Risiken und Kosten verbunden [Ditt99, S. 2]. Aus diesem Grund wird ein evolutionäres, prototypisches Vorgehen empfohlen, bei dem die Phasen Analyse, Design und Implementierung mehrfach zyklisch durchlaufen werden. Die Anforderungen der Endanwender, die zumeist erst während der Nutzungsphase konkretisiert werden können, lassen sich auf diese Weise sukzessive realisieren [Jone98, S. 46], [Kurz99, S. 309], [Sinz02, S. 317 f.], [WiDV04, S. 112 f.].
- *Betriebsphase*: Auch nach der Produktivsetzung ist ein DWH-System durch Anpassungs- und Erweiterungsarbeiten stets an veränderte Rahmenbedingungen und Anforderungen der Endanwender anzupassen. Während der Betriebsphase sind zudem Kriterien wie Verfügbarkeit, Leistungsfähigkeit und Zuverlässigkeit sowie Bedien- und Wartbarkeit des Gesamtsystems Faktoren, die neben der Aufrechterhaltung einer kontinuierlich hohen Datenqualität entscheidend zur Akzeptanz des DWH-Systems beitragen [Ditt99, S. 101], [AdMo00, S. 79 f.], [WiDV04, S. 90].

Studien zufolge scheitert ungefähr jedes zweite DWH-Projekt [BuKD06, S. 35], [Pott98, S. 94]. Die hohe Komplexität sowie die Kosten und die Dauer einer (organisationsweiten) DWH-Einführung zählen hier mit zu den Hauptgründen. Diese Faktoren sind außerdem für die Entstehung bereichsspezifischer DWH-Insellösungen verantwortlich, von denen eine schnelle, kostengünstige und auf die eigenen Bedürfnisse zugeschnittene Realisierung erwartet wird. Auch die Wahrung der eigenen Datenhoheit sowie die Notwendigkeit, einen Konsens mit anderen Organisationseinheiten (z. B. über die Datensemantik) zu finden, sind ursächlich für das Aufkommen bereichsspezifisch isolierter DWH-Systeme.

Ein nicht-standardisiertes Vorgehen bei DWH-Einführungsprojekten trägt zusätzlich zur Heterogenität der entwickelten DWH-Systeme bei [Kell97, S. 195]. Heterogenität bei DWH-Systemen kann auch durch historisch bedingtes Wachstum entstehen. An INMONS Schilderung des *Day 1 to Day n Phenomenon*, bei dem die Evolution von DWH-Systemen in einer Organisation beleuchtet wird, lässt sich dieses gut nachvollziehen [Inmo05, S. 39 ff.].

2.3.5 Data-Warehouse-Systeme in einer verteilten Organisation

Insbesondere bei der verteilten Organisationsform summieren sich die Treiber für das Entstehen multipler DWH-Systeme. Jede Sparte betreibt, nicht zuletzt aufgrund der ihr zugesprochenen Autonomie, ein eigenes DWH-System. Dieses ermöglicht eine auf die spezifischen Anforderungen der Sparte ausgerichtete Analyse der Daten aus den eigenen operativen Systemen. Zudem ist eine unabhängig von der Gesamtorganisation initiierte DWH-Einführung schneller und zunächst kostengünstiger zu realisieren. Aus denselben Gründen können auch innerhalb einer Sparte mehrere DWH-Systeme entstehen. Abbildung 2-8 visualisiert diese Konstellation.

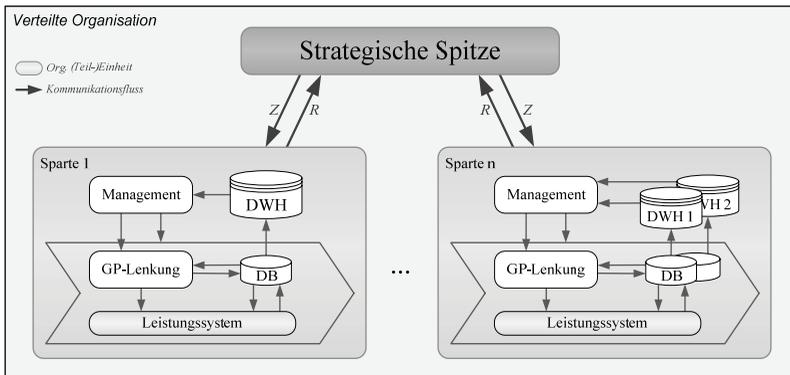


Abbildung 2-8: Regelkreisorientierte Einordnung eines DWH bei einer verteilten Organisation

Die Realisierung eines konsolidierten Berichtswesens zur Kommunikation mit der Führungsebene (Strategischen Spitze) erfordert somit eine aufwändige Analyse und Verknüpfung der Berichtsergebnisse aus den Sparten. Diskrepanzen zwischen den Berichten sowie ein hoher Interpretationsspielraum der Resultate sind durch ein fehlendes gemeinsames Verständnis der Terminologie vorprogrammiert [LeE197, S. 171].

2.3.6 Spezifische Data-Warehouse-Konzepte

Bei klassischen DWH-Systemen wird der Datenbestand in periodischen Abständen erweitert und kann von Analysewerkzeugen lesend abgefragt werden. Für einige Anwendungsfälle wie z. B. *Customer Relationship Management (CRM)*³² oder *Echtzeitanalysen* ist dieses Paradigma jedoch nicht ausreichend. Aus diesem Grund sind einige Implementierungsvarianten des klassischen DWH entstanden.

CLOSED-LOOP-DATA-WAREHOUSE

Beim *Closed-Loop-DWH* werden die operativen Systeme mit Informationen aus den Analyseergebnissen ergänzt. Es existiert somit eine Rückkopplung vom DWH zu den operativen Systemen. Besonders bei der Verknüpfung von analytischem und operativem CRM kommt der Closed-Loop-Ansatz zum Tragen [KeMU06, S. 87], [HiRW06, S. 49].

REAL-TIME-DATA-WAREHOUSE

Mitunter ist die durch den definierten Ladezyklus erzielte Datenaktualität im DWH nicht hinreichend. Beim so genannten *Real-Time-DWH* werden Daten der operativen Systeme nahezu in Echtzeit, also möglichst zum Zeitpunkt ihres Entstehens, im DWH bereitgestellt. Zur Sicherstellung der inhaltlichen Konsistenz des Real-Time-DWH können bspw. dedizierte EAI-Infrastrukturen³³ verwendet werden, die eine zeitnahe Transformation der operativen Daten unterstützen [ChGI04, S. 126], [KeMU06, S. 87 f.].

ACTIVE-DATA-WAREHOUSE

Mit dem Konzept des *Active-DWH* wird eine zusätzliche Operationalisierung des klassischen DWH und somit eine stärkere Unterstützung der unteren Managementebenen angestrebt. Aufgrund der Tatsache, dass operative Entscheidungssituationen zumeist besser strukturiert sind als strategische, können bestimmte Aktionen (teil-)automatisiert vom Active-DWH veranlasst werden. Ein mit dem Active-DWH eng verwobenes Paradigma ist das *Event-Condition-Action-Model (ECA-model)*; z. B. [ElNa07, S. 809]) [KeMU06, S. 87 f.], [ChGI04, S. 126 f.], [ThSM01].

³² *Customer Relationship Management (CRM)* dient der Abstimmung von Kundenkommunikation, Produkten und Dienstleistungen auf die individuellen Kundenbedürfnisse. Für eine Einführung in analytisches und operatives CRM wird z. B. auf [HiRW06] verwiesen.

³³ Unter *Enterprise Application Integration (EAI)* werden allgemein Produkte zur Integration von Anwendungssystemen subsumiert. Erläuterungen zu EAI offerieren u. a. [MaSc02], [LeNa07a, S. 401 ff.].

ENTERPRISE-DATA-WAREHOUSE

Ein *Enterprise-Data-Warehouse* (EDW) soll in einer gemeinsamen Datenbank alle entscheidungsrelevanten Daten konsolidiert vorhalten, sodass jeder dispositive Prozess der unterschiedlichen Organisationseinheiten darauf Zugriff hat. Die Realisierung eines EDW ist aus organisatorischen, technischen oder wirtschaftlichen Gründen meist nur schwer möglich. Die Einführung eines EDW wird von betrieblichen Organisationen folglich nur zögerlich verfolgt [Jones98, S. 43 f.], [ChGI04, S. 126], [Bess04], [Garr05].

WEITERE DATA-WAREHOUSE-VARIANTEN

Neben den vorgestellten Implementierungsvarianten des klassischen DWH existieren auch Bestrebungen zur Anreicherung eines DWH mit unstrukturierten Daten bspw. aus einem *Document-Warehouse* [Bang02]. Die zusätzliche Einbindung von Internetquellen wurde unter dem Pseudonym *Web Farming* erforscht [Hack98]. Eine DWH-Erweiterung stellt auch der Ansatz des *Knowledge-Warehouse* dar [Ditt04], [HaTB01].

2.4 Zusammenfassung

Nach der Einführung in betriebliche Organisationsstrukturen und der Darlegung der Führungsstruktur im Hochschulwesen, wurden in diesem Kapitel FIS im Allgemeinen und DWH-Systeme im Speziellen erörtert. Dabei konnten insbesondere folgende Ursachen für das Entstehen multipler DWH-Lösungen identifiziert werden:

- Organisationale Rahmenbedingungen (z. B. die Autonomie der Sparten)
- Individuelle Ansprüche der Endanwendergruppen
- Rechtliche sowie datenschutzrechtliche Vorgaben
- Strukturveränderungen in einer Organisation (z. B. durch M&A)
- Berücksichtigung bestehender IT-Infrastrukturen
- Komplexität, Dauer und Kosten der DWH-Einführung
- Nicht standardisiertes Vorgehen bei der DWH-Einführung
- Historisches Wachstum der DWH-Landschaft

Zudem wurde die Notwendigkeit einer zentralen DWH-Lösung aufgezeigt, um die Nutzentpotentiale eines solchen Systems voll ausschöpfen zu können.

3 Metadaten in Data-Warehouse-Systemen

Ziel des vorliegenden Kapitels ist, die Bedeutung von Metadaten für DWH-Systeme herauszustellen und den gegenwärtigen Entwicklungs- und Forschungsstand bei der Verwaltung von DWH-Metadaten aufzuzeigen. Nach einer Begriffsabgrenzung wird die Rolle der Metadaten in DWH-Systemen erörtert. Daran schließt sich die Betrachtung gängiger Metadaten-Klassifikationen sowie maßgebender Konzepte und Standards für den Metadatenaustausch an. Abschließend werden Aspekte der Metadaten-Verwaltung beleuchtet.

3.1 Begriffsbestimmung

Mit der Vorsilbe *Meta* werden diejenigen Daten gekennzeichnet, die auf einer höheren Abstraktionsebene die Daten einer darunterliegenden Ebene beschreiben. *Metadaten* erfassen folglich die Bedeutung und Eigenschaften von Daten, um diese korrekt interpretieren, einordnen, verwalten und nutzen zu können.

Data that describes the meaning and structure of business data, as well as how it is created, accessed and used [Dev197, S. 52].

Im IT-Kontext werden unter dem Begriff Metadaten gemeinhin alle Arten von Informationen subsumiert, die für die Analyse, den Entwurf, die Implementierung und Nutzung eines Computersystems erforderlich sind [StVV99a, S. 7], [VaVe01, S. 273], [BaGü04, S. 328], [Auth04, S. 27 ff.], [KeMU06, S. 42 f.].

Die Speicherung und Bereitstellung von Metadaten erfolgt mithilfe eines *Metadaten-Repositorys*. Dieses ist im Allgemeinen auf Basis eines Datenbankmanagementsystems (DBMS) realisiert [BaGü04, S. 330].

3.2 Bedeutung von Metadaten für Data-Warehouse-Systeme

Erst in Verbindung mit Metadaten gewinnen Daten an Bedeutung. Bei DWH-Systemen werden daher bei der Übertragung der Daten aus den Quellsystemen Metadaten hinzugefügt, die dem Informationsnachfrager die korrekte Interpretation und Nutzung der ursprünglichen Daten ermöglichen. Zudem unterstützen Metadaten ein schnelles und

sicheres Auffinden der benötigten Daten. Die Anwender eines DWH-Systems sind demzufolge zwingend auf Metadaten angewiesen.

Metadaten in DWH-Systemen beschreiben neben den im DWH gespeicherten Daten auch die zum Aufbau und Betrieb des DWH-Systems erforderlichen Prozesse, Architekturen und Organisationsstrukturen. Um ein konsistentes Verständnis über die Daten eines DWH-Systems zu gewährleisten, ist die homogene Repräsentation der Metadaten eine notwendige Bedingung [ScDi06a, S. 87], [Auth04, S. 34 ff.], [Holt99b, S. 95 ff.], [KeMU06, S. 42 f.], [HaWe08, S. 118].

NUTZENPOTENTIALE VON METADATEN IN DWH-SYSTEMEN

Für die Arbeit mit einem DWH-System ist die Verfügbarkeit von Metadaten eine Voraussetzung. Aus Anwendersicht dienen Metadaten vor allem der Steigerung von Effektivität und Effizienz bei der Nutzung eines DWH-Systems, aus Administrator- oder Entwicklersicht führen sie zu einer Aufwandsreduktion bei der jeweiligen Aufgabendurchführung. Tabelle 3 gibt einen Überblick über Einsatz- und Nutzenpotentiale von Metadaten im DWH-Umfeld [VaVe01, S. 277 f.], [KeMU06, S. 45 f.].

| Anwendersicht | Administrator-/Entwicklersicht |
|--|--|
| Verbesserung der DWH-Nutzung durch: <ul style="list-style-type: none"> • Schnelles Auffinden relevanter Daten (z. B. durch Orientierungshilfen über das Datenangebot) • Kenntnis der vorliegenden Datenstrukturen (z. B. Vermeidung ungültiger Berichtsdefinitionen oder ineffizienter Abfragen) • Unterstützung bei der korrekten Interpretation der Daten • Informationen über den Ursprung der Daten (Data Lineage) • Erhöhung der Anwenderakzeptanz | Reduzierung des Entwicklungsaufwands durch: <ul style="list-style-type: none"> • Unterstützung bei der Datenintegration (z. B. durch Metadaten über Datenschemata) • Automatisierung von Entwicklungsprozessen • Automatisierung von Betriebsprozessen • Unterstützung bei der Fehlersuche • Wiederverwendung bestehender Metadaten (z. B. Datenstrukturen, Transformationen) und damit Erhöhung der Flexibilität bei der Anpassung/Erweiterung eines DWH-Systems • Zentrale Verwaltung der Zugriffsrechte |
| Übergreifende Sicht | |
| Verbesserung übergreifender DWH-Prozesse durch: <ul style="list-style-type: none"> • Bereitstellung eines einheitlichen Begriffsverständnisses • Konsistente und integrierte Dokumentation der DWH-Inhalte • Verkürzung von Einarbeitungszeiten, Wissenstransfer durch Experten | |

Tabelle 3: Einsatz- und Nutzenpotentiale von DWH-Metadaten (in Anlehnung an [Auth04, S. 36])

Metadaten werden auf allen Ebenen eines DWH-Systems produziert und konsumiert. Eine signifikante Herausforderung besteht daher in der Bereitstellung konsistenter und aktueller Metadaten für die unterschiedlichen Anwendergruppen [JLVV03, S. 13], [BaGü04, S. 328], [KeMU06, S. 43].

3.3 Klassifikation von Metadaten

Die Metadaten eines DWH-Systems können aus verschiedenen Blickwinkeln klassifiziert werden (vgl. z. B. [StVV99b, S. 18 ff.], [Holt99b, S. 98 ff.], [Lehm01, S. 83 ff.], [Auth04, S. 38 ff.]). Aus Sicht der Nutzung von DWH-Metadaten wird zwischen folgenden drei Kategorien unterschieden [StVV99b, S. 5 f.], [KeMU06, S. 44]:

- *Passive Metadaten* dienen einer konsistenten Dokumentation der Struktur, des Entwicklungsprozesses und der Verwendung eines DWH-Systems. Passive Metadaten unterstützen alle Beteiligten eines DWH-Prozesses.
- *Aktive Metadaten* stellen semantische Beschreibungen von DWH-Komponenten dar, die zur Laufzeit des DWH-Systems interpretiert und zur unmittelbaren Ausführung bspw. von Transformations- oder Analyseprozessen genutzt werden.
- Zu den *semi-aktiven Metadaten* gehören statische Informationen (z. B. Strukturinformationen), die zur Laufzeit von anderen Softwarekomponenten gelesen, aber nicht aktiv zu deren Steuerung genutzt werden.

Ebenfalls eine hohe Akzeptanz besitzt die Unterteilung in *fachliche* und *technische* Metadaten. Als Differenzierungskriterium dienen dabei die Benutzertypen, die Zugang zum DWH besitzen. Betriebswirtschaftliche Informationen (z. B. fachliche Begriffs- und Kennzahldefinitionen) zählen zu den fachlichen Metadaten, die primär von den Endanwendern genutzt werden. Produzenten und Konsumenten technischer Metadaten (z. B. Transformationsregeln, Zugriffssteuerung) sind vornehmlich Administratoren und Entwickler. Eine eindeutige Zuordnung der Metadaten in eine dieser Klassen ist nicht immer möglich [StVV99a, S. 44 f.], [Lehm01, S. 84 ff.].

Die vorgestellten Klassifikationsansätze bleiben eine Aufstellung der für die hinreichende Beschreibung eines spezifischen Einsatzbereichs notwendigen Metadaten schuldig. Daher wird an dieser Stelle die *verwendungszweckgerichtete Metadatenkategorisierung* nach AUTH eingeführt. AUTH identifiziert für DWH-Systeme acht Metadaten-Kategorien, die eine eingehende Metadaten-Beschreibung erlauben [Auth04, S. 44 ff.]:

- Metadaten der Kategorie *Terminologie* dienen der Verwaltung und dem korrekten Verständnis von Fachbegriffen. Hierzu zählt auch die Unterstützung bei der Homogenisierung betrieblicher Begriffssysteme. Zudem dienen Metadaten dieser Kategorie als Orientierungshilfe in Bezug auf das Datenangebot.
- Bei Datenflüssen zwischen DWH-Komponenten stellen Datenstrukturen das Vehikel zum Transport der Daten dar. Der Kategorie *Datenstruktur und -bedeutung* zugeordnete Metadaten beschreiben diese Datenstrukturen und ermöglichen deren Wiederverwendung.
- Metadaten, die zur Definition und Steuerung von Datentransformationsprozessen eingesetzt werden, gehören der Kategorie *Datentransformation* an. Vor allem bei der Einplanung, Überwachung und Erweiterung von Datentransformationsprozessen unterstützen diese Metadaten Administratoren und Entwickler.
- Metadaten können Informationen bezüglich der Datenqualität (z. B. Korrektheit, Vollständigkeit der Daten) auf Datenschema- und Datenwertebene erfassen (Kategorie *Datenqualität*). Die Verfügbarkeit von Qualitätsmetadaten ist Voraussetzung für ein proaktives Datenqualitätsmanagement [Helf02, S. 100 ff.].
- Die Entstehung und Verwendung der Daten im Rahmen der Geschäftsprozesse sowie Zugriffsberechtigungen und Vertraulichkeitseinstufungen für die Daten dokumentieren Metadaten der Kategorie *Organisationsbezug*. Sie stellen die Verknüpfung von Daten mit einer Organisationseinheit her [Dev197, S. 279 f.].
- Bei der Metadaten-Kategorie *Historisierung* steht das Erfassen von Veränderungen an Metadaten im Mittelpunkt. Durch Historisierung (z. B. Einsatz von Gültigkeitszeiträumen) oder Versionierung (z. B. Erzeugen einer neuen Version eines Metadatenelements) können Veränderungen an Datenelementen über die Zeit nachvollzogen werden.
- Die Beschreibung der individuellen Ausprägung einer DWH-Architektur durch Metadaten der Kategorie *Systembezug* bildet die Grundlage für eine durchgängige Dokumentation des DWH-Systems sowie für eine Darstellung der Interdependenzen der DWH-Komponenten.
- Einen Überblick über die Aufbereitung und Nutzung der in einem DWH gespeicherten Daten offerieren Metadaten der Kategorie *Datenanalyse*. Sie richten sich vornehmlich an Endanwender mit dem Ziel, die Effektivität und Effizienz der Datenanalyse zu steigern.

In Tabelle 4 wird ein Überblick über die Verwendungszweckgerichteten Metadatenkategorien sowie deren zentralen Metadatenelemente gegeben. Eine detaillierte Erläuterung der einzelnen Kategorien ist in [Auth04, S. 47 ff.] nachzulesen.

| Metadaten-Kategorie | Erläuterung |
|----------------------------------|---|
| Terminologie | Informationen zur Verwaltung von Fachbegriffen (eindeutige Begriffsbenennung, Definition und Ursprung des Fachbegriffs, Begriffsbeziehungen, Synonyme, Homonyme, Verantwortliche für den Fachbegriff) |
| Datenstruktur und Datenbedeutung | Metadaten zur Beschreibung von Datenstrukturen (Name, Beschreibung, Typ einer Datenstruktur, Elemente, Datentyp der Elemente, Ersteller etc.) |
| Datentransformation | Angaben über Datentransformationsprozesse (Datenquellen und -ziele, Transformationsschritte etc.) |
| Datenqualität | Datenqualitätsmerkmale für die Ebenen Datenschema und Datenwerte (Semantik, Identifizierbarkeit, Vollständigkeit etc.) |
| Organisationsbezug | Metadaten über die datenproduzierenden und -konsumierenden Organisationseinheiten (Datenproduzent, Datenkonsument, Berechtigungen, Entstehungs- und Verwendungskontext etc.) |
| Metadatenhistorie | Informationen zur Historisierung von Metadaten (Version, Erläuterung der Änderung, Änderungsdatum etc.) |
| Systembezug | Beschreibung der individuellen Ausprägung der konzeptionellen DWH-Architektur (Softwarehersteller, -komponente, -version, Hardware etc.) |
| Datenanalyse | Metadaten zur Beschreibung der Analysemöglichkeiten (Cube, Dimension und Dimensionselement, Kennzahl, Kennzahlenbeziehung etc.) |

Tabelle 4: Verwendungszweckgerichtete Metadaten-Kategorien [Auth04, S. 44 ff.]

Bei der Ausgestaltung dieser Metadatenkategorien wurden von AUTH zunächst Metadaten-Objekte identifiziert, die für die Realisierung des Nutzens von Metadaten in DWH-Systemen erforderlich sind (vgl. Tabelle 3). Diese wurden inhaltlich zu Metadatenkategorien zusammengefasst. Der Vorschlag der Metadatenkategorisierung stützt sich auf die konsolidierten Ergebnisse von zwei Workshops des *Kompetenzzentrums Data Warehousing 2* (Universität St. Gallen) zum Metadatenmanagement [Auth04, S. 44 f.]. Empirische Ergebnisse aus dem jahrelangen Betrieb von DWH-Systemen im Projekt CEUS^{HB} unterstreichen die Anwendbarkeit dieser Metadatenklassifizierung in der Praxis. Bei den nachfolgenden Untersuchungen wird daher durchgängig auf diese Metadatenklassifizierung zurückgegriffen.

3.4 Standardisierungen und Austauschformate für Metadaten

Zur Wahrung der Integrität eines Metadaten-Repositorys sowie zur Sicherstellung eines syntaktisch und semantisch korrekten Austauschs von Metadaten zwischen DWH-Komponenten ist ein standardisiertes Format für die Beschreibung von Metadaten notwendig. Im Folgenden werden aktuelle Metadatenstandards sowie etablierte Konzepte des Metadatenaustauschs vorgestellt und diskutiert.

3.4.1 Repository-Standards

Metadaten werden mithilfe eines Repositorys verwaltet. Ein Metadaten-Repository sollte hierfür grundlegend folgende Funktionalitäten aufweisen [VaVe01, S. 278 ff.]:

- Zur Unterstützung des *Anwenderzugriffs* verfügt ein Repository über Mechanismen zur Navigation im Metadatenbestand sowie zur Filterung und Selektion der Metadaten.
- Zur Erfüllung der Kriterien *Interoperabilität* und *Werkzeugunterstützung* sind Programmierschnittstellen und ein definiertes Austauschformat vorhanden.
- Zur Verwaltung von Veränderungsprozessen (*Change Management*) weist ein Repository eine Versions- und Konfigurationsverwaltung auf.

Die Anforderungen eines spezifischen Einsatzgebiets (z. B. DWH-Systeme) bestimmen die Struktur des Repositorys und die Semantik der zu speichernden Metadaten [HaLe93, S. 15 ff.], [BaGü04, S. 330]. Um die Interoperabilität zwischen Metadaten-Repositories zu verbessern, wurden u. a. die beiden folgenden Repository-Standards vorgeschlagen:

- Um einen konsolidierten Datenaustausch zwischen Repositories zu ermöglichen, wurde 1990 von der *International Organization for Standardization* (ISO) der Standard *Information Resource Dictionary System* (IRDS)³⁴ verabschiedet. Ein *Information Resource Dictionary* (IRD) stellt ein gemeinsames Repository für Metadaten dar, die für alle Bereiche einer betrieblichen Organisation von Relevanz sind. Der Standard ist in erster Linie für den Einsatz bei kommerzieller Software-Entwicklung geeignet. Jedoch ist davon auszugehen, dass insbesonde-

³⁴ Die Originalbeschreibung des *ISO-IRDS-Standards* sowie dessen Erweiterungen können unter http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=17985 (Abruf am 10.06.2008) angefordert werden. Eine in Grundzügen ähnliche IRDS-Standardisierung wurde 1988 vom American National Standards Institute (ANSI) verabschiedet (ANSI X3.138:1988).

re durch die auf oberster Abstraktionsebene vorgegebenen Restriktionen (der IRDS-Standard schlägt eine 4-Ebenen-Architektur vor) der IRDS-Standard bei den meisten Repository-Produkten nicht erfüllt werden kann [Stül91, S. 20 ff.], [HaLe93, S. 61 ff.], [Byrn97], [StVV99b, S. 10 f.], [Lehm01, S. 97].

- Der Repository-Standard *Portable Common Tool Environment* (PCTE)³⁵ beschreibt die Basis für eine standardisierte Softwareentwicklungsumgebung, einschließlich eines Repositories und der Kommunikation zwischen den Werkzeugen. Der Hauptanwendungsbereich des PCTE-Standards liegt in der technischen und wissenschaftlichen Software-Entwicklung. Im Standard ist keine Kommunikation zwischen PCTE-Repositories oder zwischen einer PCTE-Installation und einem anderen System vorgesehen [Byrn97], [StVV99b, S. 11].

Die Verbreitung dieser Repository-Standards ist im Bereich von DWH-Systemen aufgrund oben genannter Merkmale (z. B. Nicht-Erfüllung der durch den Standard vorgegebenen Restriktionen, Fokus auf technische und wissenschaftliche Software-Entwicklung) als gering einzustufen.

3.4.2 Metadaten-Standards und -Rahmenwerke

Mit Standardisierungen für die Auszeichnung von Metadaten im DWH-Kontext wird versucht, eine homogene Metadaten-Beschreibung zu erzielen [HaWe08, S. 119]. Als Standardisierung für DWH-Metadaten ist gegenwärtig nur noch das *Common Warehouse Metamodel* (CWM)³⁶ von Bedeutung [BaGü04, S. 334]. Das CWM wird im direkten Anschluss beleuchtet. Zudem wird mit dem *Zachman-Framework* ein generisches Organisationsschema für Metadaten im Rahmen der Systementwicklung vorgestellt.

³⁵ Detailinformationen zum *PCTE-Standard* bietet u. a. folgendes Dokument: <http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-149.pdf> (Abruf am 10.06.2008).

³⁶ Die *CWM-Spezifikation* kann unter <http://www.omg.org/technology/cwm> eingesehen werden (Abruf am 10.06.2008).

- Das CWM ist ein von der *Object Management Group* (OMG)³⁷ entwickelter Standard für Metadaten von DWH-Systemen. Insbesondere die Beschreibung technischer Metadaten sowie der Austausch von Metadaten zwischen verschiedenen DWH-Komponenten stehen im Fokus des CWM-Standards. Für den datenbasierten Metadatenaustausch stehen *XML Metadata Interchange* (XMI; vgl. Abschnitt 3.4.3), für den funktionsbasierten Zugriff *CORBA-Schnittstellen*³⁸ zur Verfügung. Schwachstellen besitzt das CWM im Bereich der Erfassung von Metadaten zur Datensicherheit, zum Datenqualitätsmanagement, zur Verwaltung eines Berichtswesens und zur Erfassung der Semantik der DWH-Inhalte. Letzteres soll durch das Paket *Business Nomenclature* abgedeckt werden. Dieses gestattet jedoch nur in rudimentärer Form die Bildung einer Taxonomie oder eines Glossars über die Daten eines DWH. Die Komplexität des CWM ist u. a. ursächlich dafür, dass der Standard nur zu einem geringen Teil in der Praxis berücksichtigt wird [Melc03, S. 91 ff.], [Ditt04, S. 404, S. 407], [Hahn07, S. 17].
- Das *Zachman-Framework* ist als Matrix organisiert, die als Organisationschema für Metadaten verstanden werden kann. Die Zeilen der Matrix umfassen fünf grundlegende Rollen der Systementwicklung (Planer, Besitzer, Designer, Konstrukteur, Unterauftragnehmer). Die Spalten visualisieren sechs Architekturkriterien (Daten, Funktionen, Netzwerk, Personen, Zeit, Motivation). Das Zachman-Framework dient vornehmlich als Leitfaden, um eine unternehmensweite Software-Architektur zu modellieren. Es besitzt daher nur einen sehr geringen Formalisierungsgrad [Zach87], [BaGü04, S. 342], [FHH+05, S. 204 ff.].

Es bleibt festzuhalten, dass ein in der Praxis durchgängig angewandter Standard mit hinreichendem Potential, die Metadaten eines DWH-Systems semantisch zu beschreiben, derzeit nicht existiert.

³⁷ Vgl. <http://www.omg.org> (Abruf am 10.06.2008).

³⁸ Eine Einführung in *CORBA* (*Common Object Request Broker Architecture*) bietet folgende Webseite der OMG: <http://www.omg.org/gettingstarted/corbafaq.htm> (Abruf am 10.06.2008).

3.4.3 Metadaten-Austauschformate

Um Interoperabilität zwischen DWH-Komponenten zu erreichen, ist die Standardisierung von Austauschformaten für Metadaten erforderlich. Im CWM-Standard ist für den Datenaustausch u. a. *XML Metadata Interchange*³⁹ (XMI) vorgesehen. Da der einzig nennenswerte, nicht auf der *Extensible Markup Language* (XML; vgl. Abschnitt 6.2.2) beruhende Austauschstandard, *Case Data Interchange Format* (CDIF), voraussichtlich zugunsten von XMI eingestellt wird, konzentrieren sich an dieser Stelle die Ausführungen auf XMI [BaGü04, S. 334 f.].

Basierend auf XML wurde XMI von der OMG als Standardformat für die Übermittlung von Metadaten in Softwaresystemen entwickelt. Mithilfe von XMI kann ein Datenaustausch zwischen MOF-basierten⁴⁰ Repositories realisiert werden. Kern des XMI-Standards sind die folgenden zwei Komponenten [StVV99b, S. 11], [Jeck04, S. 29 ff.]:

- *XML DTD/Schema Production Rules* beschreiben die Regeln für die Erstellung von *XML Document Type Definitions* (DTD) oder *XML Schema-Definitions* (vgl. Abschnitt 6.2.2) zu XMI-kodierten Metadaten.
- *XML Document Production Rules* beschreiben Regeln, um Metadaten in ein XML-kompatibles Format zu überführen.

Eine mögliche Anwendung von XMI zeigt der Vorschlag einer Software-Architektur für einen XML-basierten Metadatenaustausch in DWH-Systemen [AuMa02]. Die Kommunikation zwischen lokalen Metadaten-Repositories einzelner DWH-Komponenten und einem zentralen, CWM-basierten Repository erfolgt hierbei durch den Einsatz von XMI. Da das zentrale Repository auf dem CWM-Standard basiert, gelten für diesen Architekturvorschlag analog die oben genannten Einschränkungen hinsichtlich der Erfassung der Datensemantik. Die Vor- und Nachteile eines rein XML-basierten Datentransfers werden in Abschnitt 6.2.2 aufgegriffen.

³⁹ Die jeweils aktuelle *XMI-Spezifikation* kann unter http://www.omg.org/technology/documents/modeling_spec_catalog.htm#XMI eingesehen werden (Abruf am 10.06.2008).

⁴⁰ OMGs *MetaObject Facility* (MOF) ist ein erweiterbares, modelgestütztes und integriertes Rahmenwerk, um Metadaten und Daten plattformunabhängig zu definieren, zu verarbeiten und zu integrieren http://www.omg.org/technology/documents/modeling_spec_catalog.htm#MOF (Abruf am 10.06.2008).

3.5 Metadaten-Verwaltung

Unter dem Begriff *Metadaten-Verwaltung* oder auch *Metadaten-Management* wird die Gesamtheit der Prozesse zur zielgerichteten Entwicklung, Organisation und Nutzung von Metadaten subsumiert [Auth04, S. 62]. Für die Implementierung eines Metadaten-Repositorys im Rahmen einer integrierten Metadaten-Verwaltung werden folgende Architekturansätze unterschieden [DoRa00, S. 8 ff.], [Auth04, S. 70 ff.]:

- Bei einer *zentralen Architektur* werden die Metadaten aller DWH-Komponenten in einem zentralen Repository verwaltet. Eine lokale, redundante Metadaten-Verwaltung der einzelnen DWH-Komponenten entfällt. Die konsistente Metadaten-Verwaltung an einem Ort stellt den wichtigsten Nutzenaspekt dieser Architekturvariante dar. In der Praxis ist die Realisierung eines Metadaten-Repositorys in dieser Form jedoch kaum möglich (z. B. aufgrund fehlender Unterstützung für eine vollständige Auslagerung der spezifischen Metadaten durch die DWH-Komponenten oder aufgrund politischer und kultureller Probleme hinsichtlich der Verantwortlichkeit für die Pflege der Metadaten [Auth04, S. 70 f.]).
- Eine *dezentrale Architektur* liegt vor, wenn jede DWH-Komponente ihre Metadaten lokal verwaltet. Dieses Szenario ist typisch für gegenwärtige DWH-Umgebungen, die sich zumeist aus ausgewählten Produkten verschiedener Hersteller zusammensetzen (*Best-of-Breed-Ansatz*). Vorteilhaft ist die Autonomie und Effizienz beim Zugriff auf die spezifischen Metadaten. Allerdings ist eine Vielzahl von Schnittstellen für den Metadaten-Austausch zwischen den DWH-Komponenten erforderlich. Schnittstellen zwischen Produkten unterschiedlicher Hersteller werden zumeist nur bei potentiellen, wirtschaftlichen Vorteilen eines Herstellers angeboten. Die Sicherstellung von Konsistenz innerhalb der Gesamtheit der Metadaten stellt daher eine besondere Herausforderung dar.

- Die Variante der *verteilten Architektur* umfasst einerseits lokale Repositories für die spezifischen Metadaten der DWH-Komponenten und andererseits ein zentrales Repository, in dem gemeinsame Metadaten verwaltet werden. Das zentrale Repository enthält Kopien der lokalen Metadatenbestände und stellt diese in einer einheitlichen Repräsentation bspw. auf Basis des CWM zur Verfügung. Jede DWH-Komponente muss somit nur die Schnittstelle zum zentralen Repository implementieren, um auf Metadaten anderer Komponenten zugreifen zu können. Die Realisierung dieser Architekturvariante erscheint in Anbetracht gegenwärtiger DWH-Strukturen als recht aussichtsreich.

Die nachstehende Abbildung skizziert die vorgestellten Architekturvarianten.

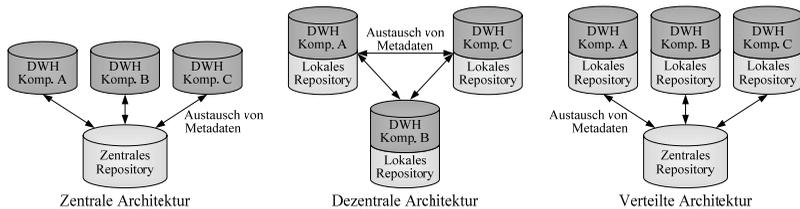


Abbildung 3-1: Architekturvarianten bei der Metadaten-Verwaltung [DoRa00, S. 8]

In der von INMON propagierten neuen Generation des Data-Warehousing (*DW 2.0™*) [Inmo06] kommt der Einführung eines *Enterprise-Metadaten-Repositorys* (EMR) eine besondere Bedeutung zu. Das EMR soll alle Metadaten eines Unternehmens umfassen. Die Metadaten lokaler Repositories sind demzufolge in das EMR zu überführen und permanent mit den zentral verwalteten Metadaten abzugleichen. Dies entspricht vom Grundsatz her der verteilten Repository-Architektur. Inwieweit beim EMR Standards – wie z. B. das CWM (vgl. Abschnitt 3.4.2) oder Funktionen für die Integration der Metadaten – vorgesehen sind, konnte nicht hinreichend ermittelt werden [Kurc06, S. 23 f.].

BARRIEREN IM RAHMEN EINER METADATEN-VERWALTUNG

Die jeweils vorliegende Architekturvariante bedingt die Aufgaben und Aktivitäten im Rahmen des Metadaten-Managements. Relevanz und Nutzenpotentiale einer stringenten Metadaten-Verwaltung sind dabei hinreichend erkannt. Jedoch entstehen nach wie vor Insellösungen bei der Verwaltung von Metadaten. Als hierfür ursächlich wurden insbesondere folgende Sachverhalte erkannt:

- auf *organisatorischer Ebene* die fehlende Abstimmung der Aktivitäten und die vage Formulierung der Projektaufträge im Bereich des Metadaten-Managements sowie die Intransparenz über bestehende Metadaten-Lösungen.
- auf *technischer Ebene* die proprietären Metadaten-Repräsentationen, die den Austausch und die Integration von Metadaten erschweren.
- auf *semantischer Ebene* die Konflikte zwischen den Fachabteilungen bei der Abstimmung der Begriffssysteme.

Zumeist mangelt es auch an der konsequenten Durchführung von Projekten zur Metadaten-Verwaltung, sodass entsprechende Vorhaben nur bereichsspezifische oder zeitlich eingeschränkte Wirkung besitzen [Bess04], [Dint06, S. 29].

PROZESSORIENTIERTE ORGANISATION DES METADATEN-MANAGEMENTS

Ein Konzept zur prozessorientierten Organisation der Metadaten-Verwaltung wurde von AUTH vorgeschlagen. Auf oberster Modellierungsebene unterscheidet dieser Ansatz folgende vier Prozesse [Auth04, S. 187 ff.]:

- Dem *Terminologie-Management* sind alle Aufgaben zum Aufbau eines unternehmensweit einheitlichen Begriffssystems zugeordnet.
- Die Katalogisierung und Standardisierung der Datenstrukturen des DWH-Systems sowie der Vordatenbanken zählen zu den Aufgaben im Bereich *Datenstruktur-Management*.
- Aufgaben zur Erfassung und Kontrolle der Datenqualität sowie das Initiieren von Verbesserungsmaßnahmen bezieht das *Datenqualitätsmanagement* ein.
- Der Prozess *Datenkontext-Management* stellt den Kontext bereit, der den System- und Organisationsbezug der Daten dokumentiert.

Der in dieser Arbeit zu entwickelnde Ansatz zur Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen konzentriert sich auf die Unterstützung der Aufgaben aus dem Bereich Terminologie-, Datenkontext- und Datenstruktur-Management. Das Datenqualitätsmanagement ist für die Überwindung von Heterogenität zunächst von nachrangiger Bedeutung.

3.6 Zusammenfassung

Die Ausführungen dieses Kapitels konzentrierten sich auf Metadaten und deren Bedeutung in DWH-Systemen. Dabei wurden Aspekte der Klassifikation, der Standardisierung und des Austauschs sowie der Verwaltung von Metadaten untersucht. Zu konstatieren ist, dass sowohl Anwender als auch Entwickler und Administratoren eines DWH-Systems zwingend auf Metadaten angewiesen sind. Die Bereitstellung konsistenter und aktueller Metadaten ist daher obligatorisch. Bestrebungen hinsichtlich einer ganzheitlichen Metadaten-Verwaltung verfehlen jedoch aus unterschiedlichen Gründen ihre Wirkung. Isolierte Metadaten-Repositories stellen die gegenwärtige Situation im DWH-Umfeld dar. Zudem konnte festgestellt werden, dass der einzig bedeutungsvolle Metadaten-Standard im DWH-Umfeld, das CWM, semantische Aspekte nur unzureichend einbezieht und aufgrund seiner Komplexität nur zu einem geringen Teil in der Praxis berücksichtigt ist.

Zu der in Kapitel 2 identifizierte Heterogenität zwischen DWH-Systemen kommt mit der Koexistenz von verschiedenartigen, nicht standardisierten Metadaten-Repositories innerhalb eines DWH-Systems eine weitere Dimension hinzu. Bei der Überwindung semantischer Heterogenität zwischen DWH-Systemen darf daher nicht von einem holistischen, zentralen Metadaten-Repository ausgegangen werden.

Als Anhaltspunkt für die Ausgestaltung einer semantisch reichhaltigen Metadaten-Beschreibung ist im Folgenden die verwendungszweckgerichtete Metadatenkategorisierung heranzuziehen. Diese weist wesentliche Elemente für eine eingehende Metadaten-Beschreibung aus. Im Hinblick auf die Überwindung der semantischen Heterogenität finden vor allem passive Metadaten Beachtung. Durch diese soll in einem ersten Schritt eine konsistente Dokumentation über die lokalen DWH-Systeme erreicht werden.

4 Führungsinformationen in betrieblichen Organisationen

Eine entscheidende Determinante des Erfolgs einer betrieblichen Organisation liegt in der Fähigkeit, die durch die Verarbeitung von Daten entstandenen Informationen zu nutzen, um neues Wissen zu generieren und zielgerichtet zu verwenden [GrZa92, S. 14].

Im vorliegenden Kapitel wird eingangs die Beziehung zwischen Daten, Information und Wissen sowie die Kodifizierbarkeit von Wissen erörtert. Dem schließt sich eine Diskussion über den Einfluss der Informationsversorgung bei der Bearbeitung einer Entscheidungsaufgabe sowie über die Auswirkungen von Informationspathologien an. Abgerundet wird das Kapitel durch die Betrachtung von Kontextinformationen bei DWH-Systemen. Die Ausführungen erfolgen jeweils im Hinblick auf die Unterstützung bei der Durchführung einer Entscheidungsaufgabe.

4.1 Daten, Information und Wissen

Das grundlegende Verständnis der Begriffe Daten, Information und Wissen wird nachfolgend anhand einer klaren Abgrenzung sowie einer Erklärung bestehender Beziehungen zwischen diesen Termini aufgezeigt. Anschließend wird Wissen im Hinblick auf seine Kodifizierbarkeit untersucht und die Bedeutung des Wissensmanagements für eine betriebliche Organisation skizziert.

4.1.1 Begriffsbestimmung und Abgrenzung

Das Ergebnis der Modellierung⁴¹ eines Ausschnitts der (betrieblichen) Realsphäre kann bspw. ein Datenschema darstellen, gemäß dem Daten gesammelt werden. *Daten* sind, für sich betrachtet, zunächst eine weitgehend bedeutungslose Ordnung von Zeichen (*syntaktische Ebene*). Sie bilden die Grundlage für einen Interpretationsprozess, welcher zugleich als initialer Schritt eines Entscheidungsprozesses aufgefasst werden kann. Durch Anwendung einer Interpretationsvorschrift α_M , welche das Wissen über die Mo-

⁴¹ Für die Charakterisierung des Begriffs *Modellierung* wird auf [FeSi06, S. 121 ff.] verwiesen.

dellabbildung M beinhaltet (z. B. das Datenschema), wird aus Daten eine *Information* abgeleitet. Informationen sind demgemäß verknüpfte und mit Bedeutung versehene Daten (*semantische Ebene*). Sie sind sowohl Eingabe- als auch Ergebnismenge von Entscheidungsprozessen. *Wissen* entsteht aufgrund individuell verarbeiteter Informationen, die in einem bestimmten Bezug zur Erfahrungswelt des jeweiligen Individuums stehen (*pragmatische Ebene*). Es ist das Ergebnis eines Lernprozesses und bildet die Gesamtheit der Kenntnisse und Fähigkeiten, auf die ein Individuum bei der Lösung von Problemen zurückgreift. Wissen unterstützt und beeinflusst somit Entscheidungsprozesse [AaNy95, S. 6 ff.], [Wirt00, S. 3 f.], [FeSi06, S. 132], [KeMU06, S. 126 ff.].

Der Weg von Daten zu Wissen lässt sich wie folgt zusammenfassen:

Daten müssen interpretiert und verknüpft, Informationen wahrgenommen, verarbeitet und weiter mit Bedeutung versehen werden, erst dann sind sie in Wissen umgewandelt [Wirt00, S. 3 f.].

EINFLUSSBEREICH VON WISSEN

Wissen obliegt eine essentielle Rolle bei der Transformation von Daten in Informationen (*Interpretation*), bei der Ableitung neuer Informationen (*Ausgestaltung* der Informationsbasis) und beim Erwerb neuen Wissens (*Lernen*) [Wirt00, S. 8 ff.].

Abbildung 4-1 skizziert die Beziehungen zwischen Realwelt, Daten, Information und Wissen. Die grau gestrichelten Linien legen dar, welche Prozesse Wissen erfordert.

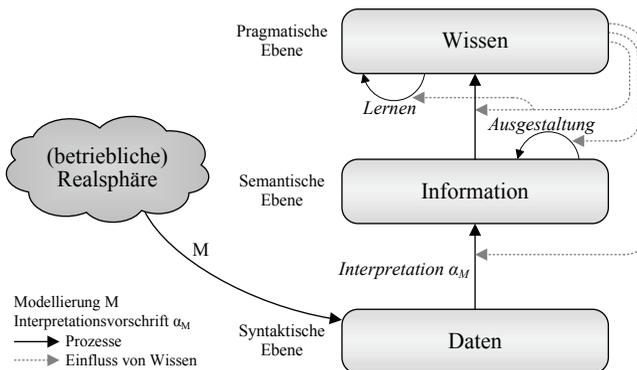


Abbildung 4-1: Daten-Information-Wissen-Modell
(in Anlehnung an [AaNy95 S. 8], [FeSi06, S. 132])

Die illustrierten Abhängigkeiten zwischen Daten, Information und Wissen sind wie folgt zu interpretieren:

- Für die *Interpretation* von Daten und somit die Ableitung von Informationen ist Wissen über die *Modellabbildung M* erforderlich. Dabei ist sowohl domänen- und kontextspezifisches als auch allgemeines (Welt-)Wissen gefordert. Um die Bedeutung von Daten kontextbezogen zu bestimmen, nutzt ein Mensch vor allem seinen kulturellen Hintergrund, seine Intuition, Erinnerungen und Erfahrungen sowie Lehrbuchwissen und domänenspezifische, heuristische Regeln.
- Mit der *Ausgestaltung* der Informationsbasis werden einerseits ein besseres Verständnis der verfügbaren Informationen und andererseits die Ableitung neuer Informationen angestrebt. Die Ausgestaltung bildet den eigentlichen Problemlöseprozess, über den eine Entscheidungsfindung erfolgt. Die Verfügbarkeit von Wissen ist an dieser Stelle obligat, da abgeleitete Informationen unter anderem auf bekannten Problemlösungen, Erklärungen und Rechtfertigungen von Annahmen oder Hypothesen beruhen. Des Weiteren bedarf es Wissen, um den Wert einer Information bspw. im Hinblick auf eine konkrete Aufgabe beurteilen zu können.
- *Lernen* ist die Integration neuer Informationen in eine bestehende Wissensstruktur, sodass diese für potentielle Entscheidungen zur Verfügung stehen. Neues Wissen entsteht auch durch Inferenzprozesse innerhalb der Wissensbasis. Wissen, das in einem spezifischen Kontext gewonnen wurde, kann dynamisch in verschiedenen, zukünftigen Situationen genutzt werden.

SUBJEKTIVITÄT UND VERNETZUNG

Mit der Transformation von Daten zu Information und Information zu Wissen ist im Allgemeinen eine fortschreitende Vernetzung bei gleichzeitiger Zunahme der *Subjektivität* zu beobachten. Damit einher geht eine Aufwertung der Relevanz. Daten stehen im Allgemeinen beliebigen Nachfragern zur Verfügung. Für Informationen gilt, dass diese von einem Individuum in Beziehung gesetzt und mit Bedeutung versehen wurden. Sie lassen sich jedoch wie im Fall der Daten auf Medien bereitstellen. Wissen dagegen ist zunächst an die Person gebunden, welche dieses Wissen erworben hat. Um das gewonnene Wissen übermitteln zu können, muss es expliziert und anschließend von Dritten rezipiert und wiederum erworben werden. Information kann folglich als subjektiv mög-

liches Wissen und Wissen als individuell angeeignete Information verstanden werden. Manches Wissen, bspw. Kompetenzwissen, ist jedoch nur schwer oder überhaupt nicht externalisierbar [Wirt00, S. 3 f.]. Hierauf Bezug nehmend erfolgt im nächsten Abschnitt eine Erläuterung zur Kodifizierbarkeit von Wissen.

4.1.2 Kodifizierbarkeit von Wissen

Wissen ist wertvoll, weil es die Voraussetzung für die Bewertung und Vernetzung von neuen Erfahrungen und Informationen ist. Informationsressourcen unterstützen die Prozesse der Schaffung und des Transfers von Wissen in Organisationen. Die Möglichkeit, Wissen zu kommunizieren gestattet sowohl dem einzelnen Menschen, seine Fertigkeiten auszubauen und zu verfeinern (*individuelles Wissen*), als auch Organisationen in ihrer Gesamtheit, Abläufe zu optimieren und somit Ziele effizienter zu erreichen (*kollektives Wissen*). Dabei ist für eine erfolgreiche Organisationsstrategie entscheidend, wie der Prozess, individuelles in kollektives Wissen und kollektives in individuelles zu überführen, gestaltet wird. Man spricht in diesem Zusammenhang auch von *organisationalem Lernen*⁴² (z. B. [PiRW03, S. 502 ff.], [KiWa07, S. 462 ff.]).

Um die im vorherigen Abschnitt dargelegten Prozesse der Wissensentstehung zu ergänzen sowie deren Beziehung zu organisationalem Lernen aufzuzeigen, soll an dieser Stelle die durch NONAKA und TAKEUCHI bekannte Dichotomie *tazites* und *explizites Wissen* eingeführt werden [NoTa95, S. 56 ff.], [Gass97, S. 150 ff.], [Krcm05, S. 479 ff.], [Pola58]:

- *Tazites Wissen* ist personengebunden und kontextspezifisch. Neben *sozialen Elementen* (z. B. Werte, Normen) zählen auch technische (z. B. Fertigkeiten und Fähigkeiten eines Individuums aufgrund persönlicher *Erfahrungen*) zu dieser Wissenskategorie. Da es sich hierbei primär um mentale Modelle handelt, die auf subjektiver Wahrnehmung der Realwelt von Individuen beruhen, ist *tazites Wissen* nur schwer formalisierbar und kommunizierbar. Aus semantischer Sicht kommt *tazitem Wissen* eine hohe Bedeutung zu, da nur durch *tazites Wissen* die Bedeutung vieler Daten erfasst werden kann.

⁴² Einen zentralen Aspekt des *organisationalen Lernens* stellt die Übertragung des Wissens der Mitarbeiter in formale Regelwerke dar, damit dieses der Organisation zur Verfügung steht [KiWa07, S. 36].

Zu ergänzen ist, dass – entgegen der in der Literatur geläufigen Gleichsetzung – LI und GAO zwischen *tazitem* und *implizitem Wissen* differenzieren [LiGa03]. Implizites Wissen unterstellt, dass ein Individuum nicht gewillt ist, sein Wissen zu kommunizieren, obwohl es dazu im Stande wäre, oder bestimmte Rahmenbedingungen die Weitergabe verhindern. Eine Kodifizierung impliziten Wissens ist offenbar unter bestimmten Voraussetzungen durchführbar. Implizites Wissen bildet somit eine Teilmenge taziten Wissens. Diese Ansicht wird auch in der vorliegenden Arbeit geteilt.

- Im Gegensatz zu tazitem Wissen kann *explizites Wissen* (leicht) artikuliert, dokumentiert und durch formale Methoden beschrieben werden. Somit besteht bspw. die Möglichkeit, wissenschaftliche Erkenntnisse zu formalisieren (*dokumentiertes Wissen*) oder Spezifikationen von Komponenten und Produkten zu kodifizieren (*Wissen als Artefakt*). Explizites Wissen kann auf unterschiedlichen Medien gespeichert, verarbeitet und übertragen werden.

Aus den obigen Ausführungen resultiert – mit Anlehnung an die Wissenspyramide von GASSMANN ([Gass97, S. 152]) – folgende Einstufung taziten und expliziten Wissens nach dem Grad der Kodifizierbarkeit. Implizites Wissen wird dabei als Teilbereich mit höherer Kodifizierbarkeit innerhalb taziten Wissens ausgewiesen.

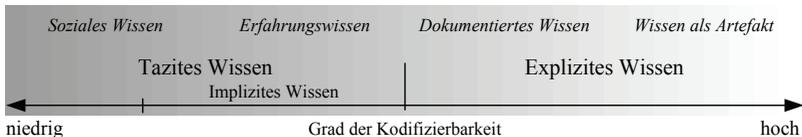


Abbildung 4-2: Kodifizierbarkeit von Wissen

WISSENSUMWANDLUNG IN ORGANISATIONEN

Im Wettbewerb zwischen Organisationen werden diejenigen langfristig überleben, welche aus den zur Verfügung stehenden Informations- und Wissensstrukturen erfolgsrelevantes Wissen schaffen und darauf Entscheidungen aufbauen. Die Internalisierung und Aktivierung bestehenden Wissens ist für Organisationen daher von großer Bedeutung. Wissensentstehung und -anwendung erfordern sowohl tazites als auch explizites Wissen, da Wissen durch die Verknüpfung dieser beiden Wissenstypen entsteht.

Im *SECI-Modell* (Socialisation, Externalization, Combination, Internalization) nach NONAKA und TAKEUCHI sind vier Arten der Wissensumwandlung in Organisationen aufgezeigt.⁴³ Die Konvertierung zwischen diesen Wissenstypen ist als kreativer Prozess zwischen den Beteiligten zu sehen, der einen Beitrag zu organisationalem Lernen leistet. Erst wenn Wissen in expliziter Form vorliegt, ist es für eine Organisation verfügbar.

We can know more than we can tell [Pola58].

Die Güte einer Organisationsstrategie kann mithin daran ermittelt werden, wie gut es gelingt, das relevante Wissen und die Erfahrung aller Organisationsmitglieder in Aktionspläne zu transformieren. Die im folgenden Abschnitt vorgestellte Disziplin *Wissensmanagement* widmet sich dieser Aufgabenstellung und untersucht Wege, Wissen in Organisationen zu erfassen und weiterzugeben [GrZa92, S. 10], [NoTa95, S. 62 ff.], [Nort05, S. 43 ff.], [Krcm05, S. 481 ff.], [Staa02], [Gass97, S. 150].

4.1.3 Wissensmanagement

Unter dem Begriff *Wissensmanagement* werden die Gestaltung und Steuerung der Prozesse, die in systematischer Weise die Wissensbasis einer Organisation beeinflussen, subsumiert. Wissensmanagement zielt dabei auf das Wissen von Menschen, auf das Teilen, Vermitteln und Weitergeben sowie eine Optimierung der Nutzung von Wissen. Dem Grad der Kodifizierbarkeit (vgl. Abschnitt 4.1.2) von Wissen kommt speziell im Hinblick auf ein IT-gestütztes Wissensmanagement eine besondere Bedeutung zu [Krcm05, S. 479 ff.], [KeMU06, S. 127 f.], [Riem05, S. 8 ff.], [Staa02].

Die nachhaltige Auseinandersetzung mit dieser Thematik ist für (betriebliche) Organisationen vor allem aus folgenden Gründen unerlässlich [Krcm05, S. 479 ff.]:

- Die Wissensintensität der Leistungen, die am Markt angeboten oder abgesetzt werden, wächst beständig. Vor allem Beratungs- und Zusatzdienstleistungen nehmen einen steigenden Stellenwert bei der Vermarktung von Produkten ein.
- Der intensive globale Wettbewerb verlangt hohe Innovationsraten sowie schnelle Neu- oder Weiterentwicklungen von Produkten und Dienstleistungen (sinkende Zykluszeiten).

⁴³ Eine detaillierte Erläuterung der *organisationalen Wissenserzeugung und -transformation* findet sich u. a. in [NoTa95, S. 62 ff.], [Nort05, S. 44 ff.], [Lehn06, S. 39 ff.] und [Krcm05, S. 481 ff.].

- Die geographische Verteilung einer Organisation, insbesondere bei der verteilten Organisationskonfiguration (vgl. Abschnitt 2.1.2), erfordert einen standortunabhängigen Zugriff auf die zum jeweiligen Zeitpunkt und den jeweiligen Nachfrager zugeschnittene Information.
- Der Mensch trägt durch sein persönliches (tazites) Wissen sowie durch kontinuierliches Lernen aktiv zur Entwicklung der Wissensbasis einer Organisation bei. Das in einer Organisation verfügbare Wissen, korreliert mit ihren Marktpotentialen (vgl. Punkt 1).

Das Thema Wissensmanagement wurde an dieser Stelle tangiert, um unter dem Aspekt der Kodifizierbarkeit von Wissen die Reichweite einer fundierten Daten- und Informationsbasis innerhalb einer betrieblichen Organisation aufzuzeigen. Für eine intensive Auseinandersetzung mit dem Terminus Wissensmanagement sowie eine Diskussion diverser Konzepte des Wissensaustauschs wird auf die Literaturquellen [Lehn06, S. 31 ff.], [Nort05, S. 174 ff.] und [Staa02] verwiesen.

4.2 Informationsversorgung von Entscheidungsträgern

Einem Individuum ist es aufgrund der heute verfügbaren Informationsmenge nahezu unmöglich, für eine spezifische Aufgabe alle produzierten Informationen zu erfassen und zu nutzen. Zudem weisen viele Informationen über die Zeit hinweg einen deutlichen Wertverlust auf [Wirt00, S. 3 f.], [KeMU06, S. 85 f.]. Diese Konstellation bedingt die Forderung nach einer für den jeweiligen Nachfrager adäquaten Informationsversorgung.

Im vorliegenden Abschnitt werden grundlegende Begriffe im Kontext der Informationsversorgung eingeführt, der Einfluss von Informationen auf Entscheidungen beleuchtet und die Auswirkungen von Informationspathologien erörtert.

4.2.1 Begriffsbestimmung und Abgrenzung

Um den Begriff *Informationsversorgung* genau abgrenzen zu können, werden zunächst verwandte Termini eingeführt und erläutert.

Die Art, Menge und Qualität von Informationen, die eine Person zur Erfüllung ihrer Aufgaben in einer bestimmten Zeit benötigt, wird als *Informationsbedarf*⁴⁴ bezeichnet. In diesem Zusammenhang ist zwischen dem *objektiven Informationsbedarf* (Art und Menge der Informationen, die ein Entscheidungsträger für eine Aufgabenerfüllung verwenden sollte), dem *subjektiven Informationsbedarf* (persönliche Einschätzung eines Entscheidungsträgers über die Art und Menge der für ihn relevanten Informationen) und der tatsächlich *nachgefragten Informationsmenge* (Teilmenge des ursprünglich geäußerten Informationsbedarfs) zu unterscheiden. Je unstrukturierter, komplexer oder veränderlicher eine Aufgabe ist, desto schwieriger erscheint die Annäherung und Abstimmung zwischen objektivem und subjektivem Informationsbedarf. Die Schnittmenge der Informationsnachfrage und dem vorliegenden *Informationsangebot* bildet die tatsächliche *Informationsversorgung*. Der objektiv zur Aufgabenerfüllung notwendige Ausschritt der Informationsversorgung stellt den *Informationsstand* dar [PiRW03, S. 81], [Metz02, S. 145 f.], [Böhn01, S. 116 ff.], [GrZa92, S. 17 f.]. Abbildung 4-3 visualisiert diesen Sachverhalt.

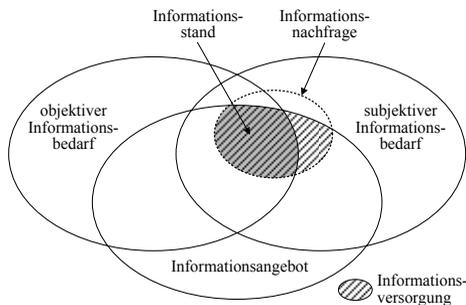


Abbildung 4-3: Informationsbedarf und Informationsversorgung [PiRW03, S. 82]

⁴⁴ Zur Bestimmung des Informationsbedarfs bei strategischen Aufgaben kommt vor allem die Methode der kritischen Erfolgsfaktoren zum Einsatz. Weitere Verfahren zur Ermittlung des Informationsbedarfs sind die Nebenprodukt-Technik, der Schlüsselindikatoren-Ansatz, der Total- und der Null-Ansatz [Metz02, S. 147 ff.].

INFORMATIONSKONGRUENZ

Steigt das Informationsangebot, ohne mit dem Informationsbedarf geeignet abgestimmt zu sein, entsteht ein Mangel im Überfluss, da die Informationsnachfrage nicht effektiv gedeckt werden kann. Gemäß dem informationslogistischen Prinzip (vgl. Abschnitt 2.2.2) ist jedoch eine Deckungsgleichheit von Informationsangebot, -nachfrage und -bedarf anzustreben. Ist diese erreicht, wird von *Informationskongruenz* gesprochen [BeBe96, S. 18], [Böhn01, S. 117 f.].

Die vorliegende Arbeit soll transitiv einen Beitrag zur Erreichung von Informationskongruenz bei Entscheidungsaufgaben leisten. Die Seite des Informationsbedarfs bleibt dabei unberührt. Stattdessen wird im Rahmen der Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen ein vollständiger und homogener Überblick über das bestehende Informationsangebot angestrebt. Hierdurch können von einem Entscheidungsträger die verfügbaren, relevanten Informationen direkt erkannt und zielgerichtet abgerufen werden. Es wird demzufolge eine Verbesserung der Informationsversorgung bzw. des Informationsstands erzielt.

4.2.2 Entscheidungsaufgaben

Einleitend wird die Struktur einer *Entscheidungsaufgabe* anhand eines Input-Output-Modells⁴⁵ beschrieben.

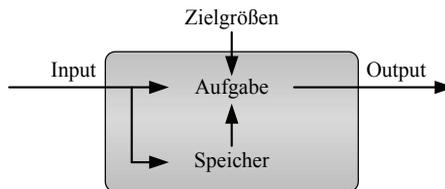


Abbildung 4-4: Struktur einer Entscheidungsaufgabe [FeSi06, S. 33]

⁴⁵ Ein *Input-Output-System* dient zur Beschreibung einer so genannten Black-Box, deren innere Struktur unbekannt oder irrelevant ist. Für die Beschreibung wird ausschließlich die Beziehung zwischen Eingabe- und Ausgabemenge des Systems herangezogen [FeSi06, S. 14].

Die von einer Aufgabe erzeugten Informationen, hier als *Output* deklariert, werden aus den für eine Aufgabe bereitgestellten Informationen, dem *Input*, abgeleitet. Zudem beeinflussen gespeicherte Informationen aus früheren Aufgabendurchführungen die Outputgenerierung (*Speicher*). Die im Speicher abgelegten Informationen werden im Rahmen der Aufgabenerfüllung modifiziert oder erweitert.

Bei der Bearbeitung einer Entscheidungsaufgabe sind zusätzliche Input-Informationen in Form von *Führungs-* oder *Zielgrößen* zu beachten. Der Output einer Entscheidungsaufgabe umfasst Entscheidungswerte, welche optimal bzw. suboptimal bezüglich dieser Größen sind [FeSi06, S. 34].

INFORMATIONSNUTZUNG BEI ENTSCHEIDUNGSAUFGABEN

Das Ergebnis einer Entscheidungsaufgabe (Output) hängt von der Nutzung der verfügbaren Informationen ab [Reil83, S. 105]. Eine Information wird bei einer Entscheidungsfindung berücksichtigt,

- je zentraler ihre Bedeutung für die Aufgabenerfüllung einer Person ist.
- je deutlicher das Informationsangebot mit relevanten Planungs-, Kontroll- und Berichtssystemen verknüpft ist.
- je leichter diese zugänglich ist. Dies umfasst organisatorische, räumliche und intellektuelle Aspekte.
- je kompakter und intuitiv verständlicher das Ergebnis einer Informationsnachfrage dargestellt wird.
- je größer das Vertrauen in die Informationsquelle ist.
- je stärker sie die persönlichen Ziele des Entscheidungsträgers begünstigt.

Diese Faktoren können auf ein Kosten-Nutzen-Kalkül des Entscheidungsträgers zurückgeführt werden: Eine Informationen wird umso eher genutzt, je geringer die Kosten für Ihre Beschaffung und je höher der Nutzen aus ihrer Beachtung ist [PiRW03, S. 84], [Reil83, S. 114 ff.].

Die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen kann einen wesentlichen Beitrag zur Senkung der Kosten bei der Informationsbeschaffung leisten (z. B. durch den Wegfall aufwändiger Recherche- und Konsolidierungsarbeit beim Zusammenführen von Informationen aus unterschiedlichen DWH-Systemen).

4.2.3 Informationspathologien

Eine ursprünglich akkurat und klar formulierte, sowie für einen konkreten Kontext relevante Information, kann durch die Übermittlung vom Sender zum ursprünglich vorgesehenen Empfänger falsch übersetzt, verkürzt oder sogar vollständig zurückgehalten werden. Um eine funktionierende, ganzheitliche Informationsversorgung sicherzustellen sind daher neben der Wahl eines adäquaten Kommunikationsmediums auch mögliche Fehlfunktionen in der organisationalen Wissensverarbeitung zu berücksichtigen. Solche Fehlfunktionen, die auch als *Informationspathologien* bezeichnet werden, beruhen auf grundsätzlich vermeidbaren Fehlern [Scho04, S. 24 ff.], [PiRW03, S. 86 ff.], [Wile67, S. 41 f.]. Hierzu zählen

- *produzierbare* Informationen, die nicht produziert,
- *beschaffbare* Informationen, die nicht beschafft,
- *vorhandene* Informationen, die nicht (korrekt) übermittelt oder
- *vorliegende* Informationen, die nicht (korrekt) verarbeitet wurden.

Informationspathologien treten insbesondere in Organisationen mit mehreren hierarchischen Ebenen sowie zahlreichen oder geographisch verteilten Einheiten auf. Aufwändige Informationssuchen, Unkenntnis existierender Informationsangebote, zu lange Informationswege oder eine verspätete Informationsübermittlung sind als mögliche Ursachen zu identifizieren. Ein Konkurrenzdenken zwischen Abteilungen oder Sparten erhöht zudem die Gefahr von Informationspathologien bspw. durch bewusste Informationszurückhaltung (vgl. implizites Wissen; Abschnitt 4.1.2) oder irreführende Informationsdarstellung [Scho04, S. 30 ff.], [Nort05, S. 47], [Wile67, S. 175 ff.]. Die Übermittlung zu vieler Informationen (*Information Overload*) kann im weiteren Sinne ebenfalls als eine Art von Informationspathologie ausgelegt werden, da auch ein Überfluss an Informationen zum Ausschluss entscheidungsrelevanter Informationen führen kann.

AUFTRETEN VON INFORMATIONSPATHOLOGIEN

Hinsichtlich des Auftretens von Informationspathologien sind drei Dimensionen zu unterscheiden [PiRW03, S. 86 ff.]:

- Bei der *aktorbezogenen Informationspathologie* fehlen dem Informationsempfänger Anknüpfungspunkte an sein vorhandenes Wissen, sodass eine potentiell neue Information nicht verstanden wird. Aktorbezogene Informationspathologien beruhen auf Unzulänglichkeiten, die auf individuelle menschliche Eigenschaften zurückzuführen sind.
- Zur *interaktionsbezogenen Dimension* von Informationspathologien zählen Verständigungsbarrieren aufgrund unterschiedlicher Grundkenntnisse, Fachsprachen o. Ä. Interaktionsbezogene Informationspathologien stellen eine Form fehlerhafter Kommunikationsprozesse dar.
- Entspricht eine Information nicht den bestehenden Vorannahmen ihres Adressaten oder passt nicht zu dessen präferiertem Informationstyp (quantitative vs. qualitative Informationen), so wird von einer *wissensbasierten Informationspathologie* gesprochen.

Informationspathologien können auch aufgrund einer nicht konsolidierten Informationsversorgung durch heterogene DWH-Systeme entstehen. Insbesondere dem fälschlichen Ausschluss einer entscheidungsrelevanten Information durch

- Nicht-Beschaffung (z. B. infolge fehlender Transparenz über das verfügbare Informationsangebot),
- verzerrte Übermittlung und falsche Verarbeitung (z. B. durch fehlerhafte Datentransformation bei der Datenweitergabe) sowie
- inkorrekte Interpretation (z. B. aufgrund individueller Begriffssysteme)

kann mit der Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen begegnet werden.

4.3 Kontextinformationen bei Data-Warehouse-Systemen

Ein Individuum kann auf effektive Weise seine Ansichten gegenüber anderen Menschen äußern sowie auf eine Mitteilung adäquat reagieren. Hierfür zeigen sich einerseits seine reichhaltige Sprache und andererseits sein Allgemeinwissen sowie das inhärente Verständnis alltäglicher Situationen verantwortlich. Menschen machen sich bei ihrer Kommunikation implizite, situationsbedingte Informationen, im allgemeinen Sprachgebrauch als *Kontext* betitelt, zu Nutze und erhöhen damit die Bandbreite ihrer Konversation.

Diese Fähigkeit lässt sich jedoch nicht ohne weiteres auf die Kommunikation zwischen Mensch und Maschine übertragen. Maschinen sind zunächst nicht in der Lage, aus dem Kontext der Mensch-Maschine-Kommunikation⁴⁶ weiterführende Informationen abzuleiten. Die Bereitstellung und der Zugang für Maschinen zu Kontextinformationen erhöht die Reichhaltigkeit der Mensch-Maschine-Kommunikation [AaNy95, S. 9], [DeAb99].

DEFINITION VON KONTEXT

Eine einheitliche Darstellung des Begriffs *Kontext* findet sich in der Literatur nicht. ISSELHORST hat eine Reihe von Kontextdefinitionen und -typisierungen untersucht und anhand seiner Ergebnisse konstatiert, dass sich je nach Anwendungszweck neue Zusammensetzungen und Definitionen von Kontext ergeben. Bekannte, generische Kontexttypisierungen (z. B. [DeAb99]) können dem Anspruch der Allgemeingültigkeit nicht gerecht werden [Isse07, S. 95 ff.].

Demzufolge ist eine für die Untersuchungssituation dieser Arbeit zutreffende Kontextdefinition und -typisierung zu erarbeiten. Im Hinblick darauf wird zunächst die von ISSELHORST vorgeschlagene Kontextdefinition für den Aufgabenbereich der Unternehmensführung eingeführt und auf ihre Anwendbarkeit geprüft.

⁴⁶ Grundlagen der *Mensch-Maschine-Kommunikation* finden sich z. B. bei [FeSi06, S. 54].

KONTEXT IM AUFGABENBEREICH DER UNTERNEHMENSFÜHRUNG

Folgende Definition des Begriffs Kontext wird von ISSELHORST für den Aufgabenbereich der Unternehmensführung vorgeschlagen. Der Begriff Unternehmensführung ist dabei umfassend zu verstehen und nicht auf eine spezifische Funktion oder Institution zu beziehen.

Zum Kontext der Unternehmensführung gehört jede Art von Information, die dazu verwendet werden kann, das Umfeld der Unternehmensführung zu charakterisieren. Die Unternehmensführung ist strukturiert durch Aufgaben bestehend aus Aufgabenobjekt und Lösungsverfahren, sowie durch Personen und Anwendungssysteme als Aufgabenträger [Isse07, S. 111].

Das Kriterium der Zugehörigkeit einer Information zum Kontext der Unternehmensführung soll sich an der Relevanz einer Information für die Aufgabe eines Entscheidungsträgers orientieren. ISSELHORST hat für den Bereich der Unternehmensführung die Kontexttypen *Aufgabenkontext*, *Unternehmenskontext*, *Sozialkontext*, *Nutzerkontext* sowie *Interaktionskontext* identifiziert [Isse07, S. 105 ff.]. Diese werden im Folgenden mit Detaillierungen aus dem von HENRICH und MORGENROTH beschriebenen Kontextmodell vorgestellt [HeMo07]:

- Der *Aufgabenkontext* charakterisiert die momentan zu bearbeitende (Teil-)Aufgabe einer Person im Rahmen der Unternehmensführung. Der Aufgabenkontext bestimmt den Informationsbedarf hinsichtlich des Aufgabenobjekts und -ziels.
- Der *Unternehmenskontext* ist durch das Umfeld des Unternehmens (z. B. Wettbewerbsumfeld, Produkt-/Marktkombination) gekennzeichnet. Der Kontext eines Unternehmens ist durch dessen Unternehmensstrategie (insbesondere Produkt- und Marktausrichtung) determiniert.
- Angaben über Kommunikationsbeziehungen zwischen Mensch und Maschine werden dem *Interaktionskontext* zugeordnet. Bei Führungsaufgaben umfassen diese bspw. Benutzeraktionen, Nachrichten sowie direkte Informationsanfragen.
- Der *Nutzerkontext* entspringt der individuellen Situation einer Person. Er ist in den *physikalischen* (z. B. geographische Daten über den Arbeitsplatz) und den *organisatorischen* Kontext (z. B. Stelle und Rolle einer Person in einer Organisation) sowie in die *Profildimension* (z. B. Wissensgebiete, Fähigkeiten einer Person) zu unterteilen.

- Der *Sozialkontext* enthält Informationen über miteinander interagierende Personen. Er ergibt sich aus der Mensch-Mensch-Kommunikation sowie der Beziehung der beiden Kommunikationspartner bei der Durchführung einer Aufgabe.

Die vorgestellten Kontextarten determinieren das Umfeld der Unternehmensführung vollständig und leisten einen Beitrag, um (latent) vorliegende Informationsbedürfnisse identifizieren zu können [Isse07, S. 108], [HeMo07].

KONTEXT BEI DER INFORMATIONSVERSORGUNG DURCH EIN DATA-WAREHOUSE

Die Kontextdefinition von ISSELHORST konzentriert sich auf den *Informationsbedarf* im Aufgabenbereich der Unternehmensführung (Top-Down-Betrachtung). Die vorliegende Arbeit fokussiert das durch DWH-Systeme bereitgestellte *Informationsangebot* (Bottom-Up-Betrachtung). Durch die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen soll für einen Entscheidungsträger Transparenz über das durch DWH-Systeme offerierte Informationsangebot geschaffen und somit seine Informationsversorgung optimiert werden. Zu diesem Zweck wird eine vollständige und semantisch kongruente Beschreibung des realen Informationsangebots angestrebt. Zu dieser gehören auch Angaben über den Kontext der Daten eines DWH.

Aufgrund der beiden unterschiedlichen Blickwinkel auf die Informationsversorgung von Entscheidungsträgern (Top-Down vs. Bottom-Up) erscheint die von ISSELHORST vorgeschlagene Kontextdefinition mit den hier vorliegenden Rahmenbedingungen nicht vollständig kongruent. Jedoch können in Bezug auf die soeben eingeführten Kontexttypen einige Überschneidungsbereiche erkannt werden. Für den Kontext eines DWH wird daher folgende Definition vorgeschlagen:

Zum Kontext eines Data-Warehouse gehört jede Art von Information, welche die gespeicherten Daten charakterisiert (Kontextart Informationsangebot) oder eine bedarfsgerechte Bereitstellung der gespeicherten Daten unterstützt (Kontextart Informationsbereitstellung).

Die beiden Kontextarten *Informationsangebot* und *Informationsbereitstellung* sind überschneidungsfrei. Aufbauend auf den von ISSELHORST im Umfeld der Unternehmensführung identifizierten Kontexttypen und differenziert nach den vorgestellten Kontextarten werden an dieser Stelle die Kontexttypen *Umweltkontext*, *Domänenkontext*, *Nutzerkontext* und *Aufgabenkontext* eingeführt und beschrieben.

Kontextart Informationsangebot:

Der Kontextart Informationsangebot sind Informationen zuzuordnen, welche die in einem DWH abgelegten Daten z. B. hinsichtlich ihres Ursprungs beschreiben. Dabei sind die Kontexttypen *Umweltkontext* und *Domänenkontext* zu unterscheiden.

- Dem *Umweltkontext* ist jede Art von Information zuzuordnen, die der Umwelt einer Organisation entstammt und über diese Angaben bereitstellt. Hierzu zählen bspw. Informationen über gesetzliche Rahmenbedingungen (z. B. Besteuerung, Gebühren) sowie Entwicklungen auf gesamt- oder finanzwirtschaftlicher Ebene (z. B. Brutto-Inlandsprodukt, Börsenkurse).
- Der *Domänenkontext* charakterisiert aus fachlicher Sicht organisationsinterne Bereiche oder Disziplinen, aus denen die Daten eines DWH abgeleitet sind (*betriebliche Diskurswelt*; z. B. [FeSi06, S. 5 f.]). Er tritt in unterschiedlicher Granularität auf. Innerhalb einer verteilten Organisation sind dabei Angaben über Fachabteilungen (z. B. Produktion, Lagerhaltung) auf unterster Ebene des Domänenkontexts zu identifizieren. Dem schließen sich auf nächst höherer Stufe Kontext-Informationen über die Geschäftseinheit (Sparte) an (z. B. Produktportfolio). Die oberste Ebene umfasst gesamtorganisatorische Informationen (z. B. Leistungsspektrum der verteilten Organisation). Abhängig von der Hierarchietiefe der Organisationskonfiguration variiert die Anzahl der Ausprägungsstufen des Domänenkontexts.

Kontextart Informationsbereitstellung:

Angaben der Kontextart Informationsbereitstellung sollen eine bedarfsgerechte Bereitstellung der Informationen für Entscheidungsträger fördern. Dies kann einerseits im Hinblick auf bestimmte DWH-Anwender (Nutzerkontext) und andererseits in Bezug auf spezifische Aufgaben erfolgen, welche aufgrund der verfügbaren Datenbasis potentiell unterstützt werden können (Aufgabenkontext).

- Analog zur Definition des *Nutzerkontexts* bei ISSELHORST bezieht sich die hier gewählte Begriffsabgrenzung auf Merkmale der individuellen Situation eines Informationsnachfragers (z. B. in Person eines Führungsverantwortlichen oder Entscheidungsträgers). Zu deren Beschreibung zeigen sich vor allem Angaben über organisatorische oder funktionale Zuständigkeiten kennzeichnend (z. B. oberes oder mittleres Management, strategische oder operative Planung). Ergänzend können Informationen zur Profildimension (z. B. Standard- oder Power-

User) Beachtung finden. Ziel der Verknüpfung dieses Kontexttyps mit der Beschreibung der Daten eines DWH ist die Beantwortung der Frage, für welche Informationskonsumenten eine Information interessant sein kann bzw. zugreifbar sein soll. Diese Kontextinformation kann bspw. auch bei der Ausgestaltung eines Berechtigungskonzepts mit einfließen. Die physikalische Komponente des Nutzerkontexts (z. B. in Form geographischer Daten über den momentanen Arbeitsplatz) wird an dieser Stelle nicht einbezogen.

- Der *Aufgabenkontext* charakterisiert, für welche Tätigkeit oder Aufgabe die Daten eines DWH relevant erscheinen (z. B. Analyse der Umsatzzahlen, Personalstandsentwicklung). Eine spezifische Verwendung der Daten kann im DWH-Umfeld nicht antizipiert werden. Jedoch kann mithilfe von Angaben dieses Kontexttyps ein rasches Auffinden der Daten unterstützt werden, die für einen bestimmten Aufgabentyp hilfreich sind. Die angestrebte Kongruenz zwischen Informationsangebot und Informationsbedarf wird somit gefördert (vgl. Abschnitt 4.2.1).

Die vorgeschlagene Kontextdefinition ermöglicht innerhalb der hier vorgegebenen Bottom-Up-Betrachtungsweise eine umfassende Beschreibung des Umfelds eines DWH. Werden bei der Untersuchung von Kontext auch die individuellen Informationsbedarfe eines Entscheidungsträgers berücksichtigt (Top-Down-Betrachtung), sind ggf. weitere Kontexttypen zu identifizieren und in die herausgearbeitete Definition mit aufzunehmen. Für die Erreichung des vorliegenden Untersuchungsziels kann diese Sichtweise jedoch ausgeklammert werden.

Nachstehende Abbildung ordnet die vier Kontexttypen zu einem DWH in die bekannte, regelkreisorientierte Darstellung eines betrieblichen Systems ein (vgl. Abschnitt 2.3.3).

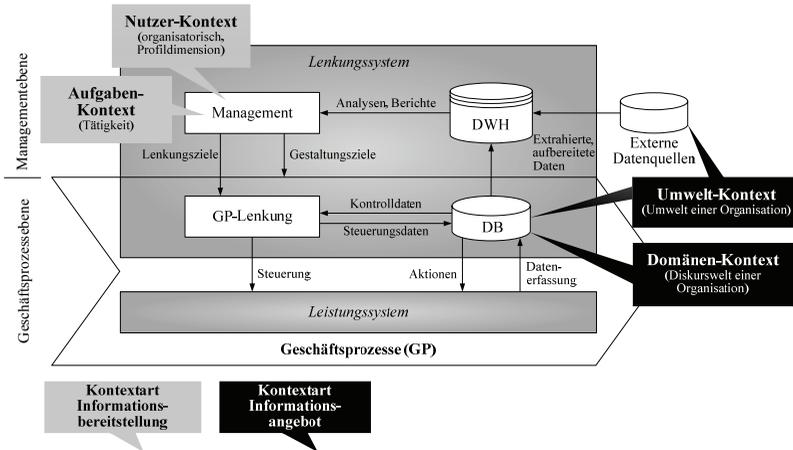


Abbildung 4-5: Kontexttypen im Umfeld eines DWH

BEDEUTUNG VON KONTEXTINFORMATIONEN IN EINEM DATA-WAREHOUSE

Die korrekte Interpretation von Daten erfordert Wissen über den Kontext, dem die Daten entstammen (vgl. Abschnitt 4.1). Die Erfassung von Kontextinformationen beim Aufbereiten der Daten im DWH bildet hierfür die Grundlage. Bei der Informationsbereitstellung sind dem Anwender diese ergänzenden Informationen mit anzubieten.

Durch die zusätzliche Bereitstellung von Kontextinformationen ergeben sich – neben der Unterstützung bei der Erfassung der Datensemantik – neue Möglichkeiten der Selektion relevanter Daten (z. B. Suche nach erforderlichen Kennzahlen für den Aufgabenbereich der Umsatzanalyse). Des Weiteren eröffnet sich die Option, dass aufgrund von Angaben über den Kontext ursprünglich als nicht relevant erachtete Daten bei der Informationsnachfrage durch einen Anwender mit einbezogen werden.

Kontextinformationen stellen einen Mehrwert sowohl für die Nutzer als auch für die Verantwortlichen eines DWH-Systems dar, da durch diese die Mensch-Maschine-Kommunikation eine Erhöhung der Reichhaltigkeit auf Maschinenseite erfährt. Die Verfügbarkeit von Kontextinformationen über ein DWH kann sich folglich unterstützend im Hinblick auf die Überwindung semantischer Heterogenität bei multiplen DWH-Systemen auswirken.

4.4 Zusammenfassung

Dieses Kapitel widmete sich Führungsinformationen in betrieblichen Organisationen. Nach der grundlegenden Einführung der Beziehungen zwischen Daten, Information und Wissen wurden Aspekte der Kodifizierbarkeit von Wissen sowie des Wissensmanagements beleuchtet. Dabei konnte festgestellt werden, dass die Bedeutung von Daten häufig nur durch tazes Wissen vollständig erfasst werden kann. Um Informationsnachfragern eine korrekte Interpretation der in einem DWH gespeicherten Daten zu ermöglichen, sind auf Basis des in einer Organisation verfügbaren Wissens reichhaltige Beschreibungen dieser Daten zu erstellen und anzubieten.

Untersuchungen bezüglich der Informationsversorgung von Entscheidungsträgern bildeten einen weiteren Schwerpunkt dieses Kapitels. Insbesondere der Einfluss von Informationen bei der Bearbeitung von Entscheidungsaufgaben sowie die Auswirkungen von Informationspathologien standen hierbei im Fokus. Abschließend wurde eine Betrachtung des Terminus Kontext bei einem DWH vorgenommen. Speziell für ein DWH erfolgte die Ausarbeitung einer Kontextdefinition mit zugehörigen Kontexttypen.

Zusammenfassend ist festzustellen, dass je reichhaltiger die Daten eines DWH beschrieben sind, desto förderlicher ist dies für die angestrebte Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen. Letztere dient auch der Erreichung von Informationskongruenz bei Entscheidungsaufgaben, indem ein vollständiger, homogener Überblick über das reale Informationsangebot geschaffen wird. Die Wahrscheinlichkeit der Beachtung einer relevanten Information bei der Entscheidungsfindung kann somit durch geringe Kosten bei der Informationsbeschaffung gesteigert und gleichzeitig das Auftreten von Informationspathologien reduziert werden.

5 Heterogenität in Datenschemata von Data-Warehouse-Systemen

Die vollständige Erfassung der vorliegenden Untersuchungssituation verlangt auch eine detaillierte Betrachtung von Heterogenitätsaspekten. Im Hinblick darauf wird zunächst auf Grundlagen der Semiotik eingegangen und die Vagheit natürlichsprachlicher Beschreibungen herausgestellt. Anschließend werden Bedingungen für eine vollständige Beschreibung eines Begriffs erläutert. Dabei werden Probleme bei der Begriffsbenennung und -interpretation sowie Ansätze zu deren Vermeidung und Auflösung dargelegt. Diesen Ausführungen folgen auf die Heterogenität von Datenschemata ausgerichtete Untersuchungen. Neben der Diskussion von Heterogenitätsarten werden deren Ursachen sowie Kriterien der Überwindung von Heterogenität dargestellt. Bestandteil der weiteren Ausführungen ist die Erörterung zentraler Elemente und Ergebnisse des Forschungsgebiets Schema-Management. Den Abschluss des Kapitels bildet eine Zusammenfassung notwendiger Merkmale, um Metadaten eines DWH-Systems semantisch reichhaltig beschreiben zu können.

5.1 Grundlagen der Semiotik

Informationen sind für die Bearbeitung einer Aufgabe eine notwendige Voraussetzung (vgl. Abschnitt 4.2.2). Damit Informationen ihre vorgesehenen Empfänger erreichen, ist eine Verständigung bzw. Kommunikation mit dem Informationsanbieter (Sender) erforderlich. Die *Semiotik* ist die wissenschaftliche Disziplin, die sich der Erforschung der Gegenstände und Funktionsweisen von Kommunikationsvorgängen widmet. Hierbei werden die drei Ebenen

- *Syntaktik* (Analyse von Zeichen und deren Beziehungen untereinander [Morr79, S. 32 ff.]),
- *Semantik* (Betrachtung der Zeichen einschließlich ihrer Bedeutung [Morr79, S. 42 ff.]) und
- *Pragmatik* (Untersuchung der (situationsabhängigen) Auswirkungen von Zeichen auf ihre Empfänger [Morr79, S. 52 ff.]) unterschieden.

Für einen Kommunikationsvorgang gilt somit auf syntaktischer Ebene die korrekte Übertragung der Zeichen sicherzustellen, auf semantischer Ebene den Zeichen die richtige Bedeutung beizumessen und auf pragmatischer Ebene die intendierte Absicht des Senders zu erfassen [PiRW03, S. 89 ff.]. Die semiotischen Ebenen eines Kommunikationsprozesses sind nicht voneinander zu trennen und bauen aufeinander auf (vgl. Abbildung 5-1).

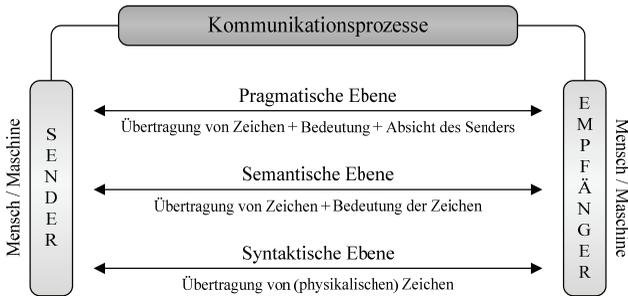


Abbildung 5-1: Semiotische Ebenen der Informationsübertragung [PiRW03, S. 90]

Die drei Ebenen befassen sich mit Zeichen, deren Beziehungen und Regeln. Die Semiotik wird daher auch als Lehre von den Zeichen oder als Zeichentheorie definiert. Demgemäß entspricht eine Sprache einem System von Zeichen, welches folgendem Paradigma unterliegt:

- Die *Syntax* wird als eine Menge von Vorschriften und Regeln angesehen, die den formal korrekten Aufbau von Wörtern und Sätzen angeben.
- Die *Semantik* beschreibt die Bedeutung eines syntaktisch korrekten Satzes. Die Bedeutung sollte eindeutig sein und von jedem, der die Sprache beherrscht, verstanden werden.
- Die *Pragmatik* umfasst subjektive Aspekte individueller Benutzer einer Sprache. Sie legt fest, was derjenige, der ein Zeichen benutzt oder entwickelt, damit meint oder meinen kann und bestimmt somit das Ziel einer sprachlichen Handlung.

Das Ziel einer sprachlichen Handlung zwischen zwei Zeichenbenutzern kann jedoch nicht generell als identisch angenommen werden. Die Pragmatik ist daher als Mittel zur Spezifikation einer Sprache ungeeignet. Im Kontext *künstlicher Sprachen*, die zur Bedienung und Programmierung von Maschinen entwickelt werden, sind lediglich Syntax und Semantik einer Sprache von Bedeutung. Die subjektive Bedeutung von Zeichen wird vernachlässigt [SaSH95, S. 98 ff.], [BrEG06, S. 4 ff.].

Bei der Übermittlung der Semantik von Zeichen mittels natürlicher Sprache kann beim Empfänger Unklarheit über deren korrekte Interpretation entstehen. Dies ist insbesondere infolge missverständlicher Formulierungen oder bei nicht vollständiger Beherrschung der gewählten, natürlichen Sprache zu beobachten (vgl. auch Abschnitt 5.2.2). Mithilfe *formaler Sprachen*⁴⁷ wird versucht, die Semantik formal, d. h. mathematisch exakt, widerspruchsfrei und vollständig zu definieren. Auf diese Weise wird die Vagheit einer natürlichsprachlichen Beschreibung umgangen. Die *formale Semantik* beschreibt ein wichtiges Gebiet der Informatik, das sich mit den Möglichkeiten der Formalisierung von künstlichen und natürlichen Sprachen beschäftigt [SaSH95, S. 98 ff.].

Um die bestehende Unschärfe der natürlichen Sprache zu begründen und den Bedarf einer formalen Beschreibung von Begriffen herauszustellen wird anschließend das semiotische Dreieck eingeführt und erläutert.

SEMIOTISCHES DREIECK

Das semiotische Dreieck illustriert die Interaktion zwischen Worten (allgemein *Symbol* oder Zeichen), *Begriffen* und *Dingen* aus der Realwelt (vgl. Abbildung 5-2).

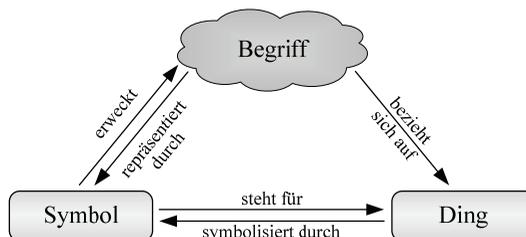


Abbildung 5-2: Semiotisches Dreieck (in Anlehnung an [LeNa07a, S. 273])

Ein *Begriff* bezieht sich auf ein bestimmtes *Ding* oder eine Klasse von Dingen. Dieser ist individuell ausgestaltet und existiert lediglich in der Vorstellung eines Menschen. *Symbole* (z. B. Worte) werden genutzt, um über Dinge zu kommunizieren. Sie können jedoch die Essenz eines Begriffs (=die Referenz auf ein Ding) oder das referenzierte Ding selbst nicht vollständig erfassen. Abhängig vom jeweiligen Erfahrungshintergrund und Kontext können aus diesem Grund unterschiedliche Empfänger einer Nachricht zu abweichenden Resultaten bezüglich der Korrespondenz zwischen einem Wort und den möglichen Begriffen und Dingen in der Realwelt gelangen [Staa02], [BrEG06, S. 3 ff.].

⁴⁷ *Formale Sprachen* werden in [Rech06] grundlegend eingeführt.

Auf DWH-Systeme übertragen bedeutet dies, dass natürlichsprachliche Beschreibungen eine eindeutige Definition der betriebswirtschaftlichen (Fach-)Begriffe und Kennzahlen nicht garantieren. Eine individuelle Interpretation durch Entscheidungsträger führt zu Unstimmigkeiten über den Aussagegehalt von Berichtsergebnissen. Auf semantischer Ebene muss daher festgelegt werden, was genau unter einem bestimmten (Fach-)Begriff oder einer Kennzahl verstanden wird [HoKB01, S. 43], [Bess04]. Diese Definition ist anschließend formal zu erfassen.

Im nachfolgenden Abschnitt werden charakteristische Merkmale eines Begriffs untersucht. Diese sollen beitragen, notwendige Kriterien für eine exakte Begriffsdefinition aufzudecken.

5.2 Merkmale eines Begriffs

Ein *Begriff* beschreibt und gliedert einen Ausschnitt der Realwelt (Gegenstandsbereich).

Ein Begriff [...] ist das Gemeinsame, das Menschen an einer Mehrheit von Gegenständen feststellen und als Mittel des gedanklichen Ordners („Begreifens“) und darum auch zur Verständigung verwenden [Wüst91, S. 8].

Zur genauen Erfassung der Spezifika eines Begriffs werden nachfolgend die Repräsentationsebenen eines Begriffs erörtert. Anschließend erfolgt eine Diskussion häufig auftretender sprachlicher Defekte. Des Weiteren wird mit der Methode zur Rekonstruktion von Fachbegriffen ein Vorgehen zum Aufbau einer konsolidierten Terminologie vorgeschlagen.

5.2.1 Repräsentationsebenen eines Begriffs

Die Repräsentationsebenen eines Begriffs dienen der exakten Spezifikation eines Begriffs. CARNAP hat erstmals einen Begriff als 3-Tupel bestehend aus *Begriffswort*, *Intension* und *Extension* definiert [Carn47, S. 1 ff.], [WeOr80, S. 15 ff.], [Ortn00, S. 102 ff.]:

- *Begriffswort (Begriffsbenennung)*: Ein Begriffswort benennt und fixiert einen Begriff.
- *Intension (Begriffsinhalt)*: Die Intension eines Begriffs umfasst dessen Merkmale (Intension nach *innen*) sowie seine Beziehungen zu anderen Begriffen (Intension nach *außen*). Anhand dieser Kriterien kann die Zugehörigkeit eines Ausdrucks (Objekts) zu einem Begriff überprüft werden.

- *Extension (Begriffsumfang)*: Zur Extension eines Begriffs gehören nach *außen* betrachtet, jene konkreten oder abstrakten Objekte, die unter den Begriff fallen, sowie mit Blick nach *innen* singuläre Beschreibungen dieser Objekte.

Die nachstehende Abbildung gibt die Struktur dieses *Begriffsmodells* in Anlehnung an [Ortn97, S. 31], [Ortn00, S. 102 ff.] und [Lehm01, S. 124 f.] wieder.

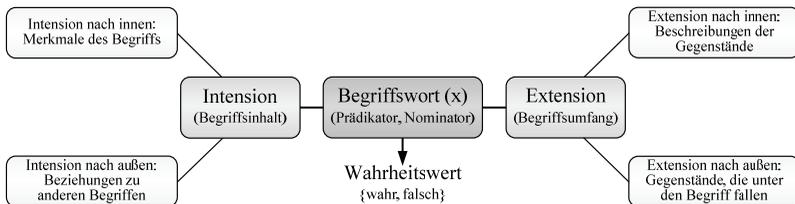


Abbildung 5-3: Begriffsmodell

(in Anlehnung an [Ortn97, S. 31], [Ortn00, S. 102], [Lehm01, S. 124])

In Ergänzung zu diesem Begriffsmodell soll die Betrachtungsweise eines Begriffs nach FREGE Beachtung finden. FREGE umschreibt einen Begriff als eine Funktion, die stets einen Wahrheitswert liefert (Begriffswort: $x \rightarrow \{w, f\}$). Durch das Einsetzen eines Eigennamens in die Funktion wird mit dem Begriffswort eine Aussage gebildet. Ein durch den Eigennamen benanntes Objekt gehört genau dann zu einem Begriff (zu seiner Extension), wenn die aufgestellte Aussage wahr ist [Freg75, S. 18 ff.], [Ortn00, S. 102], [Lehm01, S. 125].

BESCHREIBUNGS- UND ORGANISATIONSMITTEL FÜR BEGRIFFE

Um Begriffe in Organisationen korrekt einführen, verwalten und nutzen zu können, bedarf es bestimmter Beschreibungs- und Hilfsmittel. Nachstehend werden in diesem Zusammenhang geläufige Ausdrücke und Systeme vorgestellt:

- Von einer *Definition* (Begriffsbestimmung) im weiteren Sinne wird gesprochen, wenn ein Begriff durch bereits eingeführte Begriffe beschrieben wird. Dabei wird zwischen einer Inhalts- und Umfangsdefinition⁴⁸ unterschieden. Eine zweckmäßige Benennung eines Begriffs ist erst dann möglich, wenn dieser geklärt und definiert worden ist [Wüst91, S. 33 ff.].

⁴⁸ *Inhalts- und Umfangsdefinition* sind in [Wüst91, S. 34 f.] spezifiziert.

- Beim Vergleich von Begriffen, lassen sich zwischen diesen sowohl logische als auch ontologische Beziehungen feststellen. Über Beziehungen, die bspw. anhand der Ähnlichkeiten in den Begriffsinhalten bestimmt sind, lassen sich Ordnungs- oder Verknüpfungsrelationen definieren. Die zur Demarkation eines Begriffs einschränkende Merkmale dienen dabei zum Aufbau von *Begriffssystemen*⁴⁹ und somit zur Strukturierung und Ordnung von Begriffen. Ein Begriffssystem repräsentiert das Fachwissen, über das Personen verfügen müssen, die in der entsprechenden Domäne Arbeiten effizient erledigen oder Entscheidungen treffen müssen [Ortn97, S. 17 ff.], [Wüst91, S. 9 ff.].
- Unter einer *Terminologie* wird die Gesamtheit der Fachausdrücke eines Fachgebiets verstanden. Die Arbeit an und mit einer Terminologie, das so genannte *Terminologiemanagement*⁵⁰, bildet die Grundlage für eine effektive und erfolgreiche Dokumentation der Fachausdrücke. Diese ist insbesondere bei einer arbeitsteiligen Aufgabendurchführung oder einer verteilten Organisationsstruktur von zentraler Bedeutung. Die Terminologiearbeit kann in die Stufen Sammlung, Analyse und Festlegung von Begriffen unterteilt werden. Für weiterführende Informationen zu den Prinzipien der Terminologiearbeit wird auf die Literaturquellen [Lehm01, S. 133 ff.], [Wüst91, S. 8 ff.], [Hell97, S. 43 ff.] sowie Abschnitt 5.2.4 hingewiesen.

IMPLIKATION AUF DIE UNTERSUCHUNGSSITUATION

Die vorliegende Arbeit konzentriert sich auf die Betrachtung der Schemaebene, um semantische Heterogenität bei multiplen DWH-Systemen zu überwinden (vgl. Abschnitt 1.2.2). Die Datenebene und demzufolge die Extension eines Begriffs wurden eingangs ausgeklammert.

⁴⁹ Anstelle des Ausdrucks *Begriffssystem* kann auch von einer *Begriffsklassifikation* gesprochen werden [Wüst91, S. 22]. Weiterführende Darstellungen zu Begriffssystemen bietet [Wüst91, S. 9 ff.].

⁵⁰ *Terminologiemanagement* umfasst die Planung, Steuerung und Organisation der in einer spezifischen Bedeutung kontrolliert verwendeten Wörter und Fachausdrücke eines Fachgebiets [Hell97, S. 43].

Unter dieser Prämisse bedarf eine vollständige Beschreibung der zur Bezeichnung von Schemaelementen eingesetzten Begriffe

- der Spezifikation der Merkmale eines Begriffs (*Intension nach innen*) sowie
- der Angabe von Beziehungen zu anderen Begriffen (*Intension nach außen*).

Durch die Angabe von Begriffsbeziehungen wird parallel ein fachgemäßes Begriffssystem aufgebaut. Um dessen Konsistenz zu wahren, ist nachhaltige Terminologierarbeit unerlässlich. Die Auswirkungen nicht aufeinander abgestimmter Begriffe werden im anschließenden Abschnitt erörtert.

5.2.2 Sprachliche Defekte

Nicht konsolidierte oder nicht allgemein akzeptierte Begriffsdefinitionen führen zu einem unscharfen oder widersprüchlichen Verständnis innerhalb einer Domäne (z. B. infolge lückenhafter Begriffssysteme oder durch fehlendes Terminologiemanagement). Konflikte bei der Benennung oder Verwendung von Begriffen werden allgemein als *sprachliche Defekte* bezeichnet. Häufig auftretende Ausprägungen sprachlicher Defekte sowie mögliche Strategien zu deren Auflösung werden nachfolgend erörtert [Ortn97, S. 32], [Lehm01, S. 21], [LeNa07a, S. 75]:

- *Synonyme* sind Wörter, die dieselbe Bedeutung (Extension und Intension) besitzen und gegeneinander ausgetauscht werden können. Sofern eine kontrollierte Existenz von Synonymen vorliegt, muss diese nicht unbedingt aufgelöst werden. Beispiel: Kunde und Auftraggeber.
- Gleich geschriebene und gesprochene Wörter, jedoch mit jeweils anderer Bedeutung (Extension und Intension), werden als *Homonyme* bezeichnet. Homonyme sind aufzulösen und auf Bezeichnungsebene (Begriffswort) zu differenzieren. Beispiel: Auftragsbestand aus Vertriebsicht (Kundenauftragsbestand) vs. Auftragsbestand aus Produktionssicht (Fertigungsauftragsbestand).
- *Äquipollenzen* liegen vor, wenn dieselben Objekte (Extension) unter verschiedenen Blickwinkeln (Intension) betrachtet und unterschiedlich bezeichnet werden. Für ein konsolidiertes Domänenverständnis ist es eine zwingende Voraussetzung, Begriffe dieses Typs zu identifizieren und mit den Fachanwendern abzustimmen. Beispiel: Lagerbestand mengenmäßig vs. wertmäßig.

- Von *Vagheiten* wird gesprochen, wenn eine klare, inhaltliche Abgrenzung (Intension) der Begriffe fehlt. Hinsichtlich der Extension treten dadurch Unklarheiten und Unsicherheiten auf. Eine Präzisierung des Begriffsinhalts führt zu einer eindeutigen Auslegung des Begriffsumfangs. Beispiel: Sind nicht bestätigte Kundenaufträge dem Kundenauftragsbestand zuzurechnen?
- *Falsche Bezeichner* betiteln Abweichungen zwischen der tatsächlichen und der suggerierten Wortbedeutung (Extension und Intension) eines Begriffs. Falsche Bezeichner führen zu Verständigungsschwierigkeiten zwischen den Anwendern. Eine Ersetzung des ggf. geläufigen, aber nicht adäquaten Begriffsworts ist unerlässlich.

Die Behandlung sprachlicher Defekte stellt eine kritische Aufgabe und zugleich notwendige Voraussetzung für die Überwindung semantischer Heterogenität im Allgemeinen dar. Die Auflösung und Vermeidung sprachlicher Defekte, insbesondere im Umfeld von DWH-Systemen, werden im nächsten Abschnitt erörtert.

5.2.3 Behandlung sprachlicher Defekte bei Data-Warehouse-Projekten

In einem DWH-System werden Daten aus unterschiedlichen Datenquellen extrahiert, konsolidiert und themenorientiert gespeichert (vgl. Abschnitt 2.3). Um die mit einem DWH-System verbundenen Nutzenpotentiale realisieren zu können (vgl. Abschnitt 2.3.4), muss ein Konsens über das Verständnis dieser Daten erreicht werden. Dabei sind sprachliche Defekte aufzulösen und bestehende Begriffssysteme zu homogenisieren. STRAUCH schlägt hierfür drei Ansätze im Kontext von DWH-Projekten vor [Stra02, S. 197 ff.].

- Die *proaktive Homogenisierung* strebt die Entwicklung einer Begriffswelt, unabhängig von einem spezifischen (DWH-)Projekt, an. Hierzu werden auftretende Begriffe in einer Datenbank zentral erfasst und regelmäßig von einem Mitarbeiterteam mit dem bestehenden Begriffssystem konsolidiert. Um eine fortwährende Nachvollziehbarkeit sicherzustellen, dürfen einmal überprüfte Begriffe nicht mehr verändert werden.

- Die *reaktive Homogenisierung* wird nach der Analyse der Informationsbedarfe oder bei bereits bestehenden Begriffssystemen durchgeführt. Ziel ist die Definition eines einheitlichen Verständnisses über den gesamten Informationsbedarf. Dies kann bspw. mithilfe der im nachstehenden Abschnitt erörterten *Methode zur Rekonstruktion von Fachbegriffen* erfolgen. Nach der Sammlung von Aussagen werden bestehende sprachliche Defekte bereinigt und die Aussagen in Bezug auf Objekte semantischer Datenmodelle⁵¹ klassifiziert.
- Ein *Glossar* legt mit sachlichen Definitionen das korrekte Verständnis und den Gebrauch von (Fach-)Begriffen im Voraus verbindlich fest. Es ist jedoch zu konstatieren, dass bestehende Glossare kaum genutzt, gepflegt oder weiterentwickelt werden.

Die vorgestellten Ansätze können insbesondere bei der Neugestaltung eines DWH-Systems unterstützend eingesetzt werden. Idealtypisch ist bereits im Vorfeld der Implementierung eines DWH-Systems ein Konsens über die verwendeten (Fach-)Begriffe zu erzielen. Andernfalls ist ein inhomogenes Verständnis der späteren Berichtsergebnisse vorprogrammiert. Insbesondere die beiden antizipativ ausgelegten Ansätze, die proaktive Homogenisierung und das Glossar, sollten hier Beachtung finden. Für die Vereinheitlichung bestehender Berichtssysteme ist die reaktive Homogenisierung heranzuziehen.

Die reaktive Homogenisierung ist auch im Rahmen dieser Arbeit einsetzbar. Zur Realisierung der Überwindung semantischer Heterogenität bei multiplen DWH-Systemen ist eine exakte Klärung der Semantik eines Begriffs eine notwendige Voraussetzung. Aus diesem Grund wird zur Unterstützung der reaktiven Homogenisierung nachfolgend die Methode zur *Rekonstruktion von Fachbegriffen* eingeführt.

⁵¹ Der Begriff *Datenmodell* bezeichnet ein Metamodell zur Datenmodellierung [FeSi06, S.138].

5.2.4 Methode zur Rekonstruktion von Fachbegriffen

Ein fundiertes Berichtswesen erfordert ein exakt definiertes, semantisch und logisch homogenes Begriffsverständnis (vgl. Abschnitt 5.1 bis 5.2.3).

Only if data is linked to clear defined business terms it obtains a meaning (semantic) for the end user – the data turn to information [LeJa99].

Insbesondere zur Behandlung sprachlicher Defekte soll im Folgenden die *Methode zur Rekonstruktion von Fachbegriffen* vorgestellt werden. Zu diesem Zweck wird vorab der *sprachkritische Ansatz* von ORTNER skizziert.

SPRACHKRITISCHER ANSATZ

Beim Gebrauch der natürlichen Sprache werden fortwährend Sachverhalte und Anweisungen wiedergegeben oder Behauptungen aufgestellt. Durch eine kritische Analyse des Spracheinsatzes, die sowohl den Inhalt der Äußerungen als auch deren Struktur systematisch rekapituliert, können die – den Äußerungen zugrunde liegenden – Begrifflichkeiten rekonstruiert werden. Bei einem Rekonstruktionsprozess werden die sprachlichen Defekte (vgl. Abschnitt 5.2.2) an und zwischen den Begriffen behoben. Auf dieser Basis können Sachverhalte und die sich darauf stützenden Fachbegriffe konsistent beschrieben werden [Ortn97, S. 15 ff., S. 70 ff.].

METHODE ZUR REKONSTRUKTION VON FACHBEGRIFFEN

In Anlehnung an den sprachkritischen Ansatz von ORTNER sowie die Ausführungen in [LeEI97, S. 176 ff.] und [LeJa99] wird die Methode zur *Rekonstruktion von Fachbegriffen* eingeführt. Diese entstammt dem Bereich der Software-Entwicklung und konzentriert sich auf die Kommunikationslücken zwischen (Fach-)Anwendern, die sich gewöhnlich in einer gewachsenen, natürlichen Sprache ausdrücken, und Entwicklern, die eine künstliche, konstruierte Sprache bevorzugen. Im Aufgabenfeld der Datenmodellierung, die hier als Teil-Disziplin der Software-Entwicklung aufgefasst wird, kann diese Methode u. a. einen schrittweisen Aufbau eines (organisationsweit) konsistenten Datenschemas begleiten. In der Literatur existiert darüber hinaus eine Erweiterung speziell für die Modellierung von OLAP-Anwendungen [LeEI97, S. 179 f.]. Abbildung 5-4 gibt einen Überblick über die Methode zur Rekonstruktion von Fachbegriffen.

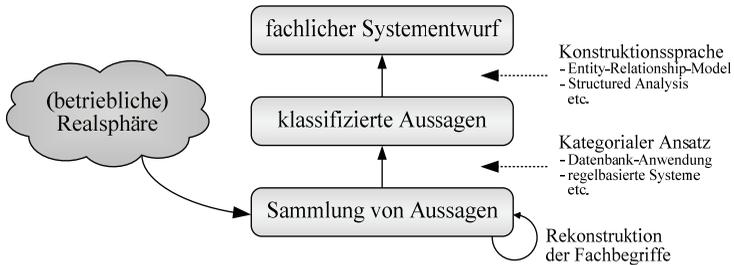


Abbildung 5-4: Methode zur Rekonstruktion von Fachbegriffen [LeE197, S. 177]

Die in der Realsphäre etablierten Begrifflichkeiten sind in einem ersten Schritt zu erfassen (*Sammlung von Aussagen*) und auf folgende Fragestellungen hin zu untersuchen:

- Was bedeutet der Fachbegriff und welche Objekte umfasst er? (Extension)
- Welche Eigenschaften besitzt der Begriff? (Intension)
- Eignet sich der Begriff für den vorgesehenen Kontext?

Dieser fachlichen Diskussion schließt sich die Identifikation und Bereinigung sprachlicher Defekte an (*Rekonstruktion der Fachbegriffe*). Die konsekutiv durchzuführende *Klassifikation* der überarbeiteten Fachbegriffe bildet die Grundlage für eine quasi natürlichsprachliche Dokumentation der Zusammenhänge. Die klassifizierten Aussagen können abschließend gemäß einer Konstruktionssprache in ein semantisches Modell überführt werden, das in den *fachlichen Systementwurf* mit eingeht.

Übertragen auf die Konstruktion von Datenbank-Anwendungen, kann die Sammlung von Aussagen nach Aussagen über Informationsobjekte und deren Eigenschaften, Operationen auf den Informationsobjekten, Ereignissen und Integritätsbedingungen unterteilt werden. Die klassifizierten Aussagen können anschließend auf Objekttypen und Beziehungen bspw. eines *Entity-Relationship-Modells* (ERM)⁵² oder eines *Strukturierten Entity-Relationship-Modells* (SERM)⁵³ abgebildet werden [LeE197, S. 178].

⁵² Das *Entity-Relationship-Modell* (ERM) wurde in seiner Grundform 1976 von CHEN eingeführt und hat sich als Standardmodell unter den semantischen Datenmodellen durchgesetzt. Detaillierte Informationen über das ERM sind z. B. in [Chen76] und [FeSi06, S. 132 ff.] nachzulesen.

⁵³ Das *Strukturierte Entity-Relationship-Modell* (SERM) von SINZ beruht auf dem Grundmodell des ERM. Ziel des SERM ist es, das Konzept des ERM mit dem Konzept quasi-hierarchischer Graphen zu verbinden und bekannte Schwachstellen des ERM (wie z. B. die Visualisierung von Existenzabhängigkeiten) zu beseitigen [Sinz88], [FeSi06, S. 146 ff.].

ORGANISATORISCHE EINBETTUNG DER BEGRIFFSREKONSTRUKTION

Um die fundierte und konsistente Definition von Fachbegriffen nachhaltig zu gewährleisten, ist eine adäquate Vorgehensweise bei der Begriffsbestimmung zu standardisieren und organisatorisch zu verankern. Gemäß dem Vorschlag von LEHMANN und ELLERAU werden im Hinblick darauf die in Abbildung 5-5 dargestellten Funktionen eingeführt und anhand der jeweiligen Zuständigkeiten erläutert [LeE197, S. 180 ff.], [Lehm01, S. 136 ff.].

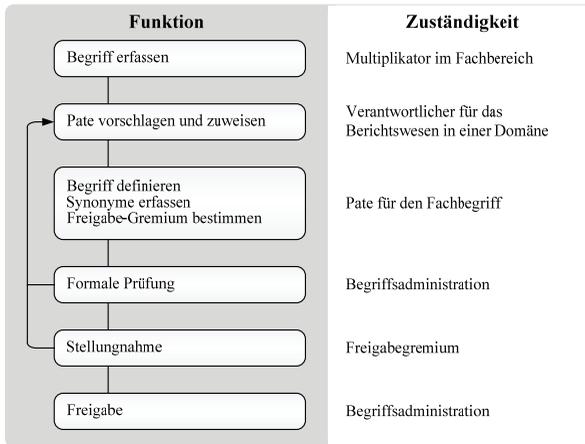


Abbildung 5-5: Organisation der Begriffsdefinition [LeE197, S. 181]

- *Multiplikator*: Innerhalb eines Fachbereichs nehmen ein oder mehrere Mitarbeiter (so genannte Prozess-Multiplikatoren) die Informationsbedarfe der Anwender auf und konsolidieren deren Anforderungen. Sie initiieren bei Bedarf die Definition neuer Begriffe.
- *Verantwortlicher für das Berichtswesen in einer Domäne*: Pro Hauptanwendung wird ein Verantwortlicher bestimmt, der die notwendigen Aktivitäten für das Berichtswesen koordiniert und einen Paten benennt, der die Definition des Fachbegriffs durchführt.
- *Pate*: Der so genannte Pate ist verantwortlich für die inhaltliche Bedeutung des Begriffs und fungiert gleichzeitig als Ansprechpartner bei entsprechenden fachlichen Fragen. Er teilt das Freigabegremium ein.

- *Begriffsadministration*: Die formale Prüfung des Fachbegriffs, vor allem im Hinblick auf Querverweise zu anderen Begriffen, obliegt der Begriffsadministration.
- *Freigabegremium*: Für jeden Fachbegriff wird ein Freigabegremium bestimmt. Nach der kritischen Prüfung des vom Paten definierten Begriffs wird dieser als global verbindlich vorgeschrieben.

Die vorgestellte organisatorische Maßnahme kann zur Etablierung eines einheitlichen Begriffssystems beitragen, da mittels dieser die Einführung und Verwaltung von Fachbegriffen global (z. B. organisationsweit) koordiniert wird. Insbesondere die Fachbereiche müssen die Notwendigkeit und Vorteile erkennen, die es mit sich bringt, einen organisationsweiten Konsens über bestehende und neue Begrifflichkeiten zu erreichen und daher aktiv an der Gestaltung mitwirken.

FACHVERANTWORTLICHE BEI DWH-SYSTEMEN

Aufgrund der personellen Ressourcenknappheit in betrieblichen Organisationen ist es kaum möglich, jede der genannten Zuständigkeiten direkt einer Person zuzuweisen. Im Bereich von DWH-Systemen hat es sich etabliert, die Rolle des Fachverantwortlichen einzuführen, der wesentliche Funktionen der Rekonstruktion von Fachbegriffen auf sich vereint.

Der Fachverantwortliche stellt den primären Wissensträger für die Datensemantik dar. Er bildet aus Systemsicht die Schnittstelle zu den Endanwendern und nimmt deren fachliche und technische Anforderungen zur Weiterentwicklung des DWH-Systems entgegen. Ihm obliegt in diesem Zusammenhang die semantische Klärung der – ggf. neu einzuführenden – Fachbegriffe. Zudem stimmt er die genauen Berichtsanforderungen mit den Nutzern ab und stellt ihnen die gewünschten Berichte zur Verfügung. Der Fachverantwortliche hat somit die Verantwortung über das Begriffs- und Berichtssystem eines DWH-Systems inne, d. h. er stellt neben dem einheitlichen Begriffsverständnis auch die semantisch korrekte Definition der Berichte sicher [SUPH07b].

5.3 Heterogenität in Datenschemata

Der Duden umschreibt *Heterogenität* als Verschiedenartigkeit, Ungleichartigkeit, Uneinheitlichkeit im Aufbau oder in der Zusammensetzung [Dude05, S. 402]. Im Hinblick auf die Erreichung des gegebenen Untersuchungsziels werden nachstehend Arten und Ausprägungen von Heterogenität vor allem auf Datenschemaebene untersucht. Dabei kommt der semantischen Heterogenität besondere Beachtung zu. Anschließend werden mögliche Gründe für Datenschema-Heterogenität erörtert und Kriterien herausgestellt, anhand derer bestimmt werden kann, ob eine spezifische Heterogenitätsart überwunden ist. Die Betrachtung eines semantischen Ähnlichkeitsmaßes rundet diesen Abschnitt ab.

5.3.1 Arten von Heterogenität

Der vorliegende Abschnitt widmet sich – mit Fokus auf die Datenschemaebene – der Erörterung grundsätzlicher Arten von Heterogenität. Folgende Heterogenitätsarten sind zu unterscheiden [LeNa07a, S. 60 ff.], [Schm98, S. 13 f.]:

- Unter *technischer Heterogenität* werden alle Fragestellungen der technischen Realisierung des Zugriffs auf die Daten der Datenquellen subsumiert (z. B. verfügbare Anfragesprachen und Funktionen, Kommunikationsprotokoll, Formate für den Datenaustausch).
- *Syntaktische Heterogenität* umfasst Gegebenheiten, die infolge unterschiedlicher Darstellung identischer Sachverhalte entstehen. Hierzu zählen bspw. abweichende Zeichenkodierung (z. B. Unicode, ASCII) oder ungleiche Trennzeichen bei flachen Dateien.
- Mit *Datenmodellheterogenität* werden Unterschiede in den zur Präsentation oder Verwaltung der Daten genutzten Datenmodellen (z. B. relational, objekt-relational, objekt-orientiert) bezeichnet. Heterogenität auf Datenmodellebene bewirkt nahezu immer auch semantische Heterogenität, da Schemaelemente in unterschiedlichen Datenmodellen jeweils die spezifische, durch das Datenmodell determinierte Semantik besitzen.

- *Strukturelle Heterogenität* bezeichnet Verschiedenheit in der strukturellen Repräsentation von Daten, d. h. zwei Schemata eines bestimmten Datenmodells weichen voneinander ab, obwohl sie den gleichen Realwelt-Ausschnitt beschreiben. Es liegt folglich ein Überschneidungsbereich bei der Intension der Schemaelemente vor.
- *Schematische Heterogenität* stellt einen Spezialfall der strukturellen Heterogenität dar. Dieser Heterogenitätsklasse sind Schemata zuzuordnen, bei denen unterschiedliche Elemente eines Datenmodells verwendet wurden, um denselben Sachverhalt zu modellieren.
- *Semantische Heterogenität* betrifft Differenzen bei der Interpretation von Begriffsworten (z. B. Bezeichner von Schemaelementen) und deren Bedeutung.

In der Praxis treten die vorgestellten Heterogenitätsarten zumeist in komplexen Kombinationen auf. Eine eindeutige Trennung der Heterogenitätsarten ist daher selten möglich. Zudem bedingen sich die Heterogenitätsarten gegenseitig.

Abbildung 5-6 ordnet die Heterogenitätsarten aufsteigend nach ihrem semantischen Anteil. Mit diesem nimmt ebenfalls die Konformität mit dem Untersuchungsschwerpunkt dieser Arbeit zu. Datenmodellheterogenität und strukturelle Heterogenität werden auch als *Modellierungsheterogenität* zusammengefasst.

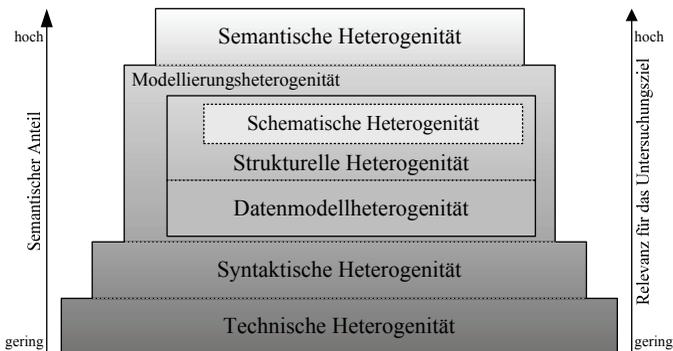


Abbildung 5-6: Heterogenitätsarten

Strukturelle und semantische Heterogenität bereiten in der Realwelt die größten Schwierigkeiten (z. B. bei Konsolidierungs- oder Integrationsvorhaben im Bereich relationaler Datenbanken). Außerdem sind die beiden Heterogenitätsarten eng miteinander verwoben. Um z. B. die semantische Gleichwertigkeit zweier Schemata beurteilen zu können, sind neben der Betrachtung der eigentlichen Schemata auch deren Integritätsbeziehungen zu analysieren [ShKa93, S. 284], [LeNa07a, S. 69]. Bei der sich anschließenden Betrachtung semantischer Heterogenität werden daher auch strukturelle Aspekte mit einbezogen.

5.3.2 Semantische Heterogenität

Die Bezeichnung von Schemaelementen erfolgt mittels Begriffen bzw. genauer mit Begriffsworten. Ein Schemaelement ist somit grundsätzlich gemäß dem vorgestellten Begriffsmodell (vgl. Abschnitt 5.2.1) in Bezug auf seine Intension und Extension zu beschreiben. Werden Schemaelemente unterschiedlich interpretiert oder verwendet, d. h. ihre Intension oder Extension weichen voneinander ab, besteht Heterogenität auf semantischer Ebene. Sprachliche Defekte sind dabei ein deutliches Indiz für semantische Heterogenität und entsprechend zu behandeln (vgl. Abschnitt 5.2.2 bis 5.2.4).

Die Aufdeckung sprachlicher Defekte und semantischer Konflikte bedarf auch der Einbeziehung von Informationen über den zugrunde liegenden Kontext (vgl. auch Abschnitt 4.1.1). Kontextinformationen zu Schemaelementen sind jedoch im Allgemeinen nicht direkt zugreifbar oder verfügbar, da in gebräuchlichen Datenmodellen keine Angaben über den Kontext vorgesehen sind. Ein Rückschluss auf die Intension eines Schemaelements z. B. anhand seiner Extension, ist somit nur indirekt möglich. Ohne weiterführende Angaben z. B. durch Attribute, Dokumentationen oder implizites Wissen kann die originäre Intension eines Schemaelements nur vage ermittelt werden. Darüber hinaus können an sich identische Schemata, abhängig von ihrem Anwendungskontext, unterschiedliche Semantik besitzen [LeNa07a, S. 73 f.].

An dieser Stelle knüpft der zu erstellende Lösungsansatz dieser Arbeit an. Mithilfe von Metadaten soll die Intension eines Begriffs, hier im Speziellen in Form eines Schemaelements, semantisch reichhaltig beschrieben werden, sodass dessen Extension hinreichend bestimmt ist. Dazu zählt auch die Einbeziehung von Kontextinformationen gemäß der in Abschnitt 4.3 gegebenen Definition.

5.3.3 Ursachen für Heterogenität

In den vorausgegangenen Kapiteln wurden Gründe für das Auftreten von Heterogenität aus organisationaler Sicht (z. B. aufgrund der vorliegenden Organisationsstruktur oder der spezifischen Aufgaben und Anforderungen von Entscheidungsträgern) sowie aus Sicht der Metadaten-Verwaltung (z. B. infolge einer vorherrschenden dezentralen Metadaten-Architektur) aufgezeigt. Des Weiteren kann Heterogenität auf Datenschemaebene durch die Beteiligung unterschiedlicher Personen bei der Schemaentwicklung entstehen. Insbesondere strukturelle Heterogenität kann durch die Präferenzen eines Modellierers oder dessen individuelle Wahrnehmung der Umweltrepräsentation begründet sein. Die folgende Aufstellung kategorisiert hierzu wesentliche Konfliktebenen [RaBe01], [SpPD92, S. 89 ff.]:

- *Semantischer Konflikt*: Die jeweiligen Modellierer nehmen eine ungleiche Menge an Realwelt-Objekten wahr (z. B. aufgrund einer ungenauen Abgrenzung des Realwelt-Ausschnitts). Infolgedessen entstehen gemäß den spezifischen Anforderungen individuelle Darstellungen überlappender Realwelt-Ausschnitte.
- *Deskriptiver Konflikt*: Bei der Beschreibung sich entsprechender Realwelt-Objekte werden von den Modellierern nicht dieselben Eigenschaften zur jeweiligen Charakterisierung erfasst. In diese Kategorie fallen auch Benennungskonflikte aufgrund von Synonymen und Homonymen (vgl. Abschnitt 5.2.2) sowie die Definition von Attribut-Wertebereichen oder von Integritätsbedingungen.
- *Heterogenitätskonflikt*: Die Anwendung verschiedenartiger Datenmodelle zählt zu dieser Konfliktklasse.
- *Struktureller Konflikt*: Bei der Nutzung desselben Datenmodells bestehen hohe Freiheitsgrade in der Abbildung und Beschreibung der Realwelt-Objekte. Ein Modellierer kann identische Gegebenheiten mithilfe unterschiedlicher Konstrukte darstellen. Des Weiteren sind spezielle Konzepte bei der Modellierung z. B. die bewusste Denormalisierung zur Anfrageoptimierung an dieser Stelle zu nennen.

Weitere Ursachen für das Entstehen von Heterogenität sind unterschiedliche zeitliche Entwicklungen oder verschiedenartige technische Anforderungen und Rahmenbedingungen [LeNa07a, S. 59, S. 67 ff.]. Die geschilderten Konfliktarten sind zueinander orthogonal und können zudem kumulativ auftreten. Differenzen in realen Schemata weisen meist einen Mix aus diesen Konflikttypen auf.

5.3.4 Überwindung von Heterogenität

Bevor ein Lösungsvorschlag zur Überwindung einer spezifischen Ausprägung von Heterogenität erarbeitet werden kann, bedarf es maßgebender Anhaltspunkte, anhand derer die Zielerreichung überprüft werden kann [LeNa07a, S. 61]:

- *Technische Heterogenität* gilt als überwunden, wenn von einem beliebigen System (z. B. einem DBMS) eine Anfrage an eine Datenquelle gesendet werden kann, diese die Anfrage versteht und eine Menge von Daten als Ergebnis bereitstellt. Dabei ist es unerheblich, ob die bei der Anfrage benutzten Schemaelemente in beiden Welten dieselbe Bedeutung besitzen.
- *Syntaktische Heterogenität* ist überbrückt, wenn alle Informationen, denen dieselbe Bedeutung anheim gegeben ist, gleich dargestellt werden.
- Nutzen zwei Systeme oder Datenquellen das gleiche Datenmodell, kann das Problem der *Datenmodellheterogenität* als gelöst angesehen werden.
- Die Überwindung der *strukturellen Heterogenität* ist erreicht, wenn semantisch identische Konzepte strukturell gleich modelliert sind.
- Wird unter den verwendeten Begriffsworten für Schemaelemente auch jeweils dasselbe verstanden, d. h. den Schemaelementen liegt jeweils die identische Bedeutung zugrunde, ist die *semantische Heterogenität* überwunden.

Das Ziel der vorliegenden Arbeit besteht nicht in der Auflösung der semantischen Heterogenität bei multiplen DWH-Systemen. Insbesondere aus organisationalem Blickwinkel darf unter bestimmten Umständen eine Auflösung der Heterogenität nicht realisiert werden (z. B. Wahrung der Autonomie einer Sparte, Einhaltung (datenschutz-)rechtlicher Vorgaben). Im Hinblick auf ein einheitliches Lenkungssystem wird daher explizit die Überwindung der semantischen Heterogenität unter Berücksichtigung eventuell bestehender Rahmenbedingungen angestrebt.

5.3.5 Semantische Ähnlichkeitsmaße

Die Erreichung des gegebenen Untersuchungsziels kann durch die Nutzung eines semantischen Ähnlichkeitsmaßes gestützt werden. Semantische Ähnlichkeitsmaße werden zur Quantifizierung des Grades der Ähnlichkeit zwischen zwei Objekten eingesetzt. Zu dessen Bestimmung wird an dieser Stelle auf einen Ansatz aus dem Bereich der Multidatenbanksysteme⁵⁴ zurückgegriffen, da eine Übertragung dieses Vorschlags auf multiple DWH-Systeme als gut realisierbar erscheint.

SEMANTISCHE ÄHNLICHKEITSRELATION

Die semantische Ähnlichkeit (*semantic proximity* (semPro)) zwischen zwei gegebenen Objekten O_1 und O_2 definieren SHETH und KASHYAP durch folgendes 4-Tupel [ShKa93]:

$$\text{semPro}(O_1, O_2) = \langle \text{KONTEXT}, \text{ABSTRAKTION}, (D_1, D_2), (S_1, S_2) \rangle$$

Dabei ist D_i die Domäne eines Objekts O_i und S_i dessen Zustand. Die Bestandteile dieses Ausdrucks bedeuten im Einzelnen:

- *Kontext*: Jedes Objekt besitzt seinen eigenen Kontext. Der Begriff Kontext in semPro bezieht sich auf den Kontext, in dem eine bestimmte semantische Ähnlichkeit zwischen den beiden gegebenen Objekten besteht.
- *Abstraktionsgrad*: Der Mechanismus, um die Domänen der Objekte aufeinander abzubilden, wird als Abstraktionsgrad angegeben. Für eine Erläuterung möglicher Abbildungsvarianten wie z. B. vollständiges 1:1-Werte-Mapping, partielles N:1-Mapping, Generalisierung, Aggregation, funktionale Abhängigkeit, beliebige (ANY), nicht definierte (NONE) oder nicht mögliche (NEG) Beziehung wird auf [ShKa93, S. 287 f.] verwiesen.
- *Domäne D_i* : Die Domäne gibt die Menge an Werten vor, die ein Objekt annehmen kann.
- *Zustand S_i* : Der Zustand eines Objekts kann als dessen Ausprägung, die bspw. in einer Datenbank gespeichert ist, interpretiert werden.

⁵⁴ Ein *Multidatenbanksystem* ist aus unabhängigen, autonomen lokalen Datenbanksystemen zusammengesetzt, die föderativ verbunden werden können [Voss08, S. 452].

Da sich die vorliegende Arbeit ausschließlich auf die Schemaebene eines DWH-Systems konzentriert, werden die Ausprägungen eines Objekts, d. h. seine Zustände S_i , bei der nachstehenden Differenzierung nicht miteinbezogen.

TAXONOMIE SEMANTISCHER ÄHNLICHKEITSGRADE

Die nachstehende Taxonomie definiert anhand obiger Relation unterschiedliche Grade semantischer Ähnlichkeit.

- *Semantische Äquivalenz (equivalence)*: Zwei Objekte sind semantisch äquivalent, wenn sie die gleiche Realwelt-Entität beschreiben. Für jeden Kontext sollte ein vollständiges 1:1-Werte-Mapping zwischen den Domänen dieser Objekte möglich sein (semantische Domänen-Äquivalenz).

$$\text{semPro}(O_1, O_2) = \langle \text{ALLE}, M, (D_1, D_2), -- \rangle$$

M = vollständiges 1:1-Werte-Mapping

- *Semantische Verwandtschaft (relationship)* liegt vor, wenn folgende Möglichkeiten der Abbildung der Domänen zweier Objekte durchführbar sind: partielles N:1-Mapping, Generalisierung, Aggregation.

$$\text{semPro}(O_1, O_2) = \langle \text{ALLE}, M, (D_1, D_2), -- \rangle$$

M = partielles N:1-Werte-Mapping, Generalisierung oder Aggregation.

- *Semantische Beziehung (relevance)* ist gegeben, wenn zwei Objekte mithilfe eines beliebigen Abbildungsmechanismus im selben Kontext zueinander in Beziehung gesetzt werden können.

$$\text{semPro}(O_1, O_2) = \langle \text{GLEICH}, \text{BELIEBIG}, (D_1, D_2), -- \rangle$$

- *Semantische Ähnlichkeit (resemblance)* legt eine zusätzliche Differenzierung nach der Rolle, die ein Objekt im gegebenen Kontext inne hat, zugrunde.

Rolle : Objekt \times Kontext \rightarrow Rollenname.

Zwei Objekte können nicht über einen Mechanismus aufeinander abgebildet werden, aber sie besitzen dieselbe Rolle in ihrem jeweiligen Kontext.

$$\begin{aligned} \text{semPro}(O_1, O_2) = & \langle \text{KONTEXT}, \text{KEINE}, (D_1, D_2), -- \rangle \\ \text{WHERE } & \text{Kontext} = \text{Kontext}(O_1) \vee \text{Kontext}(O_2) \\ \text{AND } & D_1 \neq D_2 \\ \text{AND } & \text{role-of}(O_1, \text{Kontext}) = \text{role-of}(O_2, \text{Kontext}) \end{aligned}$$

- *Semantische Unvereinbarkeit (incompatibility)*: Es existiert weder ein Kontext noch eine Abstraktion in der die Domänen zweier Objekte in Beziehung gesetzt werden können. Zudem können die Objekte keine ähnlichen Rollen, in dem Kontext, in dem sie bestehen, besitzen.

```

semPro(O1, O2) = < KEIN, NICHT MÖGLICH, (D1, D2), -- >
WHERE Kontext = Kontext(O1) v Kontext(O2)
           ist nicht definiert
AND D1 vielleicht oder vielleicht auch nicht gleich D2
AND role-of(O1, Kontext) und role-of(O2, Kontext)
           sind nicht vergleichbar.

```

Eine weitere Detaillierung dieser Taxonomie ist [ShKa93] zu entnehmen. Die skizzierten Ähnlichkeitsgrade sollen im Folgenden der Klassifizierung von Beziehungen zwischen Schemaelementen dienen. Sie bilden somit einen Grundstein für die Überwindung semantischer Heterogenität bei multiplen DWH-Systemen.

5.4 Schema-Management

Unter dem Begriff *Schema-Management* werden Techniken zur Erkennung und Überwindung struktureller Heterogenität auf Schemaebene subsumiert. Strukturelle Heterogenität liegt vor, wenn zwei Schemata differieren, obwohl sie den gleichen Realwelt-Ausschnitt beschreiben (vgl. Abschnitt 5.3). Schema-Management zielt darauf ab, Daten zusammenzuführen, die aufgrund von Heterogenität auf Schemaebene inhomogen sind. Die hierzu notwendige Datentransformation erfolgt anhand von Korrespondenzen zwischen Schemaelementen.

Die Überführung von Schemata in unterschiedliche Datenmodelle sowie die Schemaintegration im Speziellen waren Gegenstand intensiver Forschungen Mitte der 1990er Jahre und erfahren gegenwärtig unter dem Namen *Model-Management* wieder verstärkt Aufmerksamkeit. Model-Management bezieht dabei zusätzlich Anwendungsgebiete wie z. B. Webseiten, Ontologien oder Web-Service-Schnittstellen in den Forschungsbereich mit ein. Die Vision des Model-Managements besteht in der Bereitstellung eines Frameworks zur generischen Metadatenverwaltung und einer daraus resultierenden vereinfachten Erstellung und Anpassung metadatengetriebener Anwendungen [BeMe07], [Rahm07], [LeNa07a, S. 81, S. 115].

Einen der ersten Prototypen eines generischen Model-Management-Systems haben MELNIK, RAHM und BERNSTEIN mit *Rondo* vorgestellt. Dieser unterstützt jedoch nur einfache (syntaktische) Mappings [MeRB03], [QuKC05, S. 7], [Rahm07].

TECHNIKEN DES SCHEMA-MANAGEMENTS

Die folgende Abbildung gibt einen Überblick über die zentralen Techniken des Schema- bzw. Modell-Managements sowie deren Verknüpfungen.

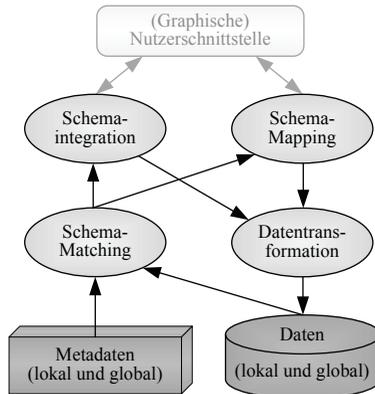


Abbildung 5-7: Aufgaben des Schema-Managements [LeNa07a, S. 115]

Ein elementares Konzept des Schema-Managements ist das *Mapping*. Basierend auf den *Metadaten* der zu verbindenden Datenquellen, beschreiben Mappings eine Menge von meist binären Beziehungen zwischen Elementen zweier Schemata. Dabei kann das Mapping entweder manuell erstellt oder mithilfe von *Schema-Matching*-Techniken (semi-)automatisch ermittelt werden. Die Mapping-Informationen können bei der *Schemaintegration* und beim *Schema-Mapping* Einsatz finden, wo sie zur Ableitung von Regeln für die *Datentransformation* dienen [LeNa07a, S. 115 ff., S. 165 ff.], [BeMe07].

5.4.1 Schema-Matching

Verfahren zur (semi-)automatischen Erkennung semantisch äquivalenter Elemente aus unterschiedlichen Schemata werden als *Schema-Matching* bezeichnet. Abhängig vom gewählten Schema-Matching-Ansatz werden dabei die Schemata, ihre Struktur und Integritätsbedingungen sowie ggf. Beispieldaten analysiert. Anhand dieser Ergebnisse werden Vorschläge für Korrespondenzen zwischen den Schemata generiert. Ein Domänenexperte verifiziert abschließend die Vorschläge und entscheidet über deren Annahme oder Ablehnung.

Aufgrund nicht formal spezifizierter oder fehlender semantischer Angaben über reale Schemata ist eine vollständig automatisierte Generierung von Schema-Korrespondenzen gegenwärtig nicht realisierbar. Die Unterstützung durch Schema-Matching beschränkt sich daher für den Anwender auf die Identifizierung geeigneter Kandidaten [BeMe07], [LeNa07a, S. 143], [Holt03, S. 43].

KLASSIFIKATION VON SCHEMA-MATCHING-ANSÄTZEN

Abbildung 5-8 offeriert eine Klassifikation gängiger Schema-Matching-Ansätze.

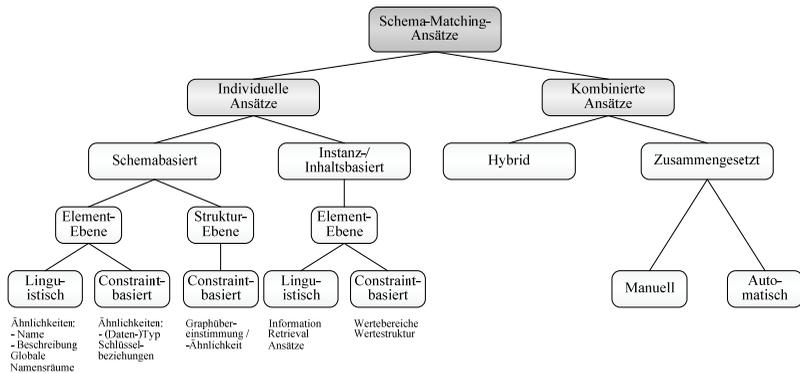


Abbildung 5-8: Klassifikation von Schema-Matching-Ansätzen [RaBe01, S. 338]

Auf oberster Ebene wird zwischen *individuellen* und *kombinierten* Ansätzen unterschieden; letztere sind weiter in *hybride* und *zusammengesetzte* (manuell oder automatisch) zu unterteilen. Die individuellen Ansätze gliedern sich in *Schema-* und *Instanz-Matching*, d. h. es wird charakterisiert, ob die Matching-Operationen Schemainformationen verarbeiten oder (auch) Elemente der Ausprägungsebene berücksichtigen. Eine Matching-Operation kann elementweise, z. B. für einzelne Attribute oder für komplexe

Datenstrukturen definiert werden (*Element-* vs. *Struktur-Ebene*). Auf unterster Ebene der Klassifikation werden jeweils *sprach-* und *constraintbasiertes* Matching differenziert. Sprachbasierte (linguistische) Matching-Operationen basieren auf Namen oder textlichen Beschreibungen; werden dagegen Schlüsselbeziehungen, Datentyp-Informationen o. Ä. herangezogen, so wird von einem constraintbasierten Ansatz gesprochen. Beispiele für Ansätze des Schema-Matchings sind bei den jeweiligen Blattknoten der geschilderten Klassifikation genannt [RaBe01], [Holt03, S. 43 f.]. Eine detaillierte Erläuterung dieser Taxonomie sowie ausgewählter Schema-Matching-Ansätze findet sich u. a. in [RaBe01] und [LeNa07a, S. 145 ff.].

OPERATOREN DES SCHEMA-MATCHINGS

Zu den elementaren Operatoren im Rahmen des Schema-Matchings (z. B. [Rahm07], [BeMe07], [QuKC05, S. 7]) zählen

- der *Match*-Operator zur Generierung eines Mappings zwischen zwei Modellen,
- der *Compose*-Operator, der zwei aufeinanderfolgende Mappings in ein Mapping zusammenführt,
- der *Merge*-Operator zum Zusammenführen zweier Modelle anhand eines vorgegebenen Mappings,
- der *Diff*-Operator, um für ein gegebenes Modell und Mapping ein Teilmodell zu bestimmen, das nicht am Mapping partizipiert, sowie
- der *ModelGen*-Operator, über den ein Modell einer Sprache in ein Modell einer anderen Sprache überführt wird.

Der Einsatz dieser Operatoren wird bspw. in [BeRa00] anhand von Szenarios zum Laden eines DWH-Systems dargelegt. Beispiele für Implementierungen im Bereich des Schema-Matchings werden in [MaBR01, S. 50], [RaBe01, S. 344 f.] gegeben. Als ein prominenter Algorithmus ist dabei *Cupid* hervorzuheben [MaBR01], dem der Anspruch eines generischen Schema-Matchings zugrunde liegt. Anwendungsbeispiele für *Cupid* stammen jedoch vornehmlich aus dem Bereich relationaler und XML-Schemata.

5.4.2 Schemaintegration

Bei der *Schemaintegration* wird ein globales, integriertes Schema aufgebaut, das möglichst alle Konzepte der lokalen, heterogenen Schemata beinhaltet. Die Festlegung, welche Elemente der zu integrierenden Schemata zu einem gemeinsamen Element des globalen Schemas zusammengefasst werden können, stützt sich auf die ermittelten Mapping-Informationen. Der Schemaintegration liegen die Formalziele Vollständigkeit, Minimalität, Korrektheit und Verständlichkeit zugrunde.

Ein bekanntes Verfahren der Schemaintegration ist die *korrespondenzbasierte Schemaintegration* (Correspondance Assertion). Hierbei werden einander entsprechende oder miteinander in Beziehung stehende Bestandteile der zu integrierenden Schemata durch so genannte Zusicherungen beschrieben. Die Formulierung von Zusicherungen als Inter-Schema-Korrespondenzen geschieht dabei unabhängig von einem konkreten Datenmodell. Basierend auf diesen Zusicherungen erfolgt die schrittweise Konstruktion des integrierten Schemas mithilfe von Integrationsregeln [LeNa07a, S. 116 ff., S. 171], [Conr97, S. 78 f., S. 85 ff.], [SpPD92].

Weitere Beispiele für Integrationsansätze sind das *Generic Integration Model* (z. B. [Schm98, S. 45 ff.], [Conr97, S. 116 ff.]) und das *Upward Inheritance-Prinzip* (z. B. [Conr97, S. 99 ff.]). Fundierte Informationen und Techniken zur Schemaintegration finden sich z. B. in [BaLN86], [Schm98, S. 17 ff.] und [RaBe01].

Die Schemaintegration stellt nach wie vor ein komplexes Problem im Bereich des Datenmanagements dar. Bis dato existiert kein Lösungsansatz, der genügend Flexibilität und Ausdrucksstärke besitzt, um mit der Komplexität realer Schemata zufriedenstellend umgehen zu können [LeNa07a, S. 122].

5.4.3 Schema-Mapping

Schema-Mapping kann einerseits eine Menge von Korrespondenzen zwischen Attributen unterschiedlicher Schemata bezeichnen sowie andererseits einen Prozess, um – basierend auf Mapping-Informationen – Vorschriften zur Datentransformation von einem Schema in ein anderes (z. B. globales) Schema abzuleiten. Im Unterschied zur Schemaintegration, die den Aufbau eines neuen, integrierten Schemas verfolgt, wird beim Schema-Mapping die Datentransformation zwischen existierenden Schemata untersucht. Eine übersichtliche Einführung in das Forschungsfeld des Schema-Mappings findet sich z. B. in [LeNa07b], [LeNa07a, S. 123 ff., S. 171 f.].

Ein bekanntes Forschungsprojekt im Bereich des Schema-Mappings ist *Clio* [HeMH01]. Im Rahmen des Clio-Projekts wurde die Generierung ausführbarer Mappings erforscht. Clio unterstützt jedoch lediglich relationale und XML-Schemata [Rahm07], [QuKC05, S. 7]. Weitere Werkzeuge für das Schema-Mapping sind in [LeNa07b, S. 456 f.] skizziert.

5.4.4 Diskussion des Schema-Managements

Der Forschungsbereich Model-Management zielt auf die Konzeption und Realisierung einer generischen Metadatenverwaltung. Ein generischer Ansatz ist erforderlich, um Abbildungen zwischen Schemata unterschiedlicher Datenmodelle zu ermöglichen. Eine automatisierte Umsetzung der Mappings in Datentransformationsregeln erfordert dabei eine hohe semantische Ausdrucksstärke. Mapping-Sprachen auf Basis logischer Regeln oder algebraischer Ausdrücke gelten auf diesem Gebiet als sehr aussichtsreich. Jedoch belegen jüngste Forschungsansätze aufs Neue, dass die generische Realisierung der Schema-Matching-Operatoren sich umso schwieriger gestaltet, je mächtiger die Mapping-Sprache ist [Rahm07]. Dies begründet auch, warum sich Anwendungsbeispiele für die unterschiedlichen Prototypen vornehmlich aus dem wohl erforschten Bereich relationaler oder XML-Schemata finden. Die Vision des Model-Managements kann daher bis heute nur als partiell realisiert angesehen werden. Es besteht nach wie vor ein erheblicher Forschungsbedarf in diesem Umfeld, insbesondere beim Einsatz mächtigerer Datenmodelle (als bspw. das relationale und XML-Datenmodell) [Rahm07], [BeMe07].

Im Hinblick auf den Einsatz mächtigerer Datenmodelle im Forschungsbereich des Schema-Managements soll mit dieser Arbeit ein Beitrag geleistet werden. Der Anspruch, einen generischen Lösungsansatz zu entwickeln, gilt hierbei innerhalb der Domäne von DWH-Systemen.

5.5 Anforderungen an eine semantisch reichhaltige Beschreibung von Data-Warehouse-Metadaten

Die Erreichung des gegebenen Untersuchungsziels besitzt als Vorbedingung die semantisch reichhaltige Beschreibung der Metadaten faktischer DWH-Systeme (vgl. Abschnitt 1.2.2). Da bei der Lösungsstrategie lediglich die Schemaebene untersucht wird, entsteht der Anspruch, die Metadaten aus intensionaler Sicht so reichhaltig zu beschreiben, dass Fachanwender deren Extension eindeutig erschließen können.

In Tabelle 4 (Seite 49) wurden Metadaten-Kategorien vorgestellt, anhand derer eine eingehende Beschreibung von DWH-Metadaten erfolgen kann (*verwendungszweckgerichtete Metadatenkategorisierung*). Von diesen erscheint die Kategorie Datenqualität zunächst von untergeordneter Bedeutung, wenn die semantische Darstellung von DWH-Metadaten fokussiert wird. Datenqualitätsmerkmale sind zur korrekten semantischen Interpretation nicht zwingend erforderlich. Metadaten-Elemente zur Erfassung der Datenqualität werden aus diesem Grund nicht in die weiteren Betrachtungen einbezogen. Die anderen Metadaten-Kategorien werden für eine semantisch reichhaltige Auszeichnung von DWH-Metadaten um folgende Beschreibungsmerkmale erweitert bzw. detailliert:

METADATEN-KATEGORIE TERMINOLOGIE

- Standardmäßig werden *allgemeine Beschreibungskomponenten* zur Erfassung des Begriffsworts (hier die Bezeichnung eines Schemaelements), dessen Beschreibung und Definition vorausgesetzt. Darüber hinaus soll in die Metadaten-Beschreibung auch das individuelle Wissen der Mitarbeiter über eine Domäne (*implizites Wissen*) einfließen, sodass dieses der Organisation auf Dauer erhalten bleibt. Implizites Wissen ist für die korrekte Interpretation vieler Daten eine Voraussetzung. Zur Kodifizierung impliziten Wissens sind entsprechende Anreize zu schaffen (z. B. durch die Bereitstellung spezieller Metadaten-Elemente sowie eines leichten und intuitiven Zugangs zu deren Pflege).

- Mit der Realisierung einer semantisch reichhaltigen Metadaten-Beschreibung für DWH-Systeme soll auch der Aufbau eines *semantisch homogenen und logisch einheitlichen Begriffssystems* erfolgen. Für die hierzu notwendige Bereinigung sprachlicher Defekte und Klärung definitorischer Aspekte von Fachbegriffen (insbesondere die Normierung der Benennung und Bedeutung) kann bspw. die Methode zur Rekonstruktion von Fachbegriffen eingesetzt werden. In diesem Zusammenhang sind auch Vorgaben hinsichtlich der *Datenstruktur* (z. B. Datentyp, Datenformat) zu vereinbaren. Die Ergebnisse sind *formal* und in *einheitlicher Form* zu erfassen, sodass sie nachhaltig determiniert sind und jederzeit gleichbleibend nachvollzogen werden können. Zur Unterstützung dieser Anforderung sind hinreichend mächtige Technologien zu wählen (vgl. Kapitel 6). Die Formalisierung stellt zudem eine Voraussetzung für *Maschinenlesbarkeit* dar.
- Bei einer semantisch reichhaltigen Metadaten-Beschreibung dürfen die, gemäß den bisher geschilderten Anforderungen, formal und somit eindeutig definierten Fachbegriffe jedoch nicht isoliert betrachtet werden. Essentiell für deren korrekte Verwendung sind die *semantische Beschreibung der Abhängigkeiten und der Beziehungen* zwischen den Fachbegriffen. Hierzu zählen die Art, die Semantik und die Restriktionen der Beziehung (z. B. Kardinalitätsvorgaben). Auch die *Angabe der Position des Begriffs im Begriffssystem* bzw. des Schemaelements im Datenschema ist zu berücksichtigen. Die Metadaten sind unter Einbeziehung dieser Beschreibungskomponenten miteinander entsprechend zu verknüpfen.

METADATEN-KATEGORIE ORGANISATIONSBEZUG

- Die korrekte Interpretation von Daten erfordert auch Angaben über den Kontext, dem die Daten entstammen. Eine semantisch reichhaltige Metadaten-Beschreibung eines DWH-Systems muss daher Elemente aufweisen, welche Informationen aus den identifizierten Kontexttypen *Domänen-, Umwelt-, Aufgaben- und Nutzerkontext* erfassen können. Neben organisationalen Kontextinformationen können hier auch Angaben über dem DWH *vorgelagerte Systeme* sowie deren *fachliche und technische Ansprechpartner* Beachtung finden.

FORMALZIELE DER SEMANTISCH REICHHALTIGEN METADATEN-BESCHREIBUNG

Das Formalziel der *Homogenität* einer semantisch reichhaltigen Metadaten-Beschreibung wurde bereits im Rahmen der obigen Ausführungen herausgestellt. Darüber hinaus sind *Vollständigkeit* und *Konsistenz* der Metadaten-Beschreibung eines DWH-Systems anzustreben. Neben den bisher genannten Anforderungen hinsichtlich des Aufbaus und der Erfassung semantisch reichhaltiger DWH-Metadaten ist die *adäquate Bereitstellung* für den Endanwender ein weiteres Formalziel. Hier sind insbesondere die Kriterien intuitives und vollständiges Auffinden und Verständnis der Metadaten sowie die semantische Navigation durch die Metadaten zu nennen.

5.6 Zusammenfassung

In diesem Kapitel wurden eingangs Grundlagen der Semiotik erörtert. In diesem Zusammenhang wurde auch die Ungenauigkeit natürlichsprachlicher Begriffsbeschreibungen aufgezeigt. Die für eine exakte Begriffsbeschreibung erforderlichen Repräsentationsebenen bildeten den Ausgangspunkt für die Diskussion sprachlicher Defekte. Strategien zu deren Vermeidung und Auflösung waren dabei ebenfalls eingeschlossen.

Des Weiteren wurden Arten und Ursachen von Heterogenität untersucht. Dabei stand insbesondere die semantische Heterogenität von Datenschemata im Fokus. Ferner konnten Gründe identifiziert werden, die gegen eine explizite Auflösung der Heterogenität sprechen und somit das gesetzte Untersuchungsziel, die Überwindung semantischer Heterogenität, bekräftigen. Letztere kann zusätzlich durch die Einbeziehung des erläuterten semantischen Ähnlichkeitsmaßes gestützt werden.

Forschungsansätze, die sich der Überwindung von Heterogenität bei Datenschemata widmen, wurden unter der Überschrift Schema-Management diskutiert. Bisherige Lösungsansätze konzentrieren sich auf Fragen struktureller Heterogenität oder scheitern infolge nicht formal spezifizierter bzw. fehlender semantischer Angaben über reale Schemata. Generische Ansätze sind in diesem Bereich noch intensiv zu erforschen.

Die Ausführungen resultierten in den abschließend vorgestellten Anforderungen an eine semantisch reichhaltige Beschreibung von DWH-Metadaten. Besondere Merkmale sind hierbei die semantische Verknüpfung der Metadaten, die Einbeziehung von Kontextinformationen und das Ziel, individuelles Mitarbeiterwissen über die Daten zu explizieren. Die Metadaten sind zudem formal zu erfassen.

6 Exkurs: Konzepte des Semantic Web

In den vorausgehenden Kapiteln wurden Aspekte der Semantik und der Heterogenität intensiv erörtert. Darauf aufbauend wurden Anforderungen an eine semantisch reichhaltige Beschreibung von DWH-Metadaten vorgestellt. Um diese Anforderungen auch von der technischen Seite her stützen zu können, wird in diesem Kapitel der Blick auf Konzepte und Technologien des Semantic Web gerichtet.

Nach einer allgemeinen Einführung in die Vision des Semantic Web werden die standardisierten Technologien Uniform Resource Identifier (URI), Extensible Markup Language (XML), Resource Description Framework (RDF) und die Web Ontology Language (OWL) vorgestellt und hinsichtlich der in Abschnitt 5.5 genannten Anforderungen beleuchtet. Abschließend erfolgt eine Diskussion über die Eignung der Semantic-Web-Technologien zur Erreichung des in dieser Arbeit angestrebten Untersuchungsziels.

6.1 Motivation und Ziele des Semantic Web

Das *World Wide Web* (WWW) hat die Art und Weise verändert, in der Menschen miteinander kommunizieren, Informationen suchen oder Geschäfte und Handel betreiben. Das heute bekannte WWW hat sich dabei zu einem Medium entwickelt, welches Dokumente für Menschen bereitstellt, ohne dass deren Inhalte automatisiert verarbeitet werden können. Es stellt gewissermaßen eine Menge unorganisierter und nur indirekt zugänglicher Daten bereit, die von Menschen gelesen werden müssen, um deren Inhalte verstehen und interpretieren zu können. Suchmaschinen können beim Auffinden relevanter Dokumente unterstützen. Jedoch sind die auf Webseiten präsentierten Informationen bislang kaum für elektronische Agenten⁵⁵ nutzbar [Zieg07, S. 173].

⁵⁵ Unter einem *Agent* wird in diesem Zusammenhang ein, als selbstständige Einheit betrachtetes, automatisiertes System verstanden [FeSi06, S. 119].

Werden Webseiten systematisch mit auf Maschinen abgestimmten Daten angereichert und speziell für Software-Agenten ausgelegte Dokumente hinzugefügt, so kann das heute bekannte WWW in das so genannte *Semantic Web* überführt werden. Der Vision von BERNERS-LEE folgend, stellt das Semantic Web Strukturen für Inhalte mit einer wohldefinierten Bedeutung bereit. Ein wichtiges Ziel ist dabei die Interoperabilität zwischen den Datenbeständen, die über das Web zugänglich sind. *Semantic-Web-Technologien* bilden hierfür die Infrastruktur.

Im Semantic Web werden Daten in *Wissensausdrücken* (*Knowledge-Statements*) organisiert. Diese erlauben strukturierte Aussagen über Ressourcen im Web. Maschinen können die Bedeutung der semantisch angereicherten Daten herleiten, indem sie Hyperlinks zu den Definitionen der Schlüsselworte und zu den Regeln für Inferenz folgen. Software-Agenten sind somit in der Lage, Webseiten gezielt zu durchsuchen sowie deren Inhalte zu verstehen und automatisch zu verarbeiten. BERNERS-LEE erachtet in diesem Zusammenhang *Ontologien* (vgl. Abschnitt 6.3) als einen elementaren Baustein des Semantic Web [BeHL01], [Bern98a], [GrLe02], [Powe03, S. 2 f.].

Die im Kontext des Semantic Web einschlägigen Technologien werden im Folgenden vorgestellt und auf ihre semantische Mächtigkeit hin untersucht.

6.2 Technologien des Semantic Web

Die Technologien des Semantic Web lassen sich gemäß Abbildung 6-1 in aufeinander aufbauende Schichten anordnen. Höher liegende Schichten besitzen gegenüber darunter liegenden eine größere Komplexität und Ausdrucksstärke.

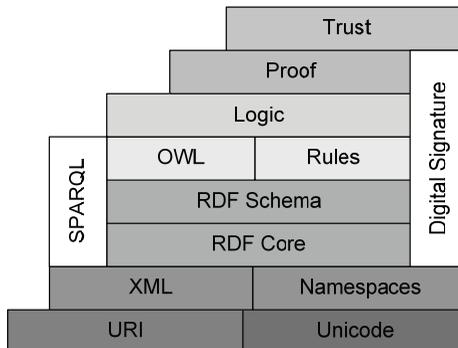


Abbildung 6-1: Semantic-Web-Schichten (in Anlehnung an [Brat07, S. 23], [Zieg07, S. 173])

Die Basisschicht bildet der Standard *Uniform Resource Identifier* (URI; vgl. Abschnitt 6.2.1). Ein URI dient der Identifikation einer Ressource. Daneben ist *Unicode*⁵⁶ zur Zeichenkodierung gesetzt. Auf der nächsthöheren Schicht sind die *Extensible Markup Language* (XML), als Vehikel für die Übertragung von Informationen im Web, sowie das Konzept der *Namensräume* (Namespaces) angeordnet. Mithilfe von Namensräumen können Auszeichnungselemente eines XML-Dokuments eindeutig festgelegt und referenziert werden (vgl. Abschnitt 6.2.2). Die übergeordnete Schicht beinhaltet das *Resource Description Framework* (RDF; vgl. Abschnitt 6.2.3). RDF gilt als das Basismodell für jegliche Inhalte des Semantic Web. Um die Bedeutung von Ressourcen sowie die Art möglicher Beziehungen zwischen diesen formal und exakt zu definieren, wird *RDF-Schema* eingesetzt (vgl. Abschnitt 6.2.3). *SPARQL* ist die Empfehlung des W3C⁵⁷ als Anfragesprache für RDF (vgl. Abschnitt 6.2.3).

Die *Web Ontology Language* (OWL) baut auf RDF/RDF-Schema auf und stellt eine vollwertige Sprache zur Definition von Ontologien dar (vgl. Abschnitt 6.3.3). Neben OWL sind *Regeln* (Rules) ebenfalls dieser Schicht zugewiesen. Regeln erlauben es, Handlungsanweisungen bspw. in der Form einer Wenn-Dann-Anweisung auszudrücken. OWL, ergänzt um Regeln, stellt die Basis für die *Logik*-Schicht dar. Die beiden oberen Schichten *Proof* und *Trust* sind bisher kaum ausgeprägt. Proof zielt unter anderem darauf ab, die Inferenzschritte einer Maschine transparent und nachvollziehbar zu gestalten. Trust als Schutzmechanismus basiert auf der Idee, bekannten Quellen und Agenten mehr Vertrauen zu schenken als unbekanntem. Die *digitale Signatur* (*Digital Signature*)⁵⁸ ist orthogonal zu den Schichten RDF bis Proof dargestellt. Durch sie kann bspw. die Urheberschaft einer Nachricht verifiziert werden [Zieg07, S. 173 ff.], [Bern98a], [DaOS03, S. 104].

⁵⁶ *Unicode* ist ein internationaler Standard, der es ermöglicht, Textzeichen unterschiedlicher Sprach- und Wirtschaftsräume sowie Bildzeichen durch einen eindeutigen Zeichencode darzustellen [ScWe07, S. 72], <http://www.unicode.org> (Abruf am 10.06.2008).

⁵⁷ Informationen zum *World Wide Web Consortium* (W3C) sind unter <http://www.w3.org> verfügbar (Abruf am 10.06.2008).

⁵⁸ Die digitale Signatur zählt zu den asymmetrischen kryptographischen Authentifikationssystemen. Sie arbeitet mit einem Schlüsselpaar, das aus einem privaten und einem öffentlichen Schlüssel besteht [ScWe07, S. 496 ff.].

6.2.1 Uniform Resource Identifier (URI)

Um auf ein Objekt bzw. eine Ressource zugreifen zu können, ist ein eindeutiger Name oder ein Identifikator notwendig. Der *URI-Standard*⁵⁹ bietet eine einheitliche Syntax, um Ressourcen zu identifizieren. Mithilfe eines URI kann ein Bezeichner festgelegt werden, den alle Beteiligten akzeptieren [Powe03, S. 21 f.], [Zieg07, S. 173].

Ein URI stellt bspw. einen Namen oder eine Ortsangabe dar. Dabei legt ein *Uniform Resource Name (URN)* einen global eindeutigen Namen fest (z. B. International Standard Book Number (ISBN)), ein *Uniform Resource Locator (URL)* bestimmt den Ort eines Objekts (z. B. Adresse einer Webseite – <http://www.semantic-metadata.com/index.htm>) [ScWe07, S. 378].

Ein URI kann sowohl einen absoluten als auch einen relativen Pfad zu einer Ressource beschreiben. Optional kann mithilfe eines Fragment-Identifikators (#) auf einen spezifischen Absatz in einem Dokument verwiesen werden [Powe03, S. 22].

6.2.2 Extensible Markup Language (XML)

Die *Extensible Markup Language (XML)* ist eine Metasprache, die es erlaubt, beliebige Auszeichnungssprachen für Dokumente bzw. Datenaustauschformate zu definieren. Sie gibt zugleich die grundsätzliche Syntax der Auszeichnungssprache vor. Bei XML wird eine klare Trennung zwischen den Metadaten, den Daten sowie der Darstellung der Daten verfolgt [EcEc04, S. 5 ff.], [LeNa07a, S. 297].

XML ist eine Teilmenge der *Standard Generalized Markup Language (SGML)*⁶⁰ und bietet Unicode-Unterstützung. Seit 1998 ist XML vom W3C standardisiert.⁶¹

Aufgrund seiner Flexibilität verfügt XML über ein breites Spektrum an Einsatzmöglichkeiten. Insbesondere bei der Strukturierung von Metadaten sowie beim Austausch von Daten im WWW hat XML hohe Bedeutung erlangt [BöU199].

⁵⁹ Die aktuelle URI-Syntaxspezifikation ist in <http://www.ietf.org/rfc/rfc3986.txt> definiert (Abruf am 10.06.2008).

⁶⁰ Informationen zur *SGML* (ISO-8879) offeriert unter anderem folgende Webseite: <http://www.w3.org/MarkUp/SGML> (Abruf am 10.06.2008).

⁶¹ Das W3C bietet detaillierte Informationen zu *XML* unter <http://www.w3.org/XML> an (Abruf am 10.06.2008).

AUFBAU EINES XML-DOKUMENTS

Jedes XML-Dokument besteht aus den Teilen *XML-Prolog* (optional), *Dokumenttyp-Definition DTD* (optional) und *Dokumentinstanz*. Im XML-Prolog werden grundlegende Eigenschaften, wie bspw. die XML-Versionsnummer oder die Art der Zeichenkodierung festgelegt. Eine DTD gibt die erlaubte Struktur der Dokumentinstanz vor, welche den strukturierten Text enthält. Alternativ zu einer DTD kann XML-Schema zur Beschreibung formal syntaktischer Anforderungen an XML-Dokumente eingesetzt werden. XML-Schema erlaubt eine differenziertere Beschreibung der gewünschten Strukturen als DTD (z. B. genaue Angabe von Kardinalitäten, Definition komplexer Typen). XML-Schemata sind ebenfalls in der Beschreibungssprache XML formuliert [EcEc04, S. 25 ff.], [ScWe07, S. 393 ff.], [HKRS08, S. 19 ff.].

GÜLTIGKEIT UND WOHLGEFORMTHEIT

XML-Dokumente können auf Korrektheit überprüft werden. Ein Dokument mit korrekter Auszeichnung durch XML-Elemente wird als *wohlgeformt* bezeichnet. Ist ein Dokument darüber hinaus konform im Hinblick auf eine spezifische DTD, so spricht man von *Gültigkeit* oder auch *Validität*. Die Überprüfung auf Korrektheit kann automatisiert durch einen so genannten *XML-Parser*⁶² erfolgen [BöU199], [Wals98].

XML-NAMENSRÄUME

Die Möglichkeit der freien Definition von Auszeichnungselementen birgt die Gefahr der Mehrdeutigkeit. Zur Vermeidung von Namenskonflikten können *Namensräume* (*Namespace*) genutzt werden. Ein Namensraum stellt eine Sammlung von Element- und Attributnamen dar, die mithilfe von URI (vgl. Abschnitt 6.2.1) identifiziert werden und verschiedene Vokabulare definieren. Die Vokabulare können über ihren Namensraum in beliebige XML-Dokumenten eingebunden und dort mit anderen kombiniert werden [DaOS03, S. 42 ff.], [EcEc04, S. 51 ff.], [BHLT06].

⁶² Eine Aufstellung verschiedener *XML-Parser* bietet bspw. <http://www.jeckle.de/xml/index.html#parser> (Abruf am 10.06.2008).

DISKUSSION DER XML

XML ist eine standardisierte und etablierte Metasprache, die Maschinenlesbarkeit ermöglicht. Insbesondere im Hinblick auf die Formalisierung von Semantik weist sie jedoch Defizite auf:

- Die Metadaten eines XML-Dokuments werden durch so genannte Auszeichnungselemente repräsentiert. Die Namen und die Bedeutung der Auszeichnungselemente können beliebig festgelegt werden. Somit kann die Bedeutung dieser Auszeichnungselemente ggf. für einen Menschen intuitiv erfassbar sein, nicht jedoch für eine Maschine. Es fehlt die Möglichkeit, die Bedeutung der Auszeichnungselemente derart zu kodieren, dass eine maschinenseitige Verarbeitung sowie eine Herleitung nicht explizit gegebenen Wissens möglich sind. Mithilfe einer DTD oder eines XML-Schemas kann lediglich die Syntax eines XML-Dokuments festgelegt werden. Es werden keine ontologiebasierten Modellierungsprimitive zur Formalisierung von Semantik angeboten. Zudem werden in XML jeweils nur einzelne Konzepte oder Objekte beschrieben [DaOS03, S. 51 f.], [EcEc04, S. 235], [AnHa04b, S. 61], [HKRS08, S. 29 f.]
- Eine spezifische Information kann in einem gültigen XML-Dokument unterschiedlich strukturiert dargestellt werden. Zudem besitzt XML eine hierarchisch geschachtelte Struktur, d. h. alle Auszeichnungselemente müssen wiederum innerhalb der Auszeichnungselemente eingebettet sein, die sie beschreiben. Insbesondere bei Abfragen auf XML-Dokumenten führen diese Eigenschaften zu hohen Kosten [Bern98b], [StHa05b, S. 5 ff.].
- Dieselbe Information kann in zwei XML-Dokumenten an unterschiedlichen Stellen des jeweiligen XML-Baums stehen. Eine syntaktische Interoperabilität auf Basis von XML ist daher nur möglich, wenn die jeweiligen Auszeichnungselemente der XML-Dokumente den beteiligten Maschinen bekannt sind und von diesen verstanden werden. Der Einsatz von XML für die Datenintegration und den Datenaustausch erscheint daher lediglich in abgegrenzten und stabilen Umgebungen praktikabel [BeFe01], [Powe03, S. 4 ff.], [HKRS08, S. 37].

Die Diskussion von XML zeigt, dass XML entscheidende Konzepte fehlen, um die gewünschte semantisch reichhaltige Beschreibung von DWH-Metadaten zu erreichen und die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen zu ermöglichen. Aus diesem Grund wird nachfolgend das auf der nächsthöheren Sprachstufe angeordnete *Resource Description Framework* dahingehend untersucht.

6.2.3 Resource Description Framework (RDF)

Das vom W3C standardisierte *Resource Description Framework* (RDF)⁶³ stellt ein Metadaten-Rahmenwerk dar, um Ressourcen, insbesondere im WWW, einheitlich zu annotieren sowie diese in einer maschinenlesbaren und -interpretierbaren Form zu speichern. Es zählt zu den grundlegenden Technologien bei der Gestaltung des Semantic Web. Im Unterschied zum hierarchischen XML ist RDF ein graphbasiertes Datenmodell [Powe03, S. 1 ff.], [LeNa07a, S. 299].

Zur korrekten maschinellen Interpretation besitzt RDF eine *formale Semantik*⁶⁴. Neben einem fundierten Datenaustausch werden hierdurch effektive Suchanfragen und Navigationsmöglichkeiten in Datenbeständen sowie die Katalogisierung der Daten unterstützt. Suchergebnisse können bspw. durch die Berücksichtigung semantischer Aspekte bei der Anfragedefinition verbessert werden [Powe03, S. 1 ff.], [Brid04, S. 52], [K1Ca04], [LeNa07a, S. 295 ff.], [HKRS08, S. 35 ff.].

Anzumerken ist, dass bei RDF von der *Open World Assumption* (OWA) ausgegangen wird, d. h. es wird unterstellt, dass eine Wissensbasis stets potentiell unvollständig ist. Eine nicht belegbare Aussage wird daher weder als wahr noch als falsch angenommen. Bei der *Closed World Assumption* (CWA) wird dagegen zugrunde gelegt, dass alle relevanten Fakten in der Wissensbasis vorhanden sind. Eine nicht explizit belegbare Aussage wird als falsch angenommen [HKRS08, S. 145 ff.].

⁶³ Die W3C-Spezifikation des *RDF* ist unter <http://www.w3.org/RDF> abrufbar (Abruf am 10.06.2008).

⁶⁴ Eine mathematische Betrachtung des Begriffs der *formalen Semantik* in Bezug auf RDF wird z. B. in [HKRS08, S. 91 ff.] gegeben.

RDF-WERKZEUGE UND REFERENZANWENDUNG

Für die Erstellung, Bearbeitung und das Parsen von RDF-Dokumenten existieren frei verfügbare Werkzeuge. Hierzu zählen bspw. *BrownSauce*⁶⁵, ein java-basierter RDF/XML-Browser, die *ICS-FORTH RDF-Suite*⁶⁶, welche u. a. einen Parser für RDF-Dokumente sowie eine RDF-Datenbank umfasst, *IsaViz*⁶⁷, ein vom W3C bereitgestelltes Entwicklungswerkzeug zur graphischen Darstellung und Bearbeitung von RDF, sowie *Closed World Machine*⁶⁸, das als Allzweck-Werkzeug für die Verarbeitung von Daten für das Semantic Web anzusehen ist. Auch kommerzielle Datenbankhersteller bieten inzwischen Unterstützung für RDF-Daten an [Czar07].

Eine prominente Referenzanwendung von RDF ist das Metadaten-Schema *Dublin Core* (DC)⁶⁹, das insbesondere für den Austausch bibliographischer Daten entwickelt wurde [EcEc04, S. 290], [LeNa07a, S. 276].

Im Folgenden werden der Aufbau sowie die Möglichkeiten der Serialisierung und der Darstellung von RDF-Dokumenten erläutert. Zudem wird RDF-Schema als Regelwerk zur Erstellung von RDF-Vokabularen und die RDF-Anfragesprache SPARQL vorgestellt. Abschließend erfolgt eine kritische Würdigung von RDF.

AUFBAU VON RDF

RDF definiert eine Sprache zur Beschreibung von *Ressourcen*. Ressourcen können sowohl physische Objekte als auch abstrakte Konzepte sein. Sie müssen lediglich eine Identität besitzen und über einen URI adressierbar sein [AnHa04b, S. 63], [Brid04, S. 51].

Anyone Can Make Statements About Any Resource [KICa04].

⁶⁵ Vgl. <http://brownsauce.sourceforge.net> (Abruf am 10.06.2008).

⁶⁶ Vgl. <http://139.91.183.30:9090/RDF> (Abruf am 10.06.2008).

⁶⁷ Vgl. <http://www.w3.org/2001/11/IsaViz> (Abruf am 10.06.2008).

⁶⁸ Vgl. <http://www.w3.org/2000/10/swap/doc/cwm.html> (Abruf am 10.06.2008).

⁶⁹ Für ausführliche Informationen zu *Dublin Core* wird auf <http://dublincore.org> verwiesen (Abruf am 10.06.2008).

Zentrales Element von RDF sind so genannte *Aussagen* (Statements), mit denen die Werte von Eigenschaftstypen von Ressourcen festgelegt werden. Ein Individuum soll mithilfe von RDF Aussagen über beliebige Ressourcen treffen können. Gleiche Aussagen werden dabei stets gleichartig dargestellt [Zieg06, S. 76 f.], [EcEc04, S. 239 ff.], [KICa04], [LeNa07a, S. 299 ff.].

[...] three pieces of information are all that's needed in order to fully define a single bit of knowledge [Powe03, S. 17].

Eine RDF-Aussage besteht aus einem Tripel, welches sich aus den Bestandteilen *Subjekt*, *Prädikat* und *Objekt* zusammensetzt. Das Subjekt ist allgemein gesprochen die Ressource, die beschrieben wird. Diese ist durch einen URI bestimmt. Das Prädikat stellt den Eigenschaftstyp (z. B. ein Attribut, eine Beziehung oder ein spezifisches Kennzeichen) der Ressource dar und ist ebenfalls durch einen URI identifiziert. Das Objekt entspricht dem Wert des Eigenschaftstyps des beschriebenen Subjekts und kann entweder durch ein *Literal* oder eine Ressource angegeben werden. Jedes RDF-Tripel kann mit weiteren RDF-Tripeln verknüpft werden. Dabei behält es seinen eindeutigen Aussageinhalt [DaOS03, S. 88 ff.], [Powe03, S. 21 f.], [Zieg06, S. 76 f.].

Abbildung 6-2 visualisiert die Zusammensetzung eines RDF-Tripels.

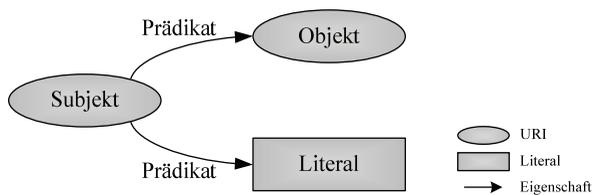


Abbildung 6-2: RDF-Tripel [DaOS03, S. 88]

RDF-GRAPH

Für RDF-Datenmodelle wurde *RDF-Graph* als Standard-Repräsentationsform festgelegt. Ein RDF-Graph besteht aus einer Menge von beschrifteten Knoten, die über gerichtete und ebenfalls beschriftete Kanten miteinander verbunden sind. Hierdurch entsteht ein Knoten-Kante-Knoten-Muster [Powe03, S. 19 ff.], [AnHa04b, S. 63 ff.].

Ellipsenförmige Knoten stellen eine Ressource dar, rechteckige literale Werte. Eine Kante repräsentiert die Eigenschaft einer Ressource. Sie beginnt bei einem Subjekt (Ressource) und endet bei dem zugehörigen Objekt. Diese Art von Graph ist im Forschungsbereich der Künstlichen Intelligenz auch als semantisches Netz bekannt [HKRS08, S. 36 f.], [Brid04, S. 53 ff.], [AnHa04b, S. 65].

RDF-Graphen können miteinander verknüpft werden. Es existiert jedoch keine Vorgabe, dass in einem RDF-Graph alle Knoten miteinander verbunden sein müssen. RDF-Graphen insistieren lediglich auf gültige Tripel und ein gültiges RDF. Nachstehende Abbildung zeigt beispielhaft einen RDF-Graphen.

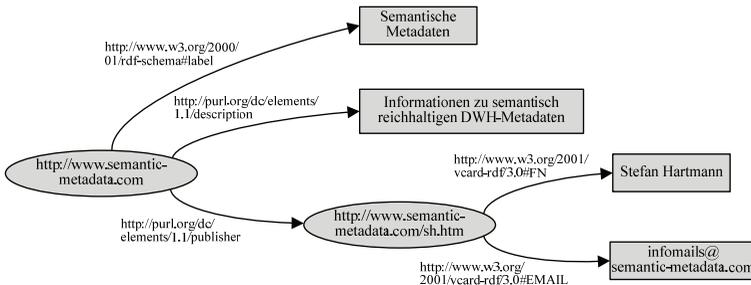


Abbildung 6-3: RDF-Graph

Beim Vergleich zweier RDF-Graphen ist wie folgt zu differenzieren [Powe03, S. 27 f.]:

- *Entailment*: Zwei RDF-Graphen sind identisch in allen Aspekten. Jede Aussage, die über den einen Graph getroffen wurde, ist ebenfalls für den anderen wahr. Die Information in den beiden Graphen ist gleich, kann aber durch unterschiedliche Repräsentationen abgebildet sein.
- *Equality*: Zwei RDF-Graphen sind identisch. Sie besitzen die gleichen, benannten Knoten.

SERIALISIERUNG ALS RDF/XML

Die Einsatzfelder von RDF korrelieren mit den Möglichkeiten der Portabilität von RDF-Daten. RDF-Graph bietet keine effizienten Ansätze zur Speicherung und Übermittlung der beschriebenen Daten [Powe03, S. 29], [HKRS08, S. 42].

Da RDF unabhängig von der Form der Darstellung ist, können RDF-Daten auf verschiedene Arten *serialisiert*⁷⁰ werden. Zu den bekanntesten Notationen zählen *N-Tripel*⁷¹, welche eine Untermenge der *N3-Notation*⁷² bezeichnen, und *RDF/XML*. Letztere steht insbesondere aufgrund ihrer weiten Verbreitung und der programmiersprachlichen Unterstützung bei der Verarbeitung von XML-Daten im Fokus der nachstehenden Untersuchungen. Darüber hinaus können zum Parsen von RDF/XML-Dokumenten herkömmliche XML-Technologien genutzt werden [HKRS08, S. 40 ff.].

Ein gemäß RDF/XML serialisiertes RDF-Dokument ist ein wohlgeformtes XML-Dokument, das über zusätzliche Merkmale und Einschränkungen verfügt, welche die Erfassung, den Austausch und die Zusammenführung von Daten unterschiedlicher Modelle unterstützen [Powe03, S. 29].

Nachstehend finden sich die im obigen RDF-Graph visualisierten Aussagen in RDF/XML-Darstellung. Innerhalb des öffnenden RDF-Wurzelements `<rdf:RDF>` werden die benötigten Namensräume deklariert. Daran schließt sich die Beschreibung der Ressourcen anhand ihrer jeweiligen Eigenschaften und Wertzuweisungen an. Ein RDF-Dokument endet mit einem schließenden RDF-Auszeichnungselement. Für eine detaillierte Erläuterung zur RDF/XML-Syntax wird auf die Literaturquellen [Powe03, S. 29 ff.], [AnHa04b, S. 69 ff.], [EcEc04, S. 241 ff.], [HKRS08, S. 42 ff.] sowie auf [Beck04] verwiesen.

⁷⁰ Unter *Serialisierung* wird u. a. die Konvertierung eines Objekts in eine persistente Form verstanden [Powe03, S. 30].

⁷¹ Eine Einführung in *N-Tripel* ist bspw. in [Powe03, S. 22 ff.], [HKRS08, S. 40 ff.] nachzulesen.

⁷² Informationen zur *N3-Notation* finden sich z. B. unter <http://www.w3.org/DesignIssues/Notation3.html> und <http://www.w3.org/2000/10/swap/Primer.html> (Abruf jeweils am 10.06.2008).

```

<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:vcard="http://www.w3.org/2001/vcard-rdf/3.0#" >
  <rdf:Description rdf:about="http://www.semantic-metadata.com">
    <rdfs:label>Semantische Metadaten</rdfs:label>
    <dc:description>Informationen zu semantisch
      reichhaltigen DWH-Metadaten</dc:description>
    <dc:publisher
      rdf:resource="http://www.semantic-metadata.com/sh.html"/>
  </rdf:Description>
  <rdf:Description
    rdf:about="http://www.semantic-metadata.com/sh.html">
    <vcard:FN>Stefan Hartmann</vcard:FN>
    <vcard:EMAIL>infomails@semantic-metadata.com</vcard:EMAIL>
  </rdf:Description>
</rdf:RDF>

```

Quellcode 1: RDF-Dokument in RDF/XML-Notation

In einem wohlgeformten RDF/XML-Dokument ist jedes Prädikat durch eine Ressource sowie jede Ressource durch ein Prädikat separiert. Die Syntaxdarstellung in RDF/XML spiegelt somit das vom RDF-Graph bekannte Knoten-Kante-Knoten-Muster in einem *Streifenmuster* (Striped Syntax)⁷³ wider [Powe03, S. 35 f.].

RDF-SCHEMA

Mithilfe von RDF können elementare Aussagen definiert, jedoch keine expliziten Angaben über deren Bedeutung getroffen werden [Brid04, S. 56], [Zieg07, S. 174]. *RDF-Schema* (RDFS) stellt daher ein domänen-neutrales Regelsystem bereit, mit dem fachspezifische RDF-Vokabulare erstellt werden können. Die Anwendungsbreite und -vielfalt möglicher Vokabulare ist dabei unbegrenzt.

⁷³ Ergänzende Erläuterungen zur *Striped Syntax* von RDF/XML offeriert <http://www.w3.org/2001/10/stripes> (Abruf am 10.06.2008).

Ein RDF-Vokabular beschreibt *Eigenschaften*, *Klassen* von Ressourcen und *Beziehungen* zwischen diesen. Die Vokabularelemente sind jeweils durch einen URI eindeutig bestimmt. RDFS organisiert das Vokabular in einer typisierten Hierarchie und erlaubt die semantische Einordnung von Begriffen. Die maßgeblichen Konzepte hierfür sind *Unterklasse-* und *Untereigenschaftsbeziehungen*. Die beiden Konstrukte sind jeweils als transitiv anzusehen [AnHa04b, S. 84 ff.], [Brid04, S. 56 ff.], [EcEc04, S. 235, S. 259], [BrGu04], [Powe03, S. 83].

VALIDITÄT VON RDF-DOKUMENTEN

RDFS schränkt zum einen die vorgegebene Struktur für RDF-Dokumente und zum anderen die Verwendung der erlaubten Bezeichner ein. Das Element *rdfs:domain* bestimmt dabei die Klassen, mit denen eine Eigenschaft kombiniert werden darf (Subjekte), das Element *rdfs:range* spezifiziert den erlaubten Wertebereich von Eigenschaften (Objekte) [Brid04, S. 59], [EcEc04, S. 264 f.], [AnHa04b, S. 86 f.] [LeNa07a, S. 305].

Anhand eines RDFS kann ein RDF-Dokument auf Validität geprüft werden. Hierfür stehen entsprechende Parser zur Verfügung (z. B. *VRP* aus der *ICS-FORTH RDF-Suite*). RDF-Schemata erlauben somit die Sicherstellung syntaktischer und semantischer Konsistenz von RDF-Dokumenten.

Weiterführende Informationen zu RDFS bieten u. a. folgende Literaturquellen: [BrGu04], [AnHa04b, S. 84 ff.], [HKRS08, S. 91 ff.].

ANFRAGESPRACHEN FÜR RDF

Für RDF-Dokumente stehen diverse Anfragesprachen zur Verfügung. Einen fundierten Überblick über RDF-Anfragesprachen bietet bspw. [Morg06, S. 177 ff.]. Zu den namhaftesten zählen *RDF Data Query Language* (RDQL)⁷⁴, *Sesame Resource Query Language* (SeRQL)⁷⁵ und *SPARQL Protocol and RDF Query Language* (SPARQL). SPARQL besitzt inzwischen den Status einer W3C-Empfehlung (W3C Recommendation)⁷⁶ als Anfragesprache für RDF und wird im Hinblick auf seine möglicherweise bedeutende Rolle nachstehend kurz erläutert.

⁷⁴ Vgl.: <http://www.w3.org/Submission/2004/SUBM-RDQL-20040109> (Abruf am 10.06.2008).

⁷⁵ Vgl.: <http://www.openrdf.org/doc/sesame/users/ch06.html#d0e1056> (Abruf am 10.06.2008).

⁷⁶ Die vom W3C vorgeschlagene *SPARQL-Spezifikation* kann unter <http://www.w3.org/TR/rdf-sparql-query> eingesehen werden (Abruf am 10.06.2008).

Eine SPARQL-Anfrage wird auf einem RDF-Graphen anhand von Graph-Mustern ausgeführt. Die Tripel des RDF-Graphen (Subjekt, Prädikat und Objekt) stellen die so genannten *Datentripel* dar, die Tripel der SPARQL-Anfrage die *Anfragetripel*. Subjekt, Prädikat und Objekt des Anfragetripels sind durch Konstanten oder Variablen definiert. Syntaktisch orientiert sich SPARQL an SQL. Innerhalb der `select`-Klausel werden die Variablenbindungen bestimmt, die als Ergebnis einer Anfrage geliefert werden sollen. In der `where`-Klausel werden diese Bindungen über Graph-Muster (Anfragetripel) hergestellt. Das Schlüsselwort `prefix` am Beginn einer Anfrage dient der Deklaration eines Namensraums [LeNa07a, S. 307 ff.].

Ein Anfragetripel stimmt mit einem Datentripel überein, wenn seine durch Konstanten definierten Bestandteile mit den zugehörigen Werten eines Datentripels übereinstimmen. Die Variablen werden dabei an die entsprechenden Werte des Datentripels gebunden. Für umfassende Anfragen können Anfragetripel zu komplexen Anfrage-Mustern kombiniert werden. Des Weiteren stehen u. a. Elemente zur Filterdefinition, zur logischen ODER-Verknüpfung sowie der UNION-Operator zur Verfügung. Mechanismen zur Sichtenbildung oder die Möglichkeit, Schemainformationen explizit anzugeben, fehlen jedoch [LeNa07a, S. 307 ff.], [Zieg07, S. 174], [HKRS08, S. 202 ff.].

Für eine eingehende Diskussion der RDF-Anfragesprache SPARQL wird auf die genannte W3C-Spezifikation sowie die Literaturquellen [LeNa07a, S. 307 ff.] und [HKRS08, S. 201 ff.] verwiesen.

VERKNÜPFUNG VON URI, XML UND RDF

Anhand der folgenden Abbildung wird das Zusammenwirken von URI, XML und RDF bei der Beschreibung von Ressourcen erläutert.

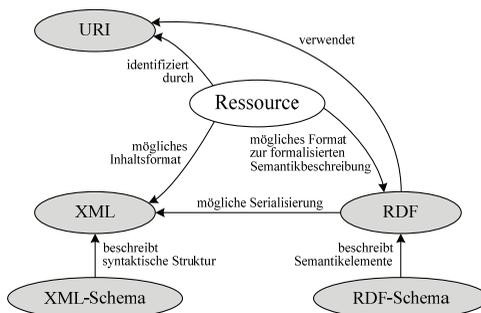


Abbildung 6-4: Beziehungen zwischen URI, XML und RDF [Jeck03, S. 23]

XML dient im Allgemeinen der standardisierten Inhaltsdarstellung. Mithilfe eines XML-Schemas kann die syntaktische Struktur eines XML-Dokuments festgelegt werden. XML stellt eine mögliche Serialisierungsform von RDF dar. RDF kann zur formalen Semantikbeschreibung einer Ressource eingesetzt werden. Hierbei dient ein RDF-Schema zur Beschreibung der Semantikelemente eines RDF-Dokuments. RDF nutzt das URI-Konzept, um eine Ressource eindeutig zu bestimmen.

DISKUSSION DES RDF

Die Semantic-Web-Technologien RDF und RDFS sind für den dezentralen Auf- und Ausbau fachspezifischer Vokabulare konzipiert. Sie unterstützen außerdem die automatische Verarbeitung von Daten, die Kombination von Daten unterschiedlicher Vokabulare und bieten gegenüber XML verbesserte Abfragemöglichkeiten [Brid04, S. 62 f.]. Vor- und Nachteile von RDF werden mit Blick auf die Unterstützung semantischer Kriterien nachstehend diskutiert:

- Die Repräsentation von Aussagen resultiert in RDF in einer flachen, tripelbasierten Struktur. Unabhängig vom Vorkommen einer Aussage in einem RDF-Dokument erzeugt diese stets dieselben Effekte sowie den jeweils selben RDF-Baum. Bei der Suche nach einer bestimmten Aussage in einem RDF-Dokument muss daher lediglich das spezifische Tripel über seinen URI gesucht werden. Bei XML-Dokumenten sind dagegen alle Elemente eines Dokuments zu prüfen, um sicherzustellen, dass der korrekte Wert gefunden wurde. Bei der Weiterverarbeitung der Ergebnismenge können bei RDF die nicht benötigten Aussagen ignoriert werden. Dies ist bei einer XML-Datei nicht möglich [Bern98b].
- RDFS kann eingesetzt werden, um eine redundante Spezifikation derselben Information zu vermeiden (Eindeutigkeit der Vokabularelemente durch das URI-Konzept). Dennoch sind Absprachen und Standardisierungen bei der Definition eines RDF-Vokabulars erforderlich, um z. B. sprachliche Defekte (vgl. Abschnitt 5.2.2) zu vermeiden. Es existieren keine Beschränkungsmöglichkeiten, die sicherstellen, dass gleiche Begriffe in unterschiedlichen RDFS-Modellen die gleiche Bedeutung haben oder dass Werte in gleicher Einheit und Notation erfasst werden [LeNa07a, S. 304].

- RDF kann in Verbindung mit RDFS als Format zur Repräsentation von Taxonomien und einfachen Ontologien dienen (vgl. Abschnitt 6.3). Für eine aussagekräftige Beschreibung von Begriffssystemen fehlen RDF und RDFS jedoch wesentliche Konzepte, zu denen insbesondere die nachfolgend genannten zählen [AnHa04b, S. 111 f.], [Lacy05, S. 132], [HKRS08, S. 118 f.]:
 - Die *Disjunktheit* von Klassen kann nicht explizit ausgedrückt werden.
 - Die *Boolesche Verknüpfung* von Klassen durch Vereinigung, Schnittmenge oder Komplement wird nicht unterstützt.
 - *Kardinalitätsbeschränkungen*, um festzulegen, wie viele eindeutige Werte eine Eigenschaft mindestens tragen muss bzw. maximal besitzen darf, können nicht definiert werden.
 - Die Auszeichnung von *Ähnlichkeiten* ist nicht möglich.
 - *Negation* und folglich auch *Kontradiktion* sind in RDFS nicht formalisierbar.
 - Spezielle Charakteristika von Eigenschaften wie beispielsweise *Transitivität*, *Einzigartigkeit* oder *Inversion* können nicht beschrieben werden.

In schwachen Formalismen, zu denen wie aufgezeigt auch RDF zählt, können nur sehr einfache Abfragen gestellt und beantwortet werden. Auch die Möglichkeiten, in diesem Rahmen neue Fakten herzuleiten, sind stark eingeschränkt bzw. nicht vorhanden [HKRS08, S. 201]. RDF und RDFS stehen daher in der Rolle, als Basis für mächtigere Ontologiesprachen (vgl. Abschnitt 6.3.2) zu fungieren und von diesen erweitert zu werden [Brid04, S. 62 f.], [AnHa04b, S. 69], [Powe03, S. 8 ff.].

Mit RDF und RDFS können bereits grundsätzliche Anforderungen an eine semantisch reichhaltige Metadaten-Beschreibung erfüllt werden (z. B. formale, semantische Angaben zu Beziehungen zwischen Konzepten). Jedoch sind einige wesentliche Konstrukte zur Festlegung von (semantischen) Restriktionen innerhalb eines Vokabulars nicht verfügbar. Die Sprachstufenhierarchie aufsteigend werden daher im nächsten Abschnitt *Ontologien* und *Ontologiesprachen* eingeführt.

6.3 Ontologien

Der Begriff *Ontologie* entstammt der Philosophie und deklariert in seiner originären Form die Lehre des Seins. Er bezieht sich auf die Wissenschaft, welche die unterschiedlichen Arten von Entitäten der Realwelt und deren Beziehungen beschreibt [Grub93], [Hess02], [SmWG04]. Seit einigen Jahren sind Ontologien auch in der (Wirtschafts-)Informatik ein wesentlicher Bestandteil der Forschung. Das Verständnis von Ontologien aus dem Blickwinkel der Informatik sowie deren Einsatzmöglichkeiten in dieser Disziplin sind Schwerpunkte der Erörterung in diesem Abschnitt. Darin einbezogen ist die Vorstellung gängiger Ontologiesprachen.

6.3.1 Begriffsbestimmung und Abgrenzung

Ontologien sind in den Bereichen der (Wirtschafts-)Informatik von Bedeutung, die sich mit der Repräsentation und Kommunikation von Wissen befassen oder Beiträge zu diesen Themen liefern.⁷⁷ Zu nennen sind hier u. a. die Künstliche Intelligenz, Datenbankanwendungen und im weiteren Sinne Systeme, die der Verarbeitung von Informationen dienen [Hess02].

Um sich gespeichertes Wissen zunutze zu machen, kann der Mensch auf Grund- und Kontextwissen eines spezifischen Wissensgebiets zurückgreifen. Anhand von Lehrbüchern, Lexika, Regelwerken o. Ä. werden exakte Anweisungen zur korrekten Interpretation fachspezifischer Inhalte gegeben. Maschinen (z. B. in Form von Software-Agenten) benötigen dagegen für Such-, Kommunikations- und Entscheidungsaufgaben über gespeichertes Wissen eine Repräsentation der zugrunde liegenden Begriffe und deren Zusammenhänge. Für derartige Anforderungen greift die (Wirtschafts-)Informatik auf Ontologien zurück [Hess02].

DEFINITION

Eine in der Literatur geläufige Definition des Ontologie-Begriffs wurde von GRUBER geprägt:

An ontology is an explicit specification of a conceptualization [Grub93].

⁷⁷ Für eine ergänzende Diskussion wissenschafts- und erkenntnisorientierter Probleme beim Umgang mit Ontologien mit Blick auf die Wirtschaftsinformatik wird auf [ScZe99] verwiesen.

Eine Ontologie erfordert demnach eine *Konzeptionalisierung* des Anwendungsbereichs, den sie beschreiben soll. Die Konzepte müssen dabei *explizit spezifiziert* sein [Grub93], [Guar98], [LeNa07a, S. 272 ff.]:

- Ein Begriff bzw. etwas weiter gefasst ein Konzept steht für ein bestimmtes Ding oder eine Klasse von Dingen. Ein Konzept existiert zunächst nur in der subjektiven Vorstellung einer Person und ist somit individuell spezifiziert. Zur Kommunikation von Konzepten wird auf Symbole wie Bilder oder Worte zurückgegriffen (vgl. Semiotisches Dreieck; Abschnitt 5.1). Konzepte können aufgrund ihres individuellen Charakters jedoch nicht immer eindeutig kommuniziert werden. Bei der *Konzeptualisierung* gilt es, alle relevanten Konzepte einer Domäne zu erfassen und Einigkeit über deren Interpretation unter den Fachverantwortlichen zu erzielen. Sprachliche Defekte wie Synonyme und Homonyme (vgl. Abschnitt 5.2.2) sind dabei aufzulösen, sodass alle Konzepte eindeutig bestimmt sind. Hier kann bspw. die *Methode zur Rekonstruktion von Fachbegriffen* dienlich sein (vgl. Abschnitt 5.2.4). Durch die Konzeptualisierung wird das Vokabular eines spezifischen Anwendungsbereichs determiniert.
- Die *Spezifikation* dient der Modellierung von Eigenschaften (z. B. Attributen, Integritätsbedingungen) und Beziehungen der Konzepte (Intension nach innen und nach außen; vgl. Abschnitt 5.2.1). Optional können Regeln definiert werden, welche die Ableitung nicht explizit angegebener Eigenschaften von Konzepten aus bestehenden Informationen erlauben. Dabei ist zu unterscheiden, ob eine Ontologie informell beschrieben, und damit wesentlich auf die Benutzung durch den Menschen ausgerichtet ist, oder formal ausgelegt ist und die automatische Ableitung von Wissen ermöglicht. Eine maschinelle Verarbeitung von Ontologien erfordert ein formales, maschinenlesbares Modell der Ontologie. Unterschiedliche Maschinen ziehen auf Basis des gleichen formalen Modells jeweils dieselben Schlüsse. Abhängig von der angestrebten Verwendung sind zur Erstellung einer Ontologie entsprechend mächtige Ontologiesprachen (z. B. Thesaurus, Taxonomie, Beschreibungslogik) zu wählen.

Die Ontologie-Definition von GRUBER lässt Raum für vielfältige Auslegungen dieses Terminus. Da in der Informatik zudem keine exakte sowie global gültige Definition des Ontologie-Begriffs existiert, hat sich ein weites Spektrum an Informationsartefakten herauskristallisiert, die als Ontologie deklariert werden [SmWe01], [LeNa07a, S. 272].

Abbildung 6-5 offeriert eine Einordnung ausgewählter Typen von Ontologien nach dem Grad ihrer semantischen Ausdrucksstärke. Mit diesem steigt ebenfalls der Komplexitätsgrad. Die Ontologie-Typen sind in Bezug auf automatische Schlussfolgerung klassifiziert.

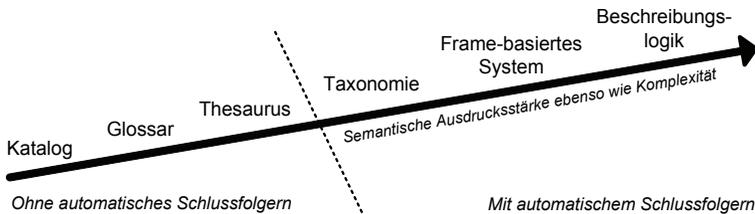


Abbildung 6-5: Typen von Ontologien (in Anlehnung an [SmWe01])

Eine – im Hinblick auf ihre semantische Reichhaltigkeit – primitive Variante einer Ontologie stellt der *Katalog* dar. Kataloge sind einer bestimmten Systematik folgende Listen oder Verzeichnisse. *Glossare* erweitern die Grundidee eines Katalogs um natürlichsprachliche Beschreibungen und Strukturierungen von Begriffen. Die Angabe von (Hierarchie-)Beziehungen zwischen Begriffen ist mit *Thesauri* möglich. Thesauri stellen hierzu vordefinierte Relationen bereit. *Taxonomien* beinhalten, in Erweiterung zum Thesaurus, das Konzept der Vererbung. Eigenschaften einer bestimmten Klasse können somit von einer spezifischeren geerbt werden. *Frame-basierte Systeme* ergänzen Taxonomien um die Möglichkeit, Relationen und Restriktionen in der Form, welche Klassen auf welche Weise miteinander in Beziehung gesetzt werden dürfen, zu definieren. Die ausdrucksmächtigste und zugleich komplexeste Form von Ontologien wird unter dem Terminus *Beschreibungslogik* zusammengefasst. Diese stützt sich u. a. auf die Logik erster Stufe (First-Order-Logic)⁷⁸. Die vorgestellten Ontologie-Typen genügen alle der Ontologie-Definition von GRUBER [SmWe01], [Zieg07, S. 173].

⁷⁸ Eine Einführung in die *Logik erster Stufe* bietet z. B. [RuNo04, S. 303 ff.]. Die Logik erster Stufe ist ausdrucksstark genug, um einen Großteil unseres Allgemeinwissens zu repräsentieren [RuNo04, S. 303].

ONTOLOGIE-KLASSIFIZIERUNG NACH GUARINO

Ontologien können auch nach dem Grad ihrer Allgemeingültigkeit bzw. ihrem Bezug zu einer spezifischen Aufgabe klassifiziert werden. GUARINO unterscheidet hierbei drei Ebenen [Guar98], [GoFC04, S. 27]:

- *Top-Level-Ontologien* definieren allgemeine Konzepte, die unabhängig von einem spezifischen Problem oder Anwendungsbereich sind. Sie dienen als Grundlage für die Ableitung spezieller Ontologien.
- *Domain- bzw. Task-Ontologien* beschreiben ein Vokabular in Bezug auf eine Domäne oder Aufgabe. Sie spezialisieren die in der Top-Level-Ontologie definierten Konzepte.
- *Applikationsontologien* beschreiben Konzepte, die sowohl auf einer determinierten Domäne als auch auf einer spezifischen Aufgabe beruhen. Sie sind zumeist Konkretisierungen entsprechender Domain- und Task-Ontologien.

Umfangreiche Ontologien können bspw. gemäß diesen Ebenen partitioniert werden. Die Komplexität einer Ontologie wird somit beherrschbar.

ENTWICKLUNGSKRITERIEN FÜR ONTOLOGIEN

Die Entwicklung einer Ontologie kann nach verschiedenartigen Ansätzen erfolgen (z. B. *induktiver vs. deduktiver Ansatz*⁷⁹). Diese werden u. a. im Rahmen des *Ontological Engineering* untersucht und vorgeschlagen. Analog zum *Software-Engineering*⁸⁰, wird beim *Ontological Engineering* das Ziel verfolgt, die Entwicklung und den Einsatz einer Ontologie über ihren gesamten Lebenszyklus effektiv zu unterstützen. Der Wert einer Ontologie ist von der Anerkennung und Zustimmung durch ihre Zielgruppe abhängig [GrLe02], [Hess02].

⁷⁹ Beim *induktiven Ansatz* werden größere Ontologien durch Verbindung mehrerer kleinerer gebildet. Beim *deduktiven Ansatz* werden zunächst allgemeine Konzepte und Regeln durch ein Gremium festgelegt, anschließend erfolgt eine Spezialisierung der Teilbereiche [Hess02].

⁸⁰ *Software Engineering* oder auch Softwaretechnik ist das Teilgebiet der Informatik, das sich mit der Bereitstellung und systematischen Verwendung von Methoden, Techniken und Werkzeugen für die Herstellung, den Betrieb und die Wartung von Software auseinandersetzt [ScWe07, S. 220 ff.].

Für die Entwicklung und Evaluierung von Ontologien schlägt GRUBER die Einhaltung folgender fünf Kriterien vor [Grub95]:

- *Klarheit*: Eine Ontologie soll auf effektive Weise die intendierte Bedeutung der beschriebenen Konzepte wiedergeben. Die Formalisierung der Konzepte ist diesbezüglich ein geeignetes Hilfsmittel. Die Definitionen sollen zudem in einer natürlichen Sprache dokumentiert werden.
- *Kohärenz*: Eine Ontologie soll Schlussfolgerungen ermöglichen, die mit den zugrunde liegenden Definitionen konsistent sind oder zumindest aus zueinander widerspruchsfreien Axiomen bestehen.
- *Erweiterbarkeit*: Die Einsatzbereiche eines gemeinsamen Vokabulars sollten beim Entwurf einer Ontologie antizipiert werden. Eine Ontologie ist dennoch so zu gestalten, dass sie gleich bleibend erweitert oder spezialisiert werden kann, d. h. eine Erweiterung darf keine Änderung bestehender Definitionen mit sich bringen.
- *Minimale Kodierungsabhängigkeit*: Die Konzeptualisierung ist auf Wissensebene (knowledge level) unabhängig von einer bestimmten Ebene der Symbolverarbeitung (symbol level)⁸¹ zu spezifizieren.
- *Minimale ontologische Bindung*: Eine Ontologie soll lediglich die Konzepte definieren, die essentiell für den Wissensaustausch in einer Domäne sind. Somit bleiben für Interessensgruppen ausreichend Freiheitsgrade zur Spezialisierung und Instanziierung der Ontologie.⁸²

Beim Aufbau einer Ontologie gilt es, einen Kompromiss bei der Ausgestaltung und Gewichtung dieser Kriterien zu finden. Eine Diskussion dieser Obliegenheit erfolgt in [Grub95].

⁸¹ Die Begriffe *knowledge level* und *symbol level* werden in [Newe81] detailliert vorgestellt.

⁸² Eine Ontologie verfolgt – im Gegensatz zu einer Wissensdatenbank – das Ziel, ein Vokabular zur Kommunikation innerhalb bzw. über eine Domäne zu beschreiben. Eine Wissensdatenbank wird eingesetzt, um Probleme zu lösen oder arbiträre Fragen zu einer Domäne zu beantworten [Grub95].

ANWENDUNGSFELDER VON ONTOLOGIEN IN DER INFORMATIK

Für den Einsatz von Ontologien in der (Wirtschafts-)Informatik haben GRUNINGER und LEE folgende Anwendungsfelder identifiziert [GrLe02], [Hess02], [LeNa07a, S. 275 f.]:

- *Kommunikation*: Eine Ontologie kann die Reichhaltigkeit der Kommunikation zwischen Menschen, zwischen Maschinen sowie zwischen Mensch und Maschine erhöhen. Kommunizieren bspw. zwei Software-Agenten miteinander, so kann die erforderliche Interpretationsvorschrift für die ausgetauschten Daten aus einer für beide Agenten zugänglichen Ontologie bezogen werden.
- *Automatisches Schließen*: Spezielle Programme, so genannte *Reasoner*⁸³, können entlang der in einer Ontologie definierten Beziehungen logische Schlüsse ziehen.
- *Organisation, Repräsentation und Wiederverwendung von Wissen*: Ontologien können als Referenzmodelle⁸⁴ eingesetzt werden, da sie alle Symbole eindeutig auf formal definierte Konzepte abbilden. Sie können, als globale Schemata verstanden, auch bei Integrationsvorhaben Anwendung finden.

Der Einsatz von Ontologien erscheint demzufolge auch zur Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen geeignet. Hervorzuheben ist neben der besonderen Reichhaltigkeit der Mensch-Maschine-Kommunikation die formale Repräsentation von Konzepten und die Möglichkeit der Wiederverwendung von Wissen z. B. durch die Nutzung einer Ontologie als Referenzmodell.

Die Verwendung von Ontologien im DWH-Umfeld wurde in der Forschung bisher kaum thematisiert. Es finden sich in diesem Kontext nur vereinzelt Beiträge (z. B. [CaNL06], [CaZL06]).

⁸³ Einen Überblick aktueller *Reasoner* für Beschreibungslogiken offeriert z. B. <http://www.cs.man.ac.uk/~sattler/reasoners.html> (Abruf am 10.02.2008).

⁸⁴ *Referenzmodelle* formulieren Sollempfehlungen für eine Klasse von Anwendungsgebieten und unterstützen den Modellierungsprozess, indem sie spezifische Ausgangslösungen anbieten [Holt03, S.47], [BDKK02, S. 25 f.].

6.3.2 Wissensrepräsentationssprachen

Zur Spezifikation von Ontologien werden so genannte *Wissensrepräsentationssprachen* eingesetzt. Wissensrepräsentationssprachen sind formale Sprachen zur Darstellung von Wissensinhalten [LeNa07a, S. 267, S. 282 ff.].

Neben *Frame-basierten Ansätzen* (z. B. F-Logic [KiLW95], [AnLa04]), *prädikatenlogischen Sprachen* (z. B. Knowledge Interchange Format (KIF)⁸⁵ [GeFi92], CycL⁸⁶ [LeGu90]) und *Beschreibungslogiken* (z. B. SHIQ [HoST99]) sind insbesondere (*Web-*)*Ontologiesprachen* wie DAML+OIL und OWL gebräuchliche Sprachen zur formalen Wissensrepräsentation. Letztere werden anschließend beleuchtet.

ALLGEMEINE ANFORDERUNGEN AN ONTOLOGIESPRACHEN

Eine Ontologiesprache sollte eine wohl definierte Syntax und Semantik besitzen, effizientes Schlussfolgern ermöglichen und zudem eine für den jeweiligen Anwendungszweck hinreichende Ausdrucksstärke besitzen. Eine *wohl definierte Syntax* ist eine notwendige Bedingung für maschinenverarbeitbare Informationen. *Formale Semantik* dient der präzisen Beschreibung von Wissen und stellt eine Voraussetzung für *Schlussfolgerungen* (Inferenz) dar. Die Möglichkeit der Inferenz erlaubt darüber hinaus die Konsistenzprüfung der Ontologie, das Aufdecken von nicht explizit definierten Beziehungen zwischen Klassen sowie die automatische Zuordnung von Instanzen zu Klassen. Hierdurch wird auch der Entwurf umfangreicher Ontologien sowie die Integration von Ontologien aus unterschiedlichen Quellen unterstützt [AnHa04a, S. 68 f.], [Zieg07, S. 173].

DARPA AGENT MARKUP LANGUAGE + ONTOLOGY INFERENCE LAYER

Das W3C hat in einer Vielzahl von charakteristischen Anwendungsfällen für Ontologien im WWW die Notwendigkeit einer, im Vergleich zu RDF und RDF-Schema (vgl. Abschnitt 6.2.3), ausdrucksmächtigeren Modellierungssprache aufgezeigt. Aus diesem Grund wurde die Ontologiesprache *DAML+OIL* (DARPA Agent Markup Language + Ontology Inference Layer) definiert. Der Name setzt sich aus dem amerikanischen Vor-

⁸⁵ Die *KIF-Spezifikation* kann unter <http://logic.stanford.edu/kif/specification.html> eingesehen werden (Abruf am 10.06.2008).

⁸⁶ Die *CycL-Syntax* ist unter <http://www.cyc.com/cycdoc/ref/cycl-syntax.html> detailliert erläutert (Abruf am 10.06.2008).

schlag *DAML-ONT*⁸⁷ [GFSH03] und der in Europa entwickelten Sprache *OIL*⁸⁸ [GoFC04, S. 258 ff.] zusammen.

DAML+OIL wurde von der W3C-Arbeitsgruppe für Web-Ontologien als Ausgangspunkt für die Entwicklung der *Web Ontology Language* (OWL) genutzt [Lacy05, S. 16], [AnHa04a, S. 67], [Powe03, S. 8]. Da wesentliche Bestandteile von DAML+OIL in OWL eingegangen sind, wird an dieser Stelle auf eine eingehende Diskussion von DAML+OIL verzichtet und stattdessen auf die Literatur sowie Online-Referenz verwiesen [GoFC04, S. 61 ff., S. 264 ff.], [CHH+01]. Eine grundlegende Einführung in OWL erfolgt im nachstehenden Abschnitt.

6.3.3 Web Ontology Language (OWL)

Die *Web Ontology Language* (OWL) ist eine vom W3C standardisierte Ontologiesprache und gehört zu den am weitesten verbreiteten Sprachen, um Ontologien zu modellieren. OWL basiert auf der Prädikatenlogik erster Stufe.⁸⁹ Sie ist von der Ontologiesprache DAML+OIL abgeleitet und stellt eine Vokabular-Erweiterung von RDF dar [BHH+04], [GoFC04, S. 65], [Lacy05, S. 133 ff.], [HKRS08, S. 125].

Ebenso wie RDF liegt auch OWL das OWA-Paradigma (vgl. Abschnitt 6.2.3) zugrunde [SmWG04], [HKRS08, S. 145].

OWL-SPRACHSTUFEN

Bei der Spezifikation einer Ontologie besteht ein Dilemma zwischen der Ausdrucksstärke und der Komplexität einer Sprache. Die Entwicklung einer detaillierten und exakten Ontologie erfordert eine möglichst reichhaltige Sprache. Die bei einer Anfrage durchzuführenden Inferenzoperationen erreichen jedoch, mit zunehmender Mächtigkeit der eingesetzten Sprache, schnell eine Komplexität, die bis zur Unentscheidbarkeit von Resolutionsproblemen führt. OWL stellt aus diesem Grund drei (Teil-)Sprachen unterschiedlicher Ausdrucksstärke bereit [AnHa04b, S. 113 ff.], [LeNa07a, S. 268, S. 311 f.], [StHa05b, S. 47 ff.], [HKRS08, S. 151 ff.].

⁸⁷ *DARPA Agent Markup Language - Ontology* (DAML-ONT): Informationen zur initialen Spezifikation sind unter <http://www.daml.org/2000/10/daml-ont.html> nachzulesen (Abruf am 10.06.2008).

⁸⁸ *Ontology Inference Layer* (OIL): Weiterführende Beschreibungen sind über die Webadresse <http://www.ontoknowledge.org/oil> verfügbar (Abruf am 10.06.2008).

⁸⁹ Eine ausführliche Darstellung der prädikatenlogischen Semantik von OWL bietet u. a. [HKRS08, S. 163 ff.].

Die Sprachstufen der OWL sind wie folgt zu charakterisieren:

- *OWL Lite* besitzt im Vergleich zu den anderen OWL-Teilsprachen die geringste semantische Reichhaltigkeit. Sie ist für hierarchische Klassifizierungen und einfache Restriktionen ausgelegt (die Angabe von Kardinalitäten ist bspw. nur mit den Werten 0 oder 1 zulässig). Da weniger Sprachelemente als bei den anderen OWL-Ebenen zur Verfügung stehen, sind effiziente Inferenzalgorithmen realisierbar. Zudem ist eine Werkzeugunterstützung verhältnismäßig leicht zu implementieren.
- *OWL DL* (description logic) bietet hohe Ausdrucksstärke bei Aufrechterhaltung der Berechenbarkeit und Entscheidbarkeit. Damit ein Algorithmus für OWL DL stets garantiert terminiert, liegen dieser OWL-Teilsprache einige Restriktionen zugrunde. Hervorzuheben ist dabei die explizite Typentrennung, d. h. eine Resource ist entweder eine Klasse oder eine Instanz. Die weiteren Restriktionen sind z. B. in [HKRS08, S. 153] erläutert. OWL DL stellt aufgrund ihrer Ausdrucksstärke bei garantierter Entscheidbarkeit die wichtigste Teilsprache von OWL dar [HKRS08, S. 163].
- *OWL Full* ist die mächtigste Variante von OWL. Es dürfen alle OWL-Sprachelemente gemeinsam mit RDF(S)-Sprachelementen uneingeschränkt benutzt werden. Zudem ist die explizite Typentrennung (vgl. OWL DL) nicht vorgeschrieben, sodass bspw. Klassen Instanzen anderer Klassen sein können. Das resultierende OWL-Full-Dokument muss lediglich ein gültiges RDF sein. OWL Full ist unentscheidbar, da Paradoxien formuliert werden können.

Anhand der auf den jeweiligen OWL-Ebenen angebotenen Sprachkonstrukte visualisiert die nachstehende Abbildung die Reichhaltigkeit von OWL Lite, OWL DL und OWL Full (in Anlehnung an [GoFC04, S. 66], [StHa05b, S. 49 ff.], [SmWG04], [AnHa04b, S. 127 ff.]):

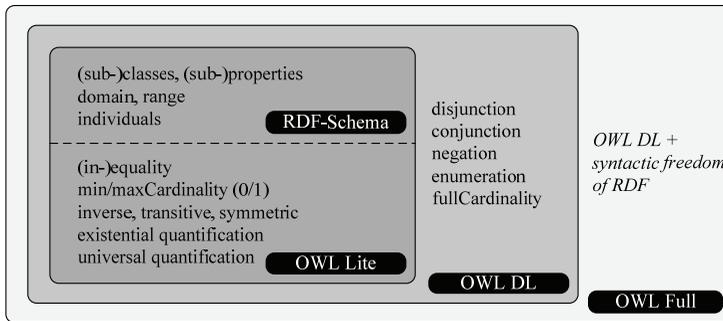


Abbildung 6-6: Sprachstufen der OWL

OWL Lite ist eine echte Teilsprache von OWL DL. Diese ist wiederum eine echte Teilsprache von OWL Full. Die in OWL Lite bereitgestellten Sprachkonstrukte (z. B. für Ähnlichkeit zweier Konzepte oder inverse, transitive, symmetrische Eigenschaft einer Beziehung) werden folglich auf Ebene der OWL DL um weitere Ausdrücke ergänzt (z. B. Disjunktion und Konjunktion, Negation, vollständige Kardinalitätsangaben).⁹⁰ In Bezug auf Syntax und Semantik besteht daher folgende Teilmengenbeziehung [StHa05b, S. 48], [HKRS08, S. 126]:

$$OWL\ Lite \subset OWL\ DL \subset OWL\ Full$$

Jede syntaktisch korrekte OWL-Lite-Aussage ist ebenfalls eine korrekte OWL-DL-Aussage. Jede gültige OWL-Lite-Schlussfolgerung ist auch eine gültige OWL-DL-Schlussfolgerung. Diese Relationen gelten analog für die Beziehung zwischen OWL DL und OWL Full [SmWG04].

Zu beachten ist, dass OWL Lite und OWL DL nicht zu RDF abwärtskompatibel sind, d. h. es können zulässige RDF-Dokumente formuliert werden, die keine zulässigen OWL-DL- oder OWL-Lite-Dokumente sind [AnHa04b, S. 70 f.], [LeNa07a, S. 312].

⁹⁰ Eine detaillierte Erläuterung der auf den jeweiligen OWL-Ebenen angebotenen Sprachkonstrukte bieten u. a. [HKRS08, S. 126 ff.] und [StHa05b, S. 49 ff.].

VERSIONIERUNG EINER OWL-ONTOLOGIE

Allen OWL-Sprachstufen gemein sind die Möglichkeiten zur Versionierung einer Ontologie. Hierzu stehen Sprachelemente für Angaben über die vorliegende Version (*owl:versionInfo*) und die Vorgängerversion (*owl:priorVersion*), die In- und Rückwärtskompatibilität zu früheren Versionen (*owl:incompatibleWith*, *owl:backwardCompatibleWith*) sowie zur Kennzeichnung nicht weiter unterstützter Klassen und Rollen bereit (*owl:DeprecatedClass*, *owl:DeprecatedProperty*) [SmWG04].

AUFBAU EINES OWL-DOKUMENTS (OWL-RDF-SYNTAX)

Für den Umgang mit OWL-Dokumenten wurde zum einen die für den Menschen leicht lesbare *abstrakte OWL-Syntax*⁹¹ (nicht für OWL Full verfügbar) und zum anderen die *OWL-RDF-Syntax* entwickelt. Letztere basiert auf RDF/XML (vgl. Abschnitt 6.2.3). Für alle OWL-Aussagen ist eine RDF-Syntax definiert. Die OWL-RDF-Syntax stellt das Standardformat für den Austausch von OWL-Dokumenten dar [SmWG04], [HKRS08, S. 126]. Im Folgenden wird daher stets die OWL-RDF-Syntax verwendet.

Als RDF-Dokument besitzt jedes OWL-Dokument ein Wurzelement (*<rdf:RDF>*). Analog zu RDF werden in diesem die benötigten Namensräume spezifiziert. Des Weiteren kann der Kopf einer OWL-Ontologie (*<owl:Ontology rdf:about="">*) Annotationen, Versionsinformationen und Import-Anweisungen zur Einbindung anderer Ontologien enthalten.

Die grundlegenden Bausteine von OWL sind – analog zu RDF/RDFS – *Klassen* (*owl:Class*), *Eigenschaftstypen* (in OWL auch als *Rollen* bezeichnet) und *Individuen*. Abstrakte Rollen verbinden Individuen mit Individuen (*owl:ObjectProperty*), konkrete Rollen koppeln Individuen mit Datenwerten (*owl:DatatypeProperty*). Domänen- und Wertebereiche für Rollen können in derselben Weise wie bei RDFS eingeschränkt werden (*rdfs:domain*, *rdfs:range*). Klassen sowie abstrakte und konkrete Rollen der OWL stellen Unterklassen zu ihrem jeweiligen Pendant im RDF/RDFS dar, d. h. *owl:Class* ist eine Subklasse zu *rdfs:Class* und *owl:ObjectProperty* bzw. *owl:DatatypeProperty* stellen Subklassen zu *rdf:Property* dar [BHH+04], [AnHa04b, S. 114 f.].

⁹¹ Für eine ausführliche Erläuterung der *abstrakten OWL-Syntax* wird auf <http://www.w3.org/TR/owl-semantics/syntax.html> (Abruf am 10.06.2008) verwiesen.

Abhängig von der gewählten OWL-Sprachstufe stehen verschiedene Sprachkonstrukte zur semantischen Spezifikation der Beziehungen zwischen Klassen, zwischen Individuen oder zwischen Klassen und Individuen zur Verfügung (vgl. Abbildung 6-6).

Ein gültiges OWL-Dokument endet mit dem schließenden Wurzelement `</rdf:RDF>`. Für detaillierte Erläuterungen zum Aufbau einer OWL-Ontologie wird auf [BHH+04], [SmWG04], [GoFC04, S. 275 ff.] und [HKRS08, S. 128 ff.] verwiesen.

ANFRAGESPRACHEN FÜR OWL

Die im Rahmen der Diskussion von RDF genannten Anfragesprachen sind nur bis zu einem gewissen Grad semantischer Ausdrucksstärke für OWL einsetzbar. Eine Erweiterung von SPARQL für OWL ist gegenwärtig noch Gegenstand vieler Diskussionen. Eine mögliche Modifikation von SPARQL hinsichtlich der Verwendung mit OWL DL ist in [SiPa07] erläutert. Ein Vorschlag einer eigenständigen Anfragesprache für OWL, die für die Ergebnisermittlung auch Inferenzmethoden mit einbezieht, wurde mit der *OWL Query Language* (OWL-QL) unterbreitet. Eine Einführung in die OWL-QL findet sich u. a. in [FiHH03].

WERKZEUGE FÜR OWL-ONTOLOGIEN

Zu den bekanntesten Editoren für Ontologien zählt *Protégé*⁹². Hierbei handelt es sich um ein frei verfügbares, java-basiertes Werkzeug zur Erstellung und Bearbeitung von (OWL-)Ontologien. Daneben sind die Editoren *SMORE – Semantic Markup, Ontology, and RDF Editor*⁹³ und *SWOOP*⁹⁴ zu nennen [Powe03, S. 248 ff.], [HKRS08, S. 155]. Insbesondere für Protégé existieren Erweiterungen, um eine OWL-Ontologie auf unterschiedliche Weise zu visualisieren (z. B. *OntoViz*, *Jambalaya*, *TGVizTab*). Eine Aufstellung und Erläuterung von Techniken zur Ontologie-Darstellung bietet [KHL+07].

Bei den Inferenzmaschinen für OWL ist insbesondere das Open-Source-Produkt *Pellet*⁹⁵ hervorzuheben, da dieses OWL DL vollständig unterstützt. Weitere bekannte Inferenzmaschinen sind *FaCT++*⁹⁶ und *RacerPro*⁹⁷ [HKRS08, S. 155 f.].

⁹² Vgl. <http://protege.stanford.edu> (Abruf am 10.06.2008).

⁹³ Vgl. <http://www.mindswap.org/2005/SMORE> (Abruf am 10.06.2008).

⁹⁴ Vgl. <http://www.mindswap.org/2004/SWOOP> (Abruf am 10.06.2008).

⁹⁵ Vgl. <http://pellet.owldl.com> (Abruf am 10.06.2008).

⁹⁶ Vgl. <http://owl.man.ac.uk/factplusplus> (Abruf am 10.06.2008).

⁹⁷ Vgl. <http://www.racer-systems.com> (Abruf am 10.06.2008).

DISKUSSION DER OWL

OWL bietet vorbehaltlich alle Eigenschaften, um die Anforderungen an eine semantisch reichhaltige Metadaten-Beschreibung für DWH-Systeme realisieren zu können. Sprachkonstrukte, die für eine aussagekräftige Beschreibung von Begriffssystemen essentiell sind und bei RDF fehlen (vgl. Abschnitt 6.2.3), werden durch OWL standardmäßig angeboten. Dabei erscheint insbesondere OWL DL, durch seine Mächtigkeit bei gleichzeitiger Entscheidbarkeit, für den vorliegenden Anwendungsbereich als geeignet. Auch die bestehende Werkzeug-Unterstützung in Form von frei verfügbaren Editoren und Referenzmaschinen spricht für den Einsatz von OWL.

6.4 Zusammenfassung

Die Einführung und Erläuterung der mit dem Semantic Web verwobenen Technologie RDF sowie der Ontologiesprache OWL bildeten den Schwerpunkt dieses Kapitels. Zudem wurden RDF und OWL jeweils auf ihre semantische Ausdrucksstärke hin untersucht und im Hinblick auf die Erfüllung der Anforderungen an eine semantisch reichhaltige Beschreibung von DWH-Metadaten beleuchtet.

Die Untersuchungsergebnisse legen nahe, die beiden genannten Konzepte des Semantic Web zur Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen einzusetzen. Insbesondere die Mächtigkeit der Ontologiesprache OWL gibt Grund zur Annahme, dass durch deren Einsatz die Ansprüche an eine semantisch reichhaltige Metadaten-Beschreibung erfüllt werden können.

7 Realisierung einer semantisch reichhaltigen Beschreibung von Data-Warehouse-Metadaten

Im vorhergehenden Kapitel wurden insbesondere die Semantic-Web-Technologien RDF und OWL auf ihre semantische Mächtigkeit hin untersucht. RDF ist demnach zur Erfüllung grundlegender Ansprüche an eine semantische Beschreibung von DWH-Metadaten geeignet. Eine vollständige Erfüllung der in Abschnitt 5.5 beschriebenen Anforderungen an eine semantisch reichhaltige Metadaten-Beschreibung von DWH-Systemen erscheint mit OWL realisierbar.

Nachstehend wird mithilfe von RDF-Schema (RDFS) ein umfassendes Vokabular für die semantische Beschreibung von DWH-Metadaten erarbeitet und dessen Anwendbarkeit demonstriert. Im Hinblick darauf erfolgt einleitend eine Untersuchung der Potentiale von RDF/RDFS zur Realisierung der spezifischen Anforderungen dieses Einsatzgebiets. Auf dem entwickelten RDF-Vokabular aufbauend wird anschließend eine OWL-Ontologie zur semantisch reichhaltigen Metadaten-Beschreibung für DWH-Systeme konstruiert.

7.1 Potentiale von RDF und RDF-Schema zur semantischen Beschreibung von Data-Warehouse-Metadaten

RDF offeriert ein fachlich neutrales Datenmodell. Um spezifische Vokabulare definieren zu können, ist ein gemeinsames Schema erforderlich. RDFS stellt hierfür ein domänen-neutrales Regelsystem bereit (vgl. Abschnitt 6.2.3).

Nachfolgend werden die von RDF/RDFS offerierten Konstrukte und Beschreibungselemente sowohl im Hinblick auf die Abbildung der bei der *verwendungszweckgerichteten Metadatenkategorisierung* identifizierten Kategorien und Metadaten-Elemente (vgl. Abschnitt 3.3) als auch hinsichtlich der Anforderungserfüllung an eine semantisch reichhaltige Metadaten-Beschreibung (vgl. Abschnitt 5.5) analysiert. Die verwendungszweckgerichtete Metadatenkategorisierung wurde als Ausgangspunkt gewählt, da anhand dieser eine eingehende Metadaten-Beschreibung für DWH-Systeme erzielt werden kann. Dies soll dem Formalziel der *Vollständigkeit* des Metadaten-Vokabulars für DWH-Systeme Genüge leisten.

Die anschließenden Untersuchungen konzentrieren sich zunächst auf die Behandlung *allgemeiner Beschreibungskonzepte und -elemente* (Abschnitt 7.1.1). Hier fließen insbesondere Kriterien der Metadaten-Kategorien Terminologie, Datenstruktur und -bedeutung sowie der Metadatenhistorisierung ein. Anschließend werden mit Bezug auf die Metadaten-Kategorie Datenanalyse grundlegende Elemente zur Beschreibung eines DWH-Systems (*DWH-Kernelemente*) und deren Abbildung in einem RDFS behandelt (Abschnitt 7.1.2). Die Metadaten-Kategorie Organisationsbezug prägt die Ausführungen des nachfolgenden Abschnitts *DWH-Organisationselemente* (Abschnitt 7.1.3). Die Möglichkeiten, Kontextinformationen durch ein RDFS zu erfassen, werden im Abschnitt *DWH-Kontextelemente* (Abschnitt 7.1.4) erörtert. Hierfür gelten die Ergebnisse aus Abschnitt 4.3 als maßgebende Grundlage. Abgerundet werden die Ausführungen mit der Untersuchung, auf welche Weise *DWH-Systemelemente* mithilfe eines RDFS dargestellt werden können (Abschnitt 7.1.5). Erkenntnisse aus den Metadaten-Kategorien Systembezug und Datentransformation bilden hierbei die Vorgabe.

Bei den Untersuchungsschritten zum Einsatz von RDF für eine semantische Beschreibung von DWH-Metadaten sind jeweils auch konkrete Vorschläge für eine Ausgestaltung des RDF-Vokabulars zu erarbeiten (vgl. auch [HaWe08]). Soweit möglich soll zudem auf standardisierte Beschreibungsfelder, wie sie z. B. der DC-Standard anbietet, zurückgegriffen werden.

7.1.1 Allgemeine Beschreibungskonzepte und -elemente

Zu den allgemeinen Beschreibungskonzepten für das zu definierende RDF-Vokabular für DWH-Metadaten gehören Elemente der Metadaten-Kategorien Terminologie, Datenstruktur und -bedeutung sowie Metadatenhistorie.

TERMINOLOGISCHE MERKMALE

Um eine Vereinheitlichung der Terminologie in betrieblichen Begriffssystemen zu erzielen, sind die verwendeten Fachbegriffe zu ermitteln und anschließend deren Benennung und Bedeutung zu normieren (vgl. Abschnitt 5.2). Selbiges ist auf Schemaebene auf die zur DWH-Modellierung eingesetzten Elemente zu übertragen. Die Bezeichnung, Bedeutung und Beziehung der Schemaelemente sind eindeutig und semantisch klar zu definieren. Nachfolgend wird aufgezeigt, wie mithilfe von RDF/RDFS diese Anforderungen erfüllt werden können.

- RDF liegt das URI-Konzept zugrunde, d. h. jeder Begriff wird mittels eines Identifiers (ID) oder URI gekennzeichnet und ist somit eindeutig bestimmt. Die Benennung eines Begriffs kann anhand des durch RDFS bereitgestellten Elements *rdfs:label*, die Begriffsdefinition mittels *rdfs:comment* oder durch das DC-Beschreibungselement *dc:description* erfolgen. Die genannten Elemente zur semantischen Charakterisierung sind sowohl auf Klassen als auch auf Eigenschaftstypen anwendbar.
- Begriffsbeziehungen können im RDFS zum einen über hierarchische Unterklassebeziehungen (*rdfs:subClassOf*) und zum anderen über individuell zu definierende, thematische Eigenschaftstypen erfolgen (*rdf:Property*). Mithilfe des RDFS-Elements *rdfs:seeAlso* kann zudem auf andere Begriffe mit ähnlichen oder weiterführenden Informationen oder Definitionen verwiesen werden.
- Die RDF-Technologie bietet darüber hinaus Unterstützung, um sprachliche Defekte (vgl. Abschnitt 5.2.2) aufzudecken. Dies gilt insbesondere für Synonyme und Homonyme. Synonyme können anhand der Ähnlichkeit oder ggf. Gleichheit der bei den Schemaelementen hinterlegten Beschreibungen erkannt werden. Die Abbildung von Homonymen wird in RDF bzw. RDFS generell durch das Konzept der Eindeutigkeit eines Begriffs (Begriffsidentität) verhindert. Die Möglichkeiten Begriffe zu identifizieren, exakt zu beschreiben und ihre Beziehungen zu anderen Begriffen semantisch darzulegen, fördern zudem die Lokalisierung weiterer Arten von sprachlichen Defekten (z. B. Äquipollenzen, falsche Bezeichner). Die Auflösung sprachlicher Defekte ist in Abstimmung mit den jeweiligen Fachexperten durchzuführen (vgl. Abschnitt 5.2.3 und Abschnitt 5.2.4).

Tabelle 5 bietet einen Überblick über die Elemente der Metadaten-Kategorie Terminologie und ihre Darstellungsmöglichkeiten mittels RDFS. Die Auflistung der Metadaten-Elemente stützt sich dabei auf die in Tabelle 3 zusammengefassten Forschungsergebnisse der *verwendungszweckgerichteten Metadatenkategorisierung*. Die speziell in einem RDFS für DWH-Metadaten zu definierenden Elemente sind mit dem Namensraum-Kürzel *dwhrdfs* (*dwhrdfs*: RDFS für DWH-Metadaten) gekennzeichnet und kursiv dargestellt. Klassen und Eigenschaftstypen, die erst bei der konkreten Ausgestaltung des Schemas spezifiziert werden können, sind durch die Platzhalter *dwhrdfs:spezProperty* und *dwhrdfs:SpezClass* vorgemerkt.

| Metadatum | Beschreibung | RDFS-Konzept | RDFS-Element |
|---------------------|---|---|---|
| Begriffsidentität | Eindeutige Bezeichnung (Identifikator) des Begriffs | (URI-Konzept) | rdf:ID |
| Begriffsbenennung | Das eigentliche Begriffswort (vgl. Abschnitt 5.2.1) | | rdfs:label |
| Begriffsdefinition | Bedeutung des Begriffs in einer Domäne | | dc:description rdfs:comment |
| Begriffsbeziehungen | Beziehungen des Begriffs zu anderen Begriffen | rdf:Property rdfs:subPropertyOf rdfs:subClassOf | <i>dwhrdfs:spezProperty</i> rdfs:seeAlso |
| Synonyme | (vgl. Abschnitt 5.2.2) | Aufdeckung anhand gleicher/ähnlicher Definition bzw. Beschreibung | |
| Homonyme | (vgl. Abschnitt 5.2.2) | Unmöglichkeit aufgrund des URI-Konzepts (vgl. Begriffsidentität) | |

Tabelle 5: RDFS-Elemente der Metadaten-Kategorie Terminologie

MERKMALE ÜBER DATENSTRUKTUR UND DATENBEDEUTUNG

Datenstrukturen entstehen durch Aggregation einfacher Datenobjekttypen. Ein Datenobjekttyp ist eine Beschreibung eines Datenobjekts anhand seines Wertebereichs und seiner Klassifizierung. Ein Datenobjekt dient der informationstechnischen Speicherung eines Datenwerts. Der zugehörige Datenobjekttyp gibt Wertebereich und Klassifizierung des Datenobjekts an [FeSi06, S. 307 ff.], [Auth04, S. 49].

Name und beschreibender Text einer Datenstruktur können über die DC-Elemente *dc:title*, *dc:description* bzw. über das RDF-inhärente Beschreibungselement *rdf:description* dargestellt werden. Zur Klassifizierung einer Datenstruktur (Typ) sind spezifische Klassen zu definieren, deren Semantik mithilfe von *rdfs:label* und *rdfs:comment* angegeben werden kann. Ebenso sind für die Beschreibung der Elemente einer Datenstruktur (Datenobjekttypen) eigene Klassen sowie zur Darstellung von Beziehungen zwischen diesen Elementen entsprechende Eigenschaftstypen erforderlich. Die herauszuarbeitenden Klassen und Eigenschaftstypen können durch die oben genannten RDFS-eigenen Beschreibungsstrukturen semantisch charakterisiert werden.

Die zusätzliche Angabe von Domänen- und Wertebereichseinschränkungen (*rdfs:domain*, *rdfs:range*) grenzt die Elemente hinsichtlich ihrer Kombinierbarkeit ein. Für die Angabe eines beliebigen Datentyps zu einem Datenstruktur-Element werden die Klasse *dwhrdfs:Datentyp* und der Eigenschaftstyp *dwhrdfs:hatDatentyp* vorgeschlagen. Darüber hinaus können Literale durch das RDF-Element *rdf:datatype* typisiert werden. Der Wert dieser Eigenschaft referenziert jeweils einen XML-Schema-Standarddatentyp [Beck04], [AnHa04b, S. 67 f.].

Angaben zum Ersteller, Erstellungsdatum und Publizisten der Datenstruktur sind über die DC-Elemente *dc:creator*, *dc:date* und *dc:publisher* möglich. Die nachfolgende Tabelle fasst die Erläuterungen dieser Metadaten-Kategorie zusammen.

| Metadatum | Beschreibung | RDFS-Konzept | RDFS-Element |
|-------------------------|---|---|--|
| Name | Benennung der Datenstruktur | | dc:title |
| Beschreibung | Zweck und Bedeutung der Datenstruktur | | rdf:description dc:description |
| Typ | Klassifizierung der Datenstruktur | rdfs:Class | <i>dwhrdfs:SpezClass</i> rdfs:label rdfs:comment |
| Elemente | Elemente der Datenstruktur | rdf:Property rdfs:Class Domänen- und Wertebereichseinschränkung | <i>dwhrdfs:spezProperty</i> <i>dwhrdfs:SpezClass</i> rdfs:label rdfs:comment rdfs:domain rdfs:range |
| Datentyp | Datentyp der Elemente | rdf:Property rdfs:Class | <i>dwhrdfs:hatDatentyp</i> <i>dwhrdfs:Datentyp</i> rdf:datatype |
| Datenstruktur-ersteller | Ersteller bzw. Verantwortlicher für die Datenstruktur | | dc:creator dc:publisher |
| Erstellungsdatum | Datum der Erstellung der Datenstruktur | | dc:date |

Tabelle 6: RDFS-Elemente der Metadaten-Kategorie Datenstruktur und -bedeutung

MERKMALE ZUR HISTORISIERUNG DER METADATEN

Die Historisierung von Metadaten dient der Dokumentation von Änderungen an Metadaten über die Zeit. Versionierung ist hierzu ein mögliches Konzept.

Um Änderungen bzw. Weiterentwicklungen am RDF-Vokabular nachvollziehen zu können, werden Eigenschaftstypen zur Pflege der Versionsnummer (*dwhrdfs:istVersion*) und zur Dokumentation der vorgenommenen Änderungen (*dwhrdfs:umfasstAenderung*) vorgeschlagen. Letztere ist hier als Untereigenschaft (*rdfs:subPropertyOf*) zur Versionsangabe ausgewiesen. Die Angabe des Änderungsdatums kann über den DC-Term⁹⁸ *dcterms:modified* erfolgen. Nachstehend sind die Ausführungen tabellarisch dargestellt.

| Metadatum | Beschreibung | RDFS-Konzept | RDFS-Element |
|----------------|---|------------------------------------|---------------------------------|
| Version | Versionsbezeichnung eines Metadaten-elements oder der Datenstruktur | rdf:Property | <i>dwhrdfs:istVersion</i> |
| Änderung | Beschreibung der Änderung zur Vorgängerversion | rdf:Property rdfs:subPropertyOf | <i>dwhrdfs:umfasstAenderung</i> |
| Änderungsdatum | Datum der vorgenommenen Änderung | | <i>dcterms:modified</i> |

Tabelle 7: RDFS-Elemente der Metadaten-Kategorie Metadatenhistorie

7.1.2 Data-Warehouse-Kernelemente

Das multidimensionale Datenmodell (z. B. [BaGü04, S. 102 ff.]) erlaubt es, auswertungsorientierte Schemata für ein DWH zu erstellen. Da jedoch gegenwärtig keine formale sowie standardisierte Beschreibung des multidimensionalen Datenmodells existiert [BaGü04, S. 176], wird zur Identifikation und Definition der Elemente dieser Kategorie folgendes Vorgehen gewählt:

⁹⁸ Zu *DC-Terms* zählen alle Metadaten-Ausdrücke, die von der Dublin Core Metadata Initiative verwaltet werden. Diese sind unter <http://dublincore.org/documents/dcmi-terms> verfügbar (Abruf am 10.06.2008).

Zum einen dienen elementare Bestandteile und deren Beziehungen, die im Metamodell⁹⁹ des *SDWM* ([Böhn01, S. 287]) sowie im Metamodell des *Multidimensionalen Entity-Relationship-Modells*¹⁰⁰ (ME/RM; z. B. [SBHD98, S. 109 ff.]) dargelegt sind, dem Aufbau eines RDF-Vokabulars zur semantischen Beschreibung der DWH-Kern-elemente. Zum anderen werden Elemente und deren Beschreibungen aus der *Metadaten-Kategorie Datenanalyse* ergänzt.

Aufgrund der in Abschnitt 3.4.2 aufgedeckten Schwächen (u. a. Komplexität, geringe Unterstützung semantischer Aspekte) wird das CWM an dieser Stelle nicht direkt referenziert. Wesentliche Elemente der beiden als Vorlage gewählten Metamodelle finden sich jedoch auch in den beiden CWM-Paketen *Multidimensional* und *OLAP* wieder.

Der multidimensionalen Modellierung liegt die Metapher¹⁰¹ eines mehrdimensionalen *Würfels* (häufig auch als *Cube* oder *Hypercube* bezeichnet) zugrunde. Ein *Cube* dient der Beschreibung von mehrdimensionalen Sichten auf ein DWH. Dieser wird durch *Dimensionen* aufgespannt, die aus einer nicht leeren Menge von *Dimensionselementen* bestehen. Dimensionen umfassen qualitative Merkmale der DWH-Daten. Quantitative Daten eines DWH werden an dieser Stelle als *Kennzahlen* bezeichnet (z. B. [Böhn01, S. 126 ff.], [BaGü04, S. 103 ff.]).

Die Elemente *Cube*, *Dimension* und *Dimensionselement* (im *SDWM* auch als *Dimensionshierarchiestufe* bezeichnet) können in einem RDFS über eigens definierte Klassen und Eigenschaftstypen abgebildet und zueinander in Beziehung gesetzt werden (z. B. *dwhrdfs:Dimension* und *dwhrdfs:hatDimension*). Dieses gilt analog für so genannte *Dimensionsattribute*, die zur Charakterisierung eines Dimensionselements eingesetzt werden können. Für die Angabe der Position eines Schemaelements in einer Dimension (vgl. Abschnitt 5.5) kann bspw. ein Dimensionsattribut *Stufe* formuliert werden. Dimensionsattribute sind ebenso wie Objekte des nachfolgend diskutierten Metadatums *Kennzahl* durch einen Verweis auf einen *Datentypen* (vgl. Abschnitt 7.1.1) zu spezifizieren.

⁹⁹ Ein *Metamodell* legt das zur Spezifikation von Modellsystemen verfügbare Begriffssystem fest. Eine fundierte Erläuterung des Begriffs *Metamodell* ist u. a. [FeSi06, S. 124 ff.] zu entnehmen.

¹⁰⁰ Das *Multidimensionale Entity-Relationship-Modell* (z. B. [SBHD98]) stellt eine Erweiterung des Entity-Relationship-Ansatzes für relationale Schemata ([Chen76]) dar. Es wurde speziell für multidimensionale Schemata entwickelt [BaGü04, S. 165 ff.].

¹⁰¹ Eine Erläuterung des Begriffs *Metapher* im Kontext der Modellierung bietet z. B. [FeSi06, S. 124 f.].

Zur Beschreibung von Kennzahlen können in einem RDFS spezifische Klassen erstellt und zur Darstellung von Kennzahlenbeziehungen Unterklassebeziehungen oder individuell ausgestaltete Eigenschaftstypen eingesetzt werden. In Anlehnung an das Metamodell des SDWM, sollen Kennzahlen zusätzlich differenziert werden in

- *Basiskennzahlen*, die ohne zusätzliche Berechnungsvorschrift aus den Fakten bereitgestellt werden,
- *gefilterte Kennzahlen*, die anhand eines Filterwerts über ein Dimensionselement eingeschränkt sind, und
- *abgeleitete Kennzahlen*, die über eine Berechnungsvorschrift determiniert sind.

Für die Repräsentation dieser Kennzahlentypen empfiehlt sich jeweils die Definition einer gleichnamigen Klasse. Die zwischen einer gefilterten Kennzahl und einem Dimensionselement erforderliche Beziehung kann durch die Bereitstellung eines hierfür explizit angelegten Eigenschaftstyps modelliert werden (z. B. *dwhrdfs:gefiltertNach*). Zudem ist jeweils ein Eigenschaftstyp zur Beschreibung des genauen Filterwerts (z. B. *dwhrdfs:hatFilterwert*) sowie zur Referenzierung der Kennzahl, auf der die gefilterte Kennzahl basiert (z. B. *dwhrdfs:basiertAufKennzahl*), notwendig. Für die Erfassung von Berechnungsvorschriften wird die Klasse *dwhrdfs:Berechnung* vorgeschlagen, die über zu spezifizierende Eigenschaftstypen einerseits zu definieren und andererseits mit entsprechenden Kennzahlen in Beziehung zu setzen ist.

Zur Definition gültiger Kombinationen von Dimensionselementen mit Kennzahlen (vgl. u. a. Metamodell des ME/RM) ist ein weiterer Eigenschaftstyp festzulegen (z. B. *kombinierbarMitKennzahl*). Die Aggregierbarkeit einer Kennzahl in Bezug auf ein Dimensionselement ist bedingt durch deren Typverträglichkeit.¹⁰² Als Beispiel sollen hier Bestandskennzahlen dienen, die in Bezug auf temporale Merkmale prinzipiell nicht zu verdichten sind. Ist eine Kennzahl hinsichtlich eines Dimensionselements nicht aggregierbar, wird von *Semiaggregierbarkeit* gesprochen [Böhn01, S. 274]. Dieser Sachverhalt kann durch Einsatz eines gesonderten Eigenschaftstyps gekennzeichnet werden (z. B. *dwhrdfs:istSemiaggregierbar*).

¹⁰² Eine Erläuterung der Typverträglichkeit in diesem Zusammenhang bietet u. a. [Böhn01, S. 139 f.]

Tabelle 8 gibt einen Überblick über die DWH-Kernelemente und deren Abbildungsmöglichkeit in einem RDFS. Dabei werden in der rechten Tabellenspalte jeweils konkrete Vorschläge für die zu erstellenden RDFS-Klassen und RDFS-Eigenschaftstypen unterbreitet. Die in der Spalte Metadatum in Klammern stehenden Begriffe sind als mögliche Oberklasse eines spezifischen Metadatums anzusehen.

Gemäß der RDF-Notationskonvention beginnt ein URI, der eine Klasse repräsentiert, mit einem Großbuchstaben. Die eindeutigen Bezeichner von Eigenschaftstypen werden dagegen klein geschrieben [HKRS08, S. 70]. Diese Vorgabe wird auch bei der Auszeichnung der vorgeschlagenen Elemente des RDFS für DWH-Metadaten angewandt.

| Metadatum | Beschreibung | RDFS-Konzept | RDFS-Element |
|--------------------|---|----------------------------|--|
| Cube | Bezeichnung und Beschreibung von materialisierten mehrdimensionalen Sichten auf ein DWH | rdfs:Class | <i>dwhrdfs:Cube</i> |
| Dimension | Beschreibung der Dimensionen, die einen Cube aufspannen | rdf:Property rdfs:Class | <i>dwhrdfs:hatDimension</i> <i>dwhrdfs:Dimension</i> |
| Dimensionselement | Elemente einer Dimension (Dimensionshierarchiestufe) | rdf:Property rdfs:Class | <i>dwhrdfs:hat</i> <i>DimensionsElement</i> <i>dwhrdfs:kombinierbar</i> <i>MitKennzahl</i> <i>dwhrdfs:</i> <i>DimensionsElement</i> |
| Dimensionsattribut | Ergänzende Informationen zu einem Dimensionselement | rdf:Property rdfs:Class | <i>dwhrdfs:hat</i> <i>DimensionsAttribut</i> <i>dwhrdfs:Dimensions</i> <i>Attribut</i> |

Fortsetzung: nächste Seite

| Metadatum | Beschreibung | RDFS-Konzept | RDFS-Element |
|---------------------------------|--|---|--|
| Kennzahl | Oberklasse der quantitativen Merkmale | rdf:Property rdfs:Class | <i>dwhrdfs:istSemiagggregierbar</i> <i>dwhrdfs:Kennzahl</i> |
| Basiskennzahl (Kennzahl) | Quantitatives Merkmal, dem keine Berechnungsvorschrift oder Einschränkung zugrunde liegt | rdfs:Class rdfs:subClassOf | <i>dwhrdfs:BasisKennzahl</i> |
| Gefilterte Kennzahl (Kennzahl) | Anhand eines Dimensionselements eingeschränkte Kennzahl | rdf:Property rdfs:Class rdfs:subClassOf | <i>dwhrdfs:basiertAufKennzahl</i> <i>dwhrdfs:gefiltertNach</i> <i>dwhrdfs:hatFilterwert</i> <i>dwhrdfs:GefilterteKennzahl</i> |
| Abgeleitete Kennzahl (Kennzahl) | Auf einer Berechnungsvorschrift basierende Kennzahl | rdf:Property rdfs:Class rdfs:subClassOf | <i>dwhrdfs:berechnetSichAus</i> <i>dwhrdfs:AbgeleiteteKennzahl</i> |
| Berechnung | Definition einer Berechnungsvorschrift für abgeleitete Kennzahlen | rdf:Property rdfs:Class | <i>dwhrdfs:hatOperator</i> <i>dwhrdfs:hatOperand1</i> <i>dwhrdfs:hatOperand2</i> <i>dwhrdfs:hatKonstante</i> <i>dwhrdfs:Berechnung</i> |
| Datentyp | Datentyp der Elemente | rdf:Property rdfs:Class | <i>dwhrdfs:hatDatentyp</i> <i>dwhrdfs:Datentyp</i> rdf:datatype |

Tabelle 8: RDFS-Elemente der Kategorie DWH-Kernelemente

7.1.3 Data-Warehouse-Organisationselemente

Die Kategorie DWH-Organisationselemente basiert auf der Metadaten-Kategorie Organisationsbezug. Die zugehörigen Elemente sollen Auskunft über Entstehung und Verwendung der Daten geben sowie Zuständigkeiten in einem DWH-System offen legen.

Für die Beschreibung von Organisationseinheiten, die im Rahmen ihrer Geschäftsprozesse Daten generieren (Datenproduzent) oder Anfragen an die Datenbasis eines DWH-Systems stellen (Datenkonsument), ist die Definition eigener Klassen und Eigenschaftstypen notwendig (z. B. *dwhrdfs:Datenproduzent* und *dwhrdfs:hatDatenproduzent*).

Eine Zugriffsberechtigung auf Schemaelemente kann bspw. durch die Ausgestaltung eines Eigenschaftstyps (z. B. *dwhrdfs:zugriffsberechtigteOrganisationseinheit*) dargelegt werden.

Zur Nennung und zum Verweis auf Ansprechpartner für bestimmte Komponenten eines DWH-Systems (Fachverantwortlicher, technischer Ansprechpartner, Berichtsverantwortlicher) sind im RDFS weitere Klassen und Eigenschaftstypen anzulegen.

Eine Übersicht der Metadaten-Elemente dieser Kategorie sowie eine Aufstellung erforderlicher Klassen und Eigenschaftstypen bietet nachfolgende Tabelle.

| Metadatum | Beschreibung | RDFS-Konzept | RDFS-Element |
|--|--|---|---|
| Datenproduzent (Organisationseinheit) | Organisationseinheit, der die Daten entstammen | rdf:Property rdfs:Class rdfs:subClassOf | <i>dwhrdfs:hat</i> <i>Datenproduzent</i> <i>dwhrdfs:</i> <i>Datenproduzent</i> |
| Datenkonsument (Organisationseinheit) | Organisationseinheit, welche die Daten verwendet | rdf:Property rdfs:Class rdfs:subClassOf | <i>dwhrdfs:hat</i> <i>Datenkonsument</i> <i>dwhrdfs:</i> <i>Datenkonsument</i> |
| Berechtigung | Zugriffsberechtigung auf Schemaelemente (Organisationseinheit) | rdf:Property | <i>dwhrdfs:</i> <i>zugriffsberechtigte</i> <i>Organisationseinheit</i> |
| Fachverantwortlicher (Person) | Für einen Begriff verantwortliche Person (Bedeutung, Ursprung des Begriffs etc.) | rdf:Property rdfs:Class rdfs:subClassOf | <i>dwhrdfs:hatFach-</i> <i>Verantwortlichen</i> <i>dwhrdfs:Fach-</i> <i>Verantwortlicher</i> |
| Technischer Ansprechpartner (Person) | In Bezug auf (system-)technische Aufgaben zuständige Person | rdf:Property rdfs:Class rdfs:subClassOf | <i>dwhrdfs:hatTechni-</i> <i>schsenAnsprechpartner</i> <i>dwhrdfs:Technischer</i> <i>Ansprechpartner</i> |
| Berichts- verantwortlicher (Person) | Ersteller bzw. Verantwortlicher für ein Berichtswesen | rdf:Property rdfs:Class rdfs:subClassOf | <i>dwhrdfs:hatBerichts-</i> <i>Verantwortlichen</i> <i>dwhrdfs:Berichts-</i> <i>Verantwortlicher</i> |

Tabelle 9: RDFS-Elemente der Kategorie DWH-Organisationselemente

7.1.4 Data-Warehouse-Kontextelemente

Die Ausgestaltung der Kategorie DWH-Kontextelemente orientiert sich an den Ergebnissen des Abschnitts 4.3. Kontextinformationen sind für die korrekte Interpretation von Daten notwendig.

Die im oben genannten Abschnitt identifizierten Kontexttypen können im RDFS mittels individuell spezifizierter Klassen und Eigenschaftstypen beschrieben und in Beziehung gesetzt werden (z. B. *dwhrdfs:DomaenenKontext*, *dwhrdfs:besitztDomaenenKontext*). Zusätzlich wird die Definition einer übergeordneten Klasse (z. B. *dwhrdfs:Kontext*) empfohlen, um allen Kontexttypen gemeinsame Eigenschaften zentral verwalten zu können. Diesbezüglich sind auch die RDFS-inhärenten Konzepte Unterklasse und Untereigenschaft zu nutzen.

Die nachstehende Tabelle skizziert die Metadaten-Elemente der hier erläuterten Kategorie sowie deren Darstellungsmöglichkeiten in einem RDFS.

| Metadatum | Beschreibung | RDFS-Konzept | RDFS-Element |
|---------------------------|---|---|--|
| Domänenkontext (Kontext) | Angaben zum Domänenkontext der Daten eines DWH (Abschnitt 4.3). | rdf:Property rdfs:subPropertyOf rdfs:Class rdfs:subClassOf | <i>dwhrdfs:besitzt</i> <i>DomaenenKontext</i> <i>dwhrdfs:DomaenenKontext</i> |
| Umweltkontext (Kontext) | Darlegung des Umweltkontexts zu den Daten eines DWH (Abschnitt 4.3). | rdf:Property rdfs:subPropertyOf rdfs:Class rdfs:subClassOf | <i>dwhrdfs:besitzt</i> <i>UmweltKontext</i> <i>dwhrdfs:UmweltKontext</i> |
| Aufgabenkontext (Kontext) | Beschreibung des Aufgabenkontexts zu den Daten eines DWH (Abschnitt 4.3). | rdf:Property rdfs:subPropertyOf rdfs:Class rdfs:subClassOf | <i>dwhrdfs:besitzt</i> <i>AufgabenKontext</i> <i>dwhrdfs:AufgabenKontext</i> |
| Nutzerkontext (Kontext) | Informationen zum Nutzerkontext der Daten eines DWH (Abschnitt 4.3). | rdf:Property rdfs:subPropertyOf rdfs:Class rdfs:subClassOf | <i>dwhrdfs:besitzt</i> <i>NutzerKontext</i> <i>dwhrdfs:NutzerKontext</i> |

Tabelle 10: RDFS-Elemente der Kategorie DWH-Kontextelemente

7.1.5 Data-Warehouse-Systemelemente

Zur Kategorie DWH-Systemelemente zählen sowohl Elemente der Metadaten-Kategorie Systembezug als auch der Datentransformation. Sie dienen der Charakterisierung systemtechnischer Eigenschaften sowie der Beschreibung der Datentransformation von einem Vorksystem in ein DWH-System.

Analog zu den bisherigen Erkenntnissen können auch in dieser Kategorie die erforderlichen Elemente durch spezielle Klassen und Eigenschaftstypen im RDFS abgebildet werden. Zudem kann bei den Elementen zur Systemklassifizierung (*Vorksystem* und *DWH-System*) die Vererbung gemeinsamer Merkmale durch Unterklasse- und Untereignisbeziehungen genutzt werden.

Bei der Beschreibung einer *Software* wird das Paradigma zugrunde gelegt, dass diese von einem bestimmten Hersteller angeboten wird und von einem *Software-Hersteller* unterschiedliche *Softwarekomponenten* erstellt werden.

Eine *Datenbank* kann als ein Bestandteil eines Vor- oder eines DWH-Systems spezifiziert werden. Für die Definition von Datenbanken sowie von *Tabellen* und *Tabellenattributen* sind eigene Klassen und für deren Abhängigkeiten und Beziehungen zugehörige Eigenschaftstypen anzulegen.¹⁰³

Die konkreten Vorschläge zur Benennung der erforderlichen RDFS-Elemente sind der nachstehenden Tabelle zu entnehmen. Lediglich das Metadatum *Transformationsschritt* soll noch an dieser Stelle explizit hervorgehoben werden. Für dieses wird die Bildung folgender Eigenschaftstypen empfohlen:

- *dwhrdfs:entstammtQuellattribut* referenziert das Quellattribut eines Datenwerts,
- *dwhrdfs:wirdTransformiertAuf* verweist auf das Zielattribut und
- *dwhrdfs:wirdTransformiert* beschreibt die eigentliche Datentransformation.

¹⁰³ An dieser Stelle wird von einem ROLAP-Ansatz und somit einer relationalen Datenspeicherung ausgegangen (vgl. Abschnitt 2.3.1). Relationale DBMS sind im Kontext von DWH-Systemen am häufigsten vertreten [QuKC05, S. 6]. Für multidimensionale Datenbanken o. Ä. sind gemäß dem jeweiligen Datenmodell entsprechende Klassen und Eigenschaftstypen zu definieren.

| Metadatum | Beschreibung | RDFS-Konzept | RDFS-Element |
|--------------------------------------|--|---|--|
| Vorsystem, DWH-System (System) | Spezifikation des Vorsystems oder des DWH-Systems | rdf:Property rdfs:subPropertyOf rdfs:Class rdfs:subClassOf | <i>dwhrdfs:basiertAuf Vorsystem dwhrdfs:basiertAuf DWHSystem dwhrdfs:Vorsystem dwhrdfs:DWHSystem</i> |
| Softwarehersteller | Name des Softwareherstellers | rdf:Property rdfs:Class | <i>dwhrdfs:angeboten DurchSoftware Hersteller dwhrdfs:Software Hersteller</i> |
| Softwarekomponente | Bezeichnung der Softwarekomponente | rdf:Property rdfs:Class | <i>dwhrdfs:erstellt SoftwareKomponente dwhrdfs:Software Komponente</i> |
| Softwareversion | Versionsangabe einer Softwarekomponente | rdf:Property | <i>dwhrdfs:hatSoftware Version</i> |
| Hardware | Typbezeichnung der eingesetzten Hardware | rdf:Property | <i>dwhrdfs:basiertAuf Hardware</i> |
| Datenbank | Spezifikation einer Datenbank (eines Vor- oder DWH-Systems) | rdf:Property rdfs:Class | <i>dwhrdfs:basiertAuf Datenbank dwhrdfs:Datenbank</i> |
| Tabelle | Benennung und Beschreibung einer Tabelle in einer Datenbank | rdf:Property rdfs:Class | <i>dwhrdfs:basiertAuf Tabelle dwhrdfs:beinhaltet Tabelle dwhrdfs:Tabelle</i> |
| Tabellenattribut | Bestimmung eines Tabellenattributs | rdf:Property rdfs:Class | <i>dwhrdfs:basiertAuf Tabellenattribut dwhrdfs:hat Tabellenattribut dwhrdfs: Tabellenattribut</i> |
| Transformations- schritt | Beschreibung eines Transformationsschritts der Daten von einem Vorsystem in ein DWH-System | rdf:Property | <i>dwhrdfs:entstammt Quellattribut dwhrdfs:wird TransformiertAuf dwhrdfs:wird Transformiert</i> |

Tabelle 11: RDFS-Elemente der Kategorie DWH-Systemelemente

7.1.6 Zusammenfassung

In diesem Abschnitt wurden die von RDF und RDFS bereitgestellten Konstrukte und Beschreibungselemente auf ihre Anwendbarkeit zur Darstellung von DWH-Metadaten untersucht. Dabei konnte aufgezeigt werden, dass ein Teil der notwendigen Beschreibungskomponenten für DWH-Metadaten über RDF- und DC-Standardelemente abbildbar ist. Für die zusätzlich zur semantischen Erfassung von DWH-Metadaten erforderlichen Elemente sind spezifisch zu definierende Klassen und Eigenschaftstypen einzusetzen. Hierfür wurden jeweils konkrete Ausprägungen von RDFS-Klassen und -Eigenschaftstypen vorgeschlagen. Diese detailliert auszugestalten und in einem integrierten RDFS in Beziehung zu setzen ist Bestandteil des nachfolgenden Abschnitts.

7.2 Ein RDF-Schema zur semantischen Beschreibung von Data-Warehouse-Metadaten

Basierend auf den Ergebnissen des vorhergehenden Abschnitts wird nachstehend ein umfassendes RDFS zur semantischen Beschreibung von DWH-Metadaten vorgestellt (DWH-RDFS). In diesem sind die oben vorgeschlagenen RDFS-Elemente detailliert ausgestaltet, ihre Abhängigkeiten erfasst und entsprechende Domänen- und Wertebereichseinschränkungen (*rdfs:range*, *rdfs:domain*) angelegt. Die Erläuterungen hierzu folgen der eingeführten Klassifizierung in allgemeine Beschreibungselemente, DWH-Kernelemente, DWH-Organisationselemente, DWH-Kontextelemente und DWH-Systemelemente. Diese Unterteilung findet sich auch im resultierenden DWH-RDFS wieder. Abschließend wird die Anwendbarkeit des DWH-RDFS aufgezeigt. Die nachfolgenden Ausführungen sind in Abbildung 7-1 (Seite 160) visualisiert.

ALLGEMEINE BESCHREIBUNGSKONZEPTE UND -ELEMENTE

Die Namensraum-Deklaration des DWH-RDFS umfasst neben den obligatorischen Verweisen zu den Namensräumen für RDF und RDFS auch die Einbindung der Namensräume DC und DC-Terms. Der dem DWH-RDFS zugewiesene Namensraum lautet:

```
xmlns:dwhrdfs="http://www.semantic-metadata.com/dwhrdfs/dwhrdfs#"
```

Auf die explizite Angabe des Namensraumkürzels *dwhrdfs* wird bei der nachfolgenden textlichen Beschreibung der DWH-RDFS-Elemente verzichtet. Die verpflichtende Angabe eines URI für jedes Element erfolgt – im Hinblick auf ein intuitives Verständnis und eine leichte Handhabung des DWH-RDFS – jeweils durch einen eindeutigen, sprechenden Namen.

Zur allgemeinen Beschreibung des DWH-RDFS wird auf Elemente des DC-Vokabulars zurückgegriffen. Dabei dient *dc:title* der Benennung, *dc:description* der Beschreibung sowie *dc:creator* und *dc:publisher* der Spezifikation des Erstellers und des Publizisten der Datenstruktur. Die Angabe des Erstellungsdatums erfolgt mittels *dc:date*. Die Elemente des DWH-RDFS sind jeweils durch die RDFS-eigenen Ausdrücke *rdfs:label* und *rdfs:comment* semantisch genau beschrieben.

Die Versionierung der Datenstruktur erfolgt anhand speziell für diesen Zweck definierter Eigenschaftstypen. Der Eigenschaftstyp *istVersion* erlaubt in Form eines Literals die Zuordnung einer Versionsnummer zum DWH-RDFS. Die Darlegung der Änderung zur Vorgängerversion ist mit dem Eigenschaftstyp *umfasstAenderung* möglich (Wertangabe als Literal). Dieser stellt eine Untereigenschaft zu *istVersion* dar.

DATA-WAREHOUSE-KERNELEMENTE

Das Wurzelement des DWH-RDFS und damit zentrale Ausgangselement zur Beschreibung eines DWH-Systems ist die Klasse *Cube*. Die einen Cube aufspannenden Dimensionen (Klasse *Dimension*) können über den Eigenschaftstyp *hatDimension* referenziert werden. Allen Eigenschaftstypen des entwickelten DWH-RDFS sind – gemäß ihren erlaubten Kombinationsmöglichkeiten mit einzelnen Klassen – entsprechende Domänen- (*rdfs:domain*) und Wertebereichseinschränkungen (*rdfs:range*) zugewiesen. Für den Eigenschaftstyp *hatDimension* sind dies die Klasse *Cube* als Domänen- und die Klasse *Dimension* als Wertebereichseinschränkung.

Zu einem Cube können mithilfe des Eigenschaftstyps *besitztKontext* Kontextinformationen (Klasse *Kontext*) hinterlegt werden (vgl. Abschnitt DWH-Kontextelemente). Des Weiteren besteht die Möglichkeit, über den Eigenschaftstyp *hatFachVerantwortlichen* auf eine für einen Cube fachlich zuständige Person zu verweisen und über *zugriffsberechtigteOrganisationseinheit* die auf einen Cube zugriffsberechtigten Organisationseinheiten zu benennen. Der Verantwortliche für das Berichtswesen, das auf einem bestimmten Cube basiert, ist mithilfe des Eigenschaftstyps *hatBerichtsVerantwortlichen* festzulegen (vgl. Abschnitt DWH-Organisationselemente).

Um die Elemente einer Dimension spezifizieren zu können, wurden die Klasse *DimensionsElement* sowie der Eigenschaftstyp *hatDimensionsElement* in das DWH-RDFS aufgenommen. Ein Dimensionselement kann anhand von Dimensionsattributen (Klasse *DimensionsAttribut*, Eigenschaftstyp *hatDimensionsAttribut*) näher beschrieben werden. Der zwischen zwei Dimensionselementen definierte Eigenschaftstyp *verdichtetZu* (Domänen- und Wertebereich sind jeweils auf die Klasse *DimensionsElement* eingeschränkt) ermöglicht es, Abhängigkeiten zwischen Dimensionselementen darzustellen. Mithilfe dieses Eigenschaftstyps können auch parallele oder unbalancierte Hierarchien (z. B. [Böhm01, S. 134 ff.]) innerhalb einer Dimension beschrieben werden.

Analog zur Klasse *Cube* besteht die Möglichkeit, einem Dimensionselement Kontextinformationen (Eigenschaftstyp *besitztKontext*) sowie einen oder mehrere Fachverantwortliche (Eigenschaftstyp *hatFachVerantwortlichen*) beizuordnen. Angaben über die *datenproduzierende* sowie *datenkonsumierende* Organisationseinheit eines Dimensionselements erfolgen mithilfe der Eigenschaftstypen *hatDatenproduzent* bzw. *hatDatenkonsument* (vgl. Abschnitt DWH-Organisationselemente).

Für die individuelle Spezifikation des Datentyps eines Dimensionsattributs wurde die gleichnamige Klasse *Datentyp* und der Eigenschaftstyp *hatDatentyp* definiert. Mittels dieses Eigenschaftstyps kann ebenso für Kennzahlen (Klasse *Kennzahl*) ein Datentyp angegeben werden. Die im vorhergehenden Abschnitt genannten Kennzahlentypen (Basiskennzahl, gefilterte und abgeleitete Kennzahl) werden im DWH-RDFS durch jeweils eigene Klassen repräsentiert, die als Unterklasse zur Klasse *Kennzahl* definiert sind (Klasse *Basiskennzahl*, *AbgeleiteteKennzahl*, *GefilterteKennzahl*). Eine Verknüpfung zur Klasse *Datentyp* ist für diese drei Kennzahlentypen somit ebenfalls zulässig.

Um Dimensionselemente mit Kennzahlen in Beziehung setzen und deren fachlich und logisch gültigen Kombinationen darlegen zu können, wurde der Eigenschaftstyp *kombinierbarMitKennzahl* angelegt (Domäne *DimensionsElement*; Wertebereich *Kennzahl*). Für die Kennzeichnung der Semiaggregierbarkeit einer Kennzahl in Bezug auf ein Dimensionselement steht der Eigenschaftstyp *istSemiaggregierbar* bereit (Domäne *Kennzahl*; Wertebereich *Dimensionselement*).

Eine gefilterte Kennzahl ist durch ein Dimensionselement hinsichtlich quantitativer Kriterien eingeschränkt. Für die Modellierung dieses Sachverhalts wurde der Eigenschaftstyp *gefiltertNach* zwischen den Klassen *GefilterteKennzahl* und *DimensionsElement* definiert. Darüber hinaus sind hierfür – gemäß den Vorschlägen des vorherigen Kapitels – die Eigenschaftstypen *basiertAufKennzahl* (Domäne *GefilterteKennzahl*; Wertebereich *Kennzahl*) und *hatFilterwert* (Wertangabe als Literal) verfügbar.

Die Berechnungsvorschrift für eine abgeleitete Kennzahl kann mithilfe der Klasse *Berechnung* und dem Eigenschaftstyp *berechnetSichAus* erfasst werden. Eine Berechnung setzt sich aus zwei Operanden, die entweder beide durch eine Kennzahl (*hatOperand1*, *hatOperand2*) oder durch eine Kennzahl und eine Konstante (*hatKonstante*) repräsentiert sind, und einem Operator (*hatOperator*) zusammen. Die beiden letztgenannten Eigenschaftstypen erwarten die Objektwerte als Literal. Auf diese Weise ist eine hohe Flexibilität bei der Abbildung der Berechnungen gewährleistet. Komplexe Rechenschritte können durch rekursive Definition abgeleiteter Kennzahlen erstellt werden.

Die Option, die einem DWH zugrunde liegenden oder vorgeschalteten Systeme und Datenbanken zu beschreiben, wird im Abschnitt DWH-Systemelemente detailliert ausgeführt. An dieser Stelle soll lediglich vermerkt werden, dass die Klassen *Cube*, *DimensionsElement* und *Kennzahl* mit entsprechenden Eigenschaftstypen kombiniert werden können, um system- und datenbanktechnische Informationen zu diesen zu erfassen (z. B. *basiertAufSystem*, *basiertAufDatenbank*).

DATA-WAREHOUSE-ORGANISATIONSELEMENTE

Die beiden elementaren Klassen der Kategorie DWH-Organisationselemente lauten *Person* und *Organisationseinheit*.

Zur Klasse *Person* sind die Unterklassen *BerichtsVerantwortlicher*, *Technischer-Ansprechpartner* und *FachVerantwortlicher* modelliert. Deren Verknüpfung mit anderen Klassen ist über die jeweils zugehörigen Eigenschaftstypen *hatBerichtsVerantwortlichen* (Domäne *Cube*), *hatTechnischenAnsprechpartner* (Domäne *System*; vgl. Abschnitt DWH-Systemelemente) und *hatFachVerantwortlichen* (Domäne *Cube*, *DimensionsElement*, *Kennzahl*, *System*) möglich. Auf diese Weise können für die im Rahmen der jeweiligen Domäneneinschränkung zulässigen Klassen Berichtsverantwortliche, technische Ansprechpartner oder Fachverantwortliche bestimmt werden.

Datenproduzent und *Datenkonsument* sind Unterklassen der Klasse *Organisationseinheit* und dienen der Beschreibung datenproduzierender bzw. datenkonsumierender Organisationseinheiten. Die zugeordneten Eigenschaftstypen lauten *hatDatenproduzent* und *hatDatenkonsument* (Domäne *DimensionsElement*, *Kennzahl*).

Des Weiteren erlaubt der Eigenschaftstyp *zugriffsberechtigteOrganisationseinheit* die Dokumentation von Zugriffsberechtigungen (Domäne *Cube*, *DimensionsElement*, *Kennzahl*; Wertebereich *Organisationseinheit*).

DATA-WAREHOUSE-KONTEXTELEMENTE

Die zur Beschreibung des Umfelds eines DWH-Systems identifizierten Kontexttypen Domänen-, Umwelt-, Aufgaben- und Nutzerkontext (vgl. Abschnitt 4.3) sind als jeweils gleichnamige Unterklassen (*rdfs:subClassOf*) zur Klasse *Kontext* modelliert. Diese können einerseits über den ihrer Oberklasse zugeordneten Eigenschaftstyp *besitztKontext* und andererseits über einen für jeden Kontexttyp spezifisch angelegten Eigenschaftstyp (z. B. *besitztAufgabenKontext*) referenziert werden. Letztere sind als Untereigenschaft zum Eigenschaftstyp *besitztKontext* ausgewiesen und erben von diesem die Domäneneinschränkung auf die Klassen *Cube*, *DimensionsElement* und *Kennzahl*. Die jeweils zugehörige Wertebereichseinschränkung erlaubt eine Verknüpfung nur mit dem individuell angesprochenen Kontexttyp (*besitztAufgabenKontext* darf z. B. als Objekte nur Instanzen der Klasse *AufgabenKontext* tragen).

DATA-WAREHOUSE-SYSTEMELEMENTE

In der Kategorie DWH-Systemelemente besteht die Möglichkeit, systemtechnische Aussagen über ein Vor- bzw. Quellsystem (Klasse *Vorsystem*) oder über das zugrunde liegende DWH-System (Klasse *DWHSystem*) zu erfassen. Beide Klassen stellen Unterklassen der Klasse *System* dar. Zu dieser können Angaben über die eingesetzte Hardware (Eigenschaftstyp *basiertAufHardware*) in Form eines Literals oder über die eingesetzte Software (Eigenschaftstyp *eingesetzteSoftware*) durch Verweise auf Objekte der Klasse *Software* erfolgen. Eine Verknüpfung ausgehend von den Klassen *Cube*, *DimensionsElement* und *Kennzahl* mit der Klasse *System* ist über den Eigenschaftstyp *basiertAufSystem* möglich. Für eine direkte Referenz auf die Klassen *Vorsystem* und *DWHSystem* wurden – jeweils als Untereigenschaft zu *basiertAufSystem* – zusätzlich die Eigenschaftstypen *basiertAufVorsystem* und *basiertAufDWHSystem* angelegt.

Zu einem System kann des Weiteren sowohl ein technischer (Eigenschaftstyp *hatTechnischenAnsprechpartner*) als auch fachlicher Ansprechpartner (Eigenschaftstyp *hatFachVerantwortlichen*) hinterlegt werden (vgl. Abschnitt DWH-Organisationselemente). Aufgrund der Unterklassebeziehung sind diese Eigenschaftstypen bei den Klassen *Vorsystem* und *DWHSystem* ebenfalls einsetzbar.

Die auf einem System betriebene Software (Klasse *Software*) kann wie folgt spezifiziert werden: Mit dem Eigenschaftstyp *angebotenDurchSoftwareHersteller* ist eine Verknüpfung zur Klasse *SoftwareHersteller* vorgesehen, um den Hersteller einer Software exakt zu erfassen. Die Beschreibung der verwendeten Software-Komponente eines Software-Herstellers ist über die Klasse *SoftwareKomponente* und den Eigenschaftstyp *erstelltSoftwareKomponente* möglich. Versionsangaben zu einer Software-Komponente können als Literal über den Eigenschaftstyp *hatSoftwareVersion* angegeben werden.

Bedeutungsvoll für die umfassende Beschreibung eines DWH-Systems sind Informationen über den Ursprungs- und Speicherort der Daten. Im Hinblick darauf können über den Eigenschaftstyp *verwaltetDatenbank* zu einer spezifischen Software (hier einem DBMS) die jeweils zugrunde liegenden Datenbanken (Klasse *Datenbank*) spezifiziert werden. Der Eigenschaftstyp *beinhaltetTabelle* sowie die Klasse *Tabelle* wurden definiert, um die in einer Datenbank enthaltenen Tabellen zu erfassen. Der Aufbau einer Tabelle kann mithilfe des Eigenschaftstyps *hatTabellenattribut* und der Klasse *Tabellenattribut* dargestellt werden.

Die Darlegung von Datentransformationsschritten zwischen einem Vorsystem und einem DWH-System erfolgt anhand der in Abschnitt 7.1.5 spezifizierten Eigenschaftstypen *entstammtQuellattribut*, *wirdTransformiertAuf* und *wirdTransformiert*. Die beiden Erstgenannten besitzen sowohl als Domänen- als auch Wertebereichseinschränkung die Klasse *Tabellenattribut*. Aufgrund der modellierbaren Abhängigkeit kann die Klasse *Tabellenattribut* transitiv entweder einem Vor- oder einem DWH-System zugeordnet sein. Auf diese Weise sind die Datentransformationsschritte von der Datenquelle ins Datenziel (dem DWH) dokumentierbar. Die eigentliche Datentransformation kann als Literal mittels des Eigenschaftstyps *wirdTransformiert* detailliert erläutert werden.

Des Weiteren wurde der Eigenschaftstyp *basiertAufDatenbank* angelegt, um ausgehend von den Klassen *Cube*, *DimensionsElement* und *Kennzahl* auf eine Datenbank (Klasse *Datenbank*) verweisen zu können. Wie beschrieben kann diese sowohl eine Datenquelle als auch die Datenbasis eines DWH-Systems darstellen. Zusätzlich wurde der Eigen-

schaftstyp *basiertAufTabelle* definiert. Dieser dient der direkten Verknüpfung mit einer Tabelle (Wertebereich *Tabelle*). Der Eigenschaftstyp *basiertAufTabellenattribut* ermöglicht die Referenzierung eines Tabellenattributs (Wertebereich *Tabellenattribut*).

GESAMTSICHT AUF DAS RDF-SCHEMA FÜR DATA-WAREHOUSE-METADATEN

In Abbildung 7-1 sind die eben vorgestellten Klassen und Eigenschaftstypen des DWH-RDFS zusammengefasst. Darin sind sowohl die beschriebenen Unterklassebeziehungen (gestrichelte Linie) als auch die für die Eigenschaftstypen erlaubten Kombinationen mit den jeweiligen Klassen (durchgezogene Linie) gekennzeichnet. Aus Gründen der Übersichtlichkeit wurde auf die Darstellung von Untereigenschaften verzichtet.

Die Aufteilung in die vier beschriebenen Kategorien ist durch eine farbliche Unterlegung herausgestellt. Die Abbildung lässt sich zudem vertikal in eine fachliche Ebene (DWH-Kernelemente und DWH-Kontextelemente) sowie eine ressourcentechnische Ebene (DWH-Organisationselemente und DWH-Systemelemente) gliedern. Letztere ist entsprechend in personelle und maschinelle Ressourcen zu unterteilen.

Des Weiteren stellt Quellcode 2 einen Ausschnitt des DWH-RDFS in RDF/XML-Notation dar. Das vollständig entwickelte DWH-RDFS ist über dessen URI abrufbar.

RDF-ANWENDUNGSBEISPIEL

Zur Darlegung der Anwendbarkeit des DWH-RDFS wurde exemplarisch der Ausschnitt eines DWH-Systems modelliert (vgl. Abbildung 7-2). Dieses besteht aus dem Cube *Studenten*, der die Dimension *Lehrangebot* beinhaltet (*hatDimension*). Die Dimension umfasst die Elemente *Studiengang*, *Lehreinheit* und *Fakultät* (*hatDimensionsElement*). Die Elemente besitzen in der genannten Reihenfolge eine jeweils höhere Aggregationsstufe (*verdichtetZu*). Das Dimensionselement *Studiengang* verweist – exemplarisch für alle hier genannten Dimensionselemente – auf die Kennzahl *Anzahl Studenten* (*kombinierbarMitKennzahl*).

Abbildung 7-2 ist horizontal in die Ebenen RDF und RDFS gegliedert. Der untere Bereich der Abbildung stellt das soeben eingeführte Beispiel eines DWH-Systems als RDF-Graph dar. Oberhalb ist der zugehörige Ausschnitt des RDFS für DWH-Metadaten aufgezeigt. Anhand vertikaler Verbindungslinien sind die jeweiligen Typbeziehungen zwischen den Instanzen im RDF und den Klassen bzw. Eigenschaftstypen des DWH-RDFS visualisiert. Quellcode 3 beinhaltet das zugehörige RDF-Dokument in RDF/XML-Notation.

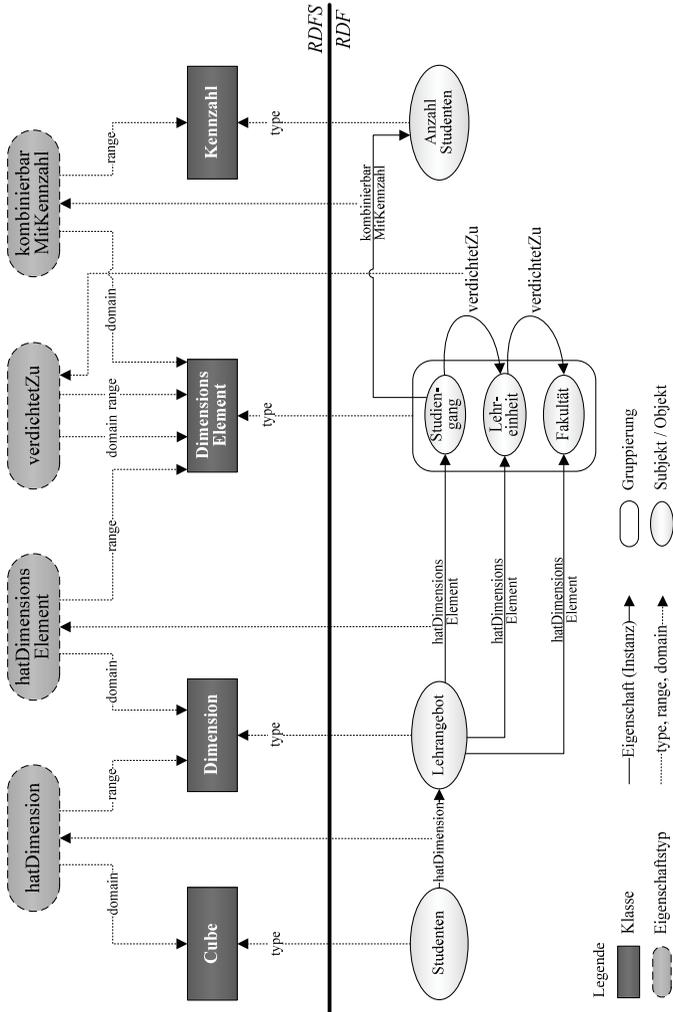


Abbildung 7-2: RDF- und RDFS-Ebene

```

<?xml version="1.0"?>
<!-- Namespace-Deklaration -->
<rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dwhrdfs="http://www.semantic-metadata.com/dwhrdfs/dwhrdfs#">

<!-- Allgemeine Dokumentation des RDF-Vokabulars -->
  <rdf:Description
    rdf:about="http://www.semantic-metadata.com/dwhrdfs/dwhrdfs#">
    <dc:title>RDF-Schema zur semantischen Metadaten-Beschreibung
      von Data-Warehouse-Systemen</dc:title>
    <dc:description>Dieses RDF-Schema stellt Klassen und
      Eigenschaftstypen fuer eine semantische
      Metadaten-Beschreibung von Data-Warehouse-Systemen dar.
    </dc:description>
    <dc:creator>Stefan Hartmann</dc:creator>
    <dc:date>08/06/2008</dc:date>
  </rdf:Description>

<!-- Definition der Klassen -->
  <rdfs:Class rdf:ID="Cube">
    <rdfs:label>Cube</rdfs:label>
    <rdfs:comment>Ein Cube dient der Beschreibung von mehr-
      dimensionalen Sichten auf ein DWH.</rdfs:comment>
  </rdfs:Class>
  <rdfs:Class rdf:ID="Dimension">
    <rdfs:label>Dimension</rdfs:label>
    <rdfs:comment>Mithilfe dieser Klasse koennen Dimensionen
      dargestellt werden.</rdfs:comment>
  </rdfs:Class>
  <rdfs:Class rdf:ID="Kennzahl">
    <rdfs:label>Kennzahl</rdfs:label>
    <rdfs:comment>Kennzahlen sind quantitative Objekte
      eines DWH.</rdfs:comment>
  </rdfs:Class>
  <rdfs:Class rdf:ID="BasisKennzahl">
    <rdfs:label>Basis-Kennzahl</rdfs:label>
    <rdfs:comment>Eine Basis-Kennzahl ist eine Kennzahl, die
      direkt aus den Fakten bereitgestellt wird.</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#Kennzahl"/>
  </rdfs:Class>

```

```
<!-- Definition der Eigenschaftstypen -->
<rdf:Property rdf:ID="hatDimension">
  <rdfs:label>hat Dimension</rdfs:label>
  <rdfs:comment>Eine Verknuepfung zwischen Cube und Dimension wird
    mithilfe dieses Eigenschaftstyps modelliert.
</rdfs:comment>
  <rdfs:domain rdf:resource="#Cube"/>
  <rdfs:range rdf:resource="#Dimension"/>
</rdf:Property>
<rdf:Property rdf:ID="verdichtetZu">
  <rdfs:label>verdichtet Zu</rdfs:label>
  <rdfs:comment>Ein Dimensionselement einer niedrigeren Stufe
    kann zu einem Dimensionselement einer hoeheren
    Stufe aggregiert werden.</rdfs:comment>
  <rdfs:range rdf:resource="#DimensionsElement"/>
  <rdfs:domain rdf:resource="#DimensionsElement"/>
</rdf:Property>
<rdf:Property rdf:ID="kombinierbarMitKennzahl">
  <rdfs:label>kombinierbar mit Kennzahl</rdfs:label>
  <rdfs:comment>Die Abbildung logisch und fachlich gueltiger
    Verknuepfungen zwischen Dimensionselementen und Kennzahlen
    erfolgt mit diesem Eigenschaftstyp.</rdfs:comment>
  <rdfs:domain rdf:resource="#DimensionsElement"/>
  <rdfs:range rdf:resource="#Kennzahl"/>
</rdf:Property>
</rdf:RDF>
```

Quellcode 2: RDF-Schema für DWH-Metadaten (Ausschnitt)

```

<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dwhrdfs="http://www.semantic-metadata.com/dwhrdfs/dwhrdfs#"
  xml:base="http://www.semantic-metadata.com/dwhrdfs/example">

  <dwhrdfs:Cube rdf:ID="Studenten">
    <rdfs:label>Studenten-Cube</rdfs:label>
    <rdfs:comment>Studentendaten einer Hochschule.</rdfs:comment>
    <dwhrdfs:hatDimension rdf:resource="#Lehrangebot"/>
  </dwhrdfs:Cube>

  <dwhrdfs:Dimension rdf:ID="Lehrangebot">
    <rdfs:label>Dimension Lehrangebot</rdfs:label>
    <rdfs:comment>Diese Dimension umfasst Merkmale aus dem
      Bereich Lehrangebot.</rdfs:comment>
    <dwhrdfs:hatDimensionsElement rdf:resource="#Studiengang"/>
    <dwhrdfs:hatDimensionsElement rdf:resource="#Lehreinheit"/>
    <dwhrdfs:hatDimensionsElement rdf:resource="#Fakultaet"/>
  </dwhrdfs:Dimension>

  <dwhrdfs:DimensionsElement rdf:ID="Studiengang">
    <rdfs:label>Studiengang</rdfs:label>
    <dwhrdfs:verdichtetZu rdf:resource="#Lehreinheit"/>
    <dwhrdfs:kombinierbarMitKennzahl
      rdf:resource="#AnzahlStudenten"/>
  </dwhrdfs:DimensionsElement>

  <dwhrdfs:DimensionsElement rdf:ID="Lehreinheit">
    <rdfs:label>Lehreinheit</rdfs:label>
    <dwhrdfs:verdichtetZu rdf:resource="#Fakultaet"/>
  </dwhrdfs:DimensionsElement>

  <dwhrdfs:DimensionsElement rdf:ID="Fakultaet">
    <rdfs:label>Fakultaet</rdfs:label>
  </dwhrdfs:DimensionsElement>

  <dwhrdfs:BasisKennzahl rdf:ID="AnzahlStudenten">
    <rdfs:label>Anzahl Studenten</rdfs:label>
    <rdfs:comment>Diese Basis kennzahl gibt die Gesamtzahl
      der Studenten aus.</rdfs:comment>
  </dwhrdfs:BasisKennzahl>
</rdf:RDF>

```

Quellcode 3: RDF-Dokument zu einem DWH-Schema (Ausschnitt)

ZUSAMMENFASSUNG UND FAZIT

Im vorliegenden Abschnitt wurde ein RDF-Vokabular zur semantischen Beschreibung von DWH-Metadaten vorgestellt und detailliert erläutert. Als Grundlage für die Entwicklung des DWH-RDFS dienten dabei die Ergebnisse der vorausgegangenen Untersuchung hinsichtlich der Potentiale von RDF/RDFS zur Erfüllung der aufgestellten Anforderungen an eine semantisch reichhaltige Beschreibung von DWH-Metadaten.

Das vollständig ausgestaltete DWH-RDFS steht als Vorlage zur DWH-Modellierung unter dem oben genannten Namensraum frei zur Verfügung. Die den Klassen und Eigenschaftstypen des DWH-RDFS zugeordneten Beschreibungselemente sowie die semantisch erfassten Beziehungen zwischen den Klassen ermöglichen einem Nutzer, den Zweck eines DWH-RDFS-Elements klar zu erfassen und dieses korrekt zu verwenden. Die Erstellung eines RDF-Dokuments auf Basis des DWH-RDFS kann durch den Einsatz von (RDF-)Editoren unterstützt werden. Mithilfe von Parsern (vgl. Abschnitt 6.2.3) sind die RDF-Dokumente anschließend auf Validität gegen das DWH-RDFS zu prüfen. Die generelle Erweiterbarkeit eines RDFS erlaubt es, in diesem Abschnitt nicht beachtete Aspekte wie z. B. Beschreibung des Berichtswesens, Angaben zur Zeitplanung von Ladeprozessen oder Erfassung von Datenqualitätskennzeichen durch Definition weiterer Klassen und Eigenschaftstypen zu berücksichtigen. Eine Erweiterung des DWH-RDFS ist dabei durch Angabe einer neuen Versionsnummer und der vorgenommenen Veränderung zu kennzeichnen.

Ein explizit herauszustellender Vorteil der RDF-basierten Beschreibung eines DWH-Systems liegt in der Möglichkeit, valide RDF-Dokumente – die jeweils unterschiedliche DWH-Systeme beschreiben – hinsichtlich struktureller und semantischer Aspekte zueinander in Beziehung setzen und miteinander vergleichen zu können. Dies ist aufgrund deren Konformität mit dem DWH-RDFS realisierbar. Die auch auf Instanzebene verfügbaren Beschreibungselemente (z. B. *rdfs:comment*, *dc:description*) bilden hierfür einen wichtigen Baustein. Die in ihrer Bedeutung eindeutig charakterisierten Beschreibungselemente können – in Kombination mit der bestehenden Werkzeugunterstützung zur intuitiven Erstellung valider RDF-Dokumente – als ein Anreiz für die Kodifizierung impliziten Wissens (z. B. eines Modellierers oder Fachexperten) interpretiert werden. Die formale Semantik von RDF unterstützt dabei die eindeutige Erfassung von Aussagen. Die Konklusion dieses Abschnitts ergibt daher, dass mithilfe des DWH-RDFS eine semantisch homogene Beschreibung von DWH-Systemen realisierbar ist.

7.3 Eine OWL-Ontologie zur semantisch reichhaltigen Beschreibung von Data-Warehouse-Metadaten

Auf den im DWH-RDFS vorgestellten Klassen und Eigenschaftstypen aufbauend wird in diesem Abschnitt eine Ontologie erarbeitet, die über zusätzliche Beschreibungskonstrukte verfügt, um im Vergleich zur RDF-basierten Lösung eine semantisch ausdrucksstärkere Darstellung von DWH-Metadaten zu ermöglichen. Für die Konstruktion eines entsprechenden ontologiebasierten Vokabulars soll hierbei OWL DL eingesetzt werden. Die semantische Mächtigkeit von OWL DL bei gleichzeitiger Entscheidbarkeit begründet diese Entscheidung (vgl. Abschnitt 6.3.3).

Einführend werden allgemeine Kriterien für die Entwicklung der OWL-Ontologie für DWH-Metadaten (DWH-OWL) diskutiert. Dem folgt eine Untersuchung der Potentiale von OWL, um die Klassen und Eigenschaftstypen des DWH-RDFS durch eine OWL-basierte Beschreibung semantisch reichhaltiger zu gestalten. Abschließend wird die resultierende DWH-OWL detailliert vorgestellt.

7.3.1 Allgemeine Entwicklungskriterien einer Ontologie für Data-Warehouse-Metadaten

Die Entwicklung der DWH-OWL folgt dem – im Rahmen der Entwicklungskriterien für Ontologien skizzierten – *deduktiven Ansatz* (vgl. Abschnitt 6.3.1). Dementsprechend soll diese lediglich allgemeingültige Konzepte zur Beschreibung eines DWH-Systems umfassen, die als Vorlage z. B. für die Beschreibung spezieller Architekturvarianten von DWH-Systemen (vgl. Abschnitt 2.3.6) dienen können. Gemäß der Klassifizierung nach GUARINO entsteht mit der DWH-OWL eine so genannte *Top-Level-Ontologie* (vgl. Abschnitt 6.3.1). Top-Level-Ontologien weisen in der Regel nur schwache ontologische Ausprägungen auf. Sie dienen vornehmlich der Modellierung. Ein direkter Mehrwert durch Inferenz steht nicht im Fokus.

Bei der Ausgestaltung einer Top-Level-Ontologie ist daher abzuwägen, in welchem Maße ontologische Beschreibungselemente zur Einschränkung der jeweiligen Klassen und Rollen eingesetzt werden. Der IRDS-Standard hat gezeigt, dass zu harte Restriktionen auf oberster Abstraktionsebene die Akzeptanz des Ansatzes schmälern können (vgl. Abschnitt 3.4.1). Für die Elemente der zu definierenden Top-Level-Ontologie soll daher insbesondere der Grundsatz der minimalen ontologischen Bindung gelten (vgl. Abschnitt 6.3.1) und nur die für die Beschreibung eines DWH-Systems zwingend notwen-

digen Elemente als verbindlich deklariert werden. Auf diese Weise bleibt es einem Modellierer zu einem bestimmten Grad freigestellt, wie umfassend er ein DWH-System erfasst (z. B. mit oder ohne Beschreibung von System- und Datenbankinformationen).

Die Forderungen nach Klarheit, Kohärenz und Erweiterbarkeit einer Ontologie (vgl. Abschnitt 6.3.1) sind auch von der DWH-OWL zu erfüllen. Einen grundlegenden Beitrag hierzu kann die dem DWH-RDFS zugrunde liegende Kategorisierung in allgemeine Beschreibungselemente, DWH-Kernelemente, DWH-Organisationselemente, DWH-Kontextelemente und DWH-Systemelemente leisten. Diese soll daher für die Entwicklung der DWH-OWL übernommen werden. In Bezug auf Erweiterbarkeit ist zu vermerken, dass OWL – wie auch RDFS – von technologischer Seite her beliebig um Klassen und Rollen erweiterbar ist. Außerdem soll durch eine möglichst allgemeingültige Beschreibung der Klassen und Rollen sichergestellt werden, dass zukünftige Erweiterungen der DWH-OWL ohne Änderung bestehender Definitionen erfolgen können.

7.3.2 Potentiale von OWL zur semantisch reichhaltigen Beschreibung von Data-Warehouse-Metadaten

Eine OWL-Ontologie besteht im Wesentlichen aus Klassen und Eigenschaftstypen (Rollen) wie sie auch bei RDF/RDFS verwendet werden. Darüber hinaus besteht in OWL die Möglichkeit, Klassen und Rollen zueinander in komplexe Beziehungen zu setzen [HKRS08, S. 126]. Klassen und Rollen der OWL stellen zudem Unterklassen ihrer Entsprechung im RDF/RDFS dar (vgl. Abschnitt 6.3.3). Aufgrund dieser Zusammenhänge ist zu folgern, dass die in Abschnitt 7.1 identifizierten Potentiale von RDF/RDFS zur semantischen Beschreibung von DWH-Metadaten grundlegend auch beim Einsatz von OWL realisierbar sind.

Die Klassen und Eigenschaftstypen des DWH-RDFS sowie ihre Bezeichnung und Beschreibung können somit in einem ersten Schritt direkt in Klassen und Rollen der DWH-OWL überführt werden. Dabei wird nahe gelegt, für deren Identifikation dieselben sprechenden Namen wie im DWH-RDFS einzusetzen.

ALLGEMEINE BESCHREIBUNGSELEMENTE

Die semantische Spezifikation von Klassen und Rollen in der DWH-OWL kann mithilfe der RDFS-eigenen Beschreibungselemente (z. B. *rdfs:comment*) sowie den Vokabeln des DC-Standards (z. B. *dc:description*) erfolgen. Des Weiteren sind analog zum DWH-RDFS Unterklasse- (*rdfs:subClassOf*) sowie Untereigenschaftsbeziehungen (*rdfs:subPropertyOf*) einsetzbar. Zur Definition gültiger Kombinationen von Klassen und Rollen können in OWL ebenfalls die RDFS-Elemente *rdfs:domain* und *rdfs:range* verwendet werden.

Die im DWH-RDFS speziell definierten Eigenschaftstypen zur Versionierung (*istVersion*, *umfasstAenderung*) sind in der OWL-basierten Beschreibung nicht erforderlich. Es kann auf die von OWL offerierten Beschreibungselemente für Versionsangaben zurückgegriffen werden (vgl. Abschnitt 6.3.3). Die Angabe des Datums einer Änderung kann wie im DWH-RDFS mittels des DC-Terms *dcterms:modified* erfolgen.

SPEZIFISCHE BESCHREIBUNGSELEMENTE DER OWL DL

Das semantische Potential von OWL im Vergleich zu RDF wurde in Abschnitt 6.3.3 herausgestellt. Diejenigen Sprachkonstrukte der OWL DL, die speziell in der DWH-OWL eingesetzt werden sollen, sind nachstehend hervorgehoben:

- Zu jeder in der DWH-OWL definierten Rolle soll eine jeweils zugehörige, *inverse Rolle* angelegt und als solche ausgezeichnet werden. OWL DL bietet dafür das Sprachelement *owl:inverseOf* an. Dieses ermöglicht die Abbildung derselben Relation zwischen zwei Klassen, jedoch mit vertauschten Argumenten. Dadurch kann bspw. anhand einer Beziehung von einer Dimension zu einem Dimensionselement auch vice versa von einem Dimensionselement auf die Dimension(en) geschlossen werden, in der (denen) es referenziert wird.
- Um die Kombinationsmöglichkeiten zwischen Klassen einzuschränken bzw. die Verwendung einer Rolle als verbindlich vorzuschreiben, soll die DWH-OWL auch Kardinalitätsrestriktionen beinhalten. OWL DL bietet dafür durch die Sprachkonstrukte *owl:minCardinality*, *owl:maxCardinality* und *owl:cardinality* Unterstützung. Konstitutive Elemente der Kategorie DWH-Kernelemente sollen dabei als verpflichtend gekennzeichnet werden. Um die Gesamt-Ontologie jedoch nicht zu restriktiv zu gestalten, sind Verknüpfungen – ausgehend von der Kategorie DWH-Kernelemente – zu Klassen der anderen Kategorien als optional

auszuweisen. Die zugehörigen inversen Rollen können allerdings als verbindlich eingestuft sein, da bspw. zu einem Cube nicht zwingend ein Fachverantwortlicher angegeben werden muss, ein Fachverantwortlicher jedoch für ein bestimmtes Objekt zuständig sein muss und daher nicht ohne Zuweisung bestehen darf.

- Die vom RDF bekannten Rollenrestriktionen (*rdfs:domain*, *rdfs:range*) können in OWL durch so genannte *Quantoren* ergänzt werden. Mithilfe eines *Allquantors* wird für eine Aussage gefordert, dass eine zu einem Subjekt zugehörige Rolle nur Objekte einer spezifischen Klasse besitzen darf (*owl:allValuesFrom*). Der *Existenzquantor* insistiert zusätzlich, dass zu einem Subjekt über eine bestimmte Rolle mindestens ein Objekt der in der Restriktion genannten Klasse zugewiesen wird (*owl:someValuesFrom*). Quantorenkonstrukte können mit Ausdrücken für logische Verknüpfung (Konjunktion *owl:intersectionOf*; Disjunktion *owl:unionOf*) kombiniert werden. Somit kann z. B. für einen Fachverantwortlichen über die Rolle *hatFachverantwortlichen* gefordert werden, dass diesem mindestens ein Objekt (*owl:someValuesFrom*) der Klassen *Cube*, *DimensionsElement*, *Kennzahl* oder *System* (*owl:unionOf*) zugewiesen sein muss.
- Mithilfe des Ausdrucks *owl:disjointWith* wird die Disjunktheit von Klassen beschrieben. Dieses Sprachelement soll in der DWH-OWL genutzt werden, um für spezifische Klassen sicherzustellen, dass ein zugehöriges Individuum nicht zugleich zu einer andern Klasse zählt. Dies betrifft z. B. die Klassen *Cube* und *Dimension*. Dagegen kann z. B. ein Berichtsverantwortlicher (Klasse *BerichtsVerantwortlicher*) gleichzeitig auch fachlich für ein Objekt zuständig sein (Klasse *FachVerantwortlicher*) und somit diesen beiden Klassen zugeordnet sein. Die Einschränkung auf Disjunktheit ist daher für diese Klassen nicht vorzunehmen.
- Die von RDF bereitgestellten Elemente zur Abbildung der Unterklasse- und Untereigenschaftsbeziehung sind transitiv. OWL DL ermöglicht darüber hinaus selbstdefinierte, thematische Rollen ebenfalls als transitiv zu kennzeichnen (*owl:TransitiveProperty*). Im Hinblick auf die Ausgestaltung der DWH-OWL kann dieses Sprachelement z. B. bei der Rolle *verdichtetZu*, die zwischen zwei Dimensionselementen gültig ist, angewandt werden.

Die Erläuterungen und Erkenntnisse dieses Abschnitts dienen als Ansatzpunkt für die Konstruktion einer OWL-Ontologie zur semantisch reichhaltigen Beschreibung von DWH-Metadaten.

7.3.3 Aufbau der OWL-Ontologie zur semantisch reichhaltigen Beschreibung von Data-Warehouse-Metadaten

Im vorangegangenen Abschnitt wurde aufgezeigt, dass die Grundstruktur des DWH-RDFS (bestehend aus Klassen und Eigenschaftstypen) direkt in eine OWL-basierte Beschreibung überführt werden kann (DWH-OWL). Die allgemeinen Beschreibungselemente (z. B. *rdfs:label*, *rdfs:comment*) können dabei ebenfalls eingesetzt werden. Angaben zur Versionierung der Ontologie erfolgen über OWL-eigene Sprachelemente (z. B. *owl:versionInfo*).

Auf dieser Grundstruktur aufbauend werden die Klassen und Rollen der DWH-OWL mithilfe der soeben herausgestellten OWL-DL-spezifischen Sprachkonstrukte angereichert. Abgesehen von der Disjunktheit von Klassen betrifft die semantische Erweiterung vor allem die in der DWH-OWL definierten Rollen. Die unten angeführten Tabellen zur detaillierten Erläuterung der DWH-OWL veranschaulichen daher insbesondere die Spezifikation der Rollen, deren zugehörige inverse Rollen sowie die jeweiligen Kardinalitätsvorgaben¹⁰⁴ (Angaben in Klammern).

Die im DWH-RDFS spezifizierten Domänen- und Wertebereichseinschränkungen werden dabei eins zu eins auf die Rollen der DWH-OWL übertragen. Diese sind den beiden mittleren Tabellenspalten zu entnehmen. Mit Blick auf die inversen Rollen gelten Domänen- und Wertebereich in vertauschter Anordnung. Darüber hinaus werden die Domänen- und Wertebereichseinschränkungen durch den Einsatz von Allquantoren bzw. Existenzquantoren bei den Klassen ergänzt. Hierdurch können zugleich wesentliche Kardinalitätsangaben abgedeckt werden (Allquantor: 0,*; Existenzquantor: 1,*). Um für einen Wertebereich mehr als eine gültige Klasse referenzieren zu können, wird auf das Sprachelement der Disjunktion zurückgegriffen.

Zusätzlich in Klammern ausgewiesene Rollen benennen die jeweils direkt übergeordnete Rolle (Superrolle). Transitive Rollen sind durch den Zusatz *transitiv* gekennzeichnet.

¹⁰⁴ Die Angabe einer Kardinalität erfolgt hier in der (min, max)-Notation (vgl. z. B. [FeSi06, S. 140 ff.]).

Die weiteren Ausführungen sowie die Tabellen zur Vorstellung der DWH-OWL sind nach den vier eingeführten Kategorien für DWH-Elemente gruppiert. Textliche Beschreibungen zu den Tabellen heben dabei nur die Besonderheiten in der jeweiligen Metadaten-Kategorie hervor. Im Übrigen gelten die obigen Ausführungen in gleicher Weise.

Auf die explizite Darlegung der Disjunktheit von DWH-OWL-Klassen wurde an dieser Stelle aufgrund der relativ hohen Anzahl an Einschränkungskombinationen verzichtet.

DATA-WAREHOUSE-KERNELEMENTE

In dieser Metadatenkategorie sollen zunächst die Kardinalitätsangaben der beiden Rollen *zugehoerigerCube* und *zugehoerigeDimension* (Tabellenspalte *inverse Rolle*) begründet werden. Die jeweilige Festlegung der Kardinalität 1,* basiert auf der Annahme, dass Dimensionselemente in (unterschiedlichen) Dimensionen unterschiedlicher Cubes verwendet werden können. Exemplarisch sei an dieser Stelle ein mögliches Dimensionselement *Monat* genannt, das – einmal definiert – in den Zeitdimensionen mehrerer Cubes referenziert werden kann. Daneben ist die Zuweisung eines Dimensionselements zu einer Dimension obligatorisch. Selbiger Sachverhalt ist analog auf die Rolle *zugehoerigerCube* zu übertragen.

Des Weiteren ist die Verknüpfung eines Dimensionselements mit einer Kennzahl als optional ausgewiesen (*kombinierbarMitKennzahl*; 0,*). Dies soll insbesondere die Möglichkeit offen halten, Dimensionselemente bspw. nur für ein so genanntes Stammdaten-Reporting einsetzen zu können. Die Verwendung der zugehörigen inversen Rolle ist dagegen verpflichtend, da eine Kennzahl mindestens einen Bezug zu einem quantitativen Merkmal besitzen muss. Angaben zur Semiaggregierbarkeit einer Kennzahl (*istSemiaggregierbar*) sind generell optional.

Eine Besonderheit tritt auch bei den Rollen *hatOperand1*, *hatOperand2* und *hatKonstante* auf. Diese dürfen zur Definition einer Berechnung jeweils nur einmal verwendet werden. Gleichzeitig soll aber eine einzelne Berechnung nur aus zwei Operanden bestehen. Folgende Nebenbedingung (NB) gewährleistet diese Restriktion:

```
((hatOperand1 exactly 1) AND (hatOperand2 exactly 1)) OR
((hatOperand1 exactly 1) AND (hatKonstante exactly 1)) OR
((hatKonstante exactly 1) AND (hatOperand2 exactly 1))
```

Der Ausdruck *exactly* entspricht dabei dem OWL-Sprachelement *owl:cardinality*. Die Rollen *hatKonstante* und *hatOperand* erwarten – wie im DWH-RDFS – die Objektwerte als Literal. Diejenigen Rollen, denen literale Objektwerte zugewiesen werden können, sind in der DWH-OWL als so genannte konkrete Rollen mithilfe des Sprachkonstrukts *owl:DatatypeProperty* definiert (vgl. Abschnitt 6.3.3).

Abschließend ist hervorzuheben, dass die Rollen *verdichtetZu* und *detailliertZu* als transitiv gekennzeichnet sind.

Exemplarisch für die weiteren Metadaten-Kategorien visualisiert Abbildung 7-3 die Klassen der Kategorie DWH-Kernelemente mit ihren zulässigen Beziehungen über die entsprechenden Rollen. Die Abbildung weist auch die zugehörigen inversen Rollen (kursiv dargestellte Rollenbezeichnung) sowie die Kardinalitätsangaben mit Nebenbedingung (NB) auf.

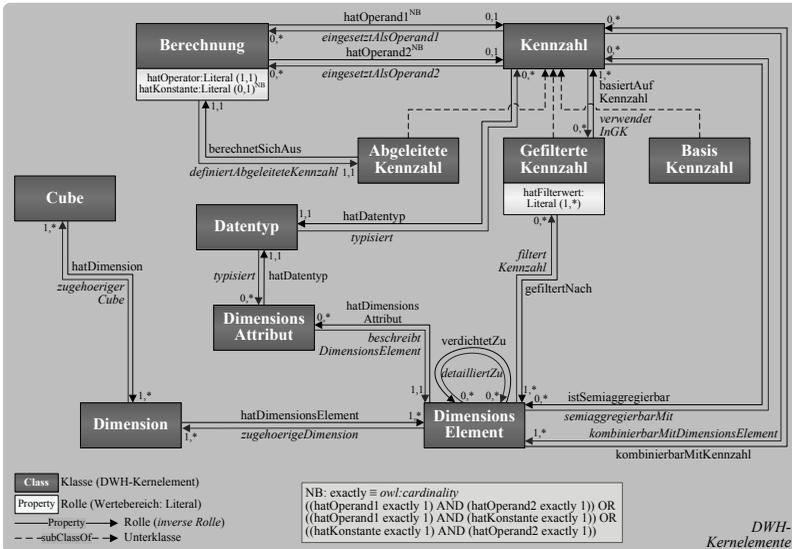


Abbildung 7-3: OWL-Ontologie für DWH-Kernelemente

Eine graphische Darstellung der integrierten Gesamtsicht der DWH-OWL erfolgt nicht. Dies wird zum einen mit der Ähnlichkeit im Aufbau zum DWH-RDFS begründet und zum anderen ist eine klare Lesbarkeit und Verständlichkeit aufgrund der doppelten Anzahl an Rollen (inverser Rollen) und der Kardinalitätsangaben nicht gewährleistet. Stattdessen wird auf die jeweilige tabellarische Beschreibung einer Kategorie von DWH-Elementen in diesem Abschnitt verwiesen. Nachstehende Tabelle gibt einen entsprechenden Überblick über die Rollen in der Kategorie DWH-Kernelemente.

| Rolle | Domäne | Wertebereich | Inverse Rolle |
|---------------------------------------|--|--------------------|---|
| hatDimension (1,*) | Cube | Dimension | zugehoerigerCube (1,*) |
| hatDimensionsElement (1,*) | Dimension | DimensionsElement | zugehoerige Dimension (1,*) |
| verdichtetZu (0,*) - <i>transitiv</i> | DimensionsElement | DimensionsElement | detailliertZu (0,*) - <i>transitiv</i> |
| hatDimensionsAttribut (0,*) | DimensionsElement | DimensionsAttribut | beschreibt DimensionsElement (1,1) |
| kombinierbar MitKennzahl (0,*) | DimensionsElement | Kennzahl | kombinierbarMit DimensionsElement (1,*) |
| istSemiaggregierbar (0,*) | Kennzahl | DimensionsElement | semiaggregierbarMit (0,*) |
| gefiltertNach (1,*) | GefilterteKennzahl | DimensionsElement | filtertKennzahl (0,*) |
| basiertAufKennzahl (1,*) | GefilterteKennzahl | Kennzahl | verwendetIn GefilterterKennzahl (0,*) |
| hatFilterwert (1,*) | GefilterteKennzahl | LITERAL | % |
| berechnetSichAus (1,1) | AbgeleiteteKennzahl | Berechnung | definiertAbgeleitete Kennzahl (1,1) |
| hatOperand1 (0,1) | Berechnung ^{NB} ^{NB} max. 2 Operanden | Kennzahl | eingesetztAls Operand1 (0,*) |
| hatOperand2 (0,1) | Berechnung ^{NB} | Kennzahl | eingesetztAls Operand2 (0,*) |
| hatKonstante (0,1) | Berechnung ^{NB} | LITERAL | % |
| hatOperator (1,1) | Berechnung | LITERAL | % |
| hatDatentyp (1,1) | DimensionsAttribut Kennzahl | Datentyp | typisiert (0,*) |

Tabelle 12: Rollen in der Kategorie DWH-Kernelemente

DATA-WAREHOUSE-ORGANISATIONSELEMENTE

Elemente der Kategorie DWH-Organisationselemente sind für die Modellierung eines DWH-Systems grundsätzlich nicht verbindlich vorgeschrieben (0,*). Der Einsatz der als invers deklarierten Rollen kann jedoch – gemäß der in Abschnitt 7.3.2 exemplarisch gegebenen Begründung – verpflichtend sein (1,*).

Eine Ausnahme bildet lediglich die Rolle *zugriffsberechtigtAuf* (0,*). Eine Organisationseinheit kann auch ohne Zugriffsrecht auf ein Objekt der Klassen *Cube*, *DimensionsElement* oder *Kennzahl* existieren. Dies liegt u. a. in der Tatsache begründet, dass die Klassen *Datenproduzent* und *Datenkonsument* jeweils Unterklassen der Klasse *Organisationseinheit* darstellen und somit bei diesen ebenfalls die Rolle *zugriffsberechtigtAuf* eingesetzt werden kann. Ein Datenproduzent muss jedoch nicht zwingend Zugriff auf die Daten in einem DWH besitzen.

Tabelle 13 beinhaltet eine vollständige Darstellung der Rollen in dieser Kategorie.

| Rolle | Domäne | Wertebereich | Inverse Rolle |
|--|---|--------------------------------|---|
| zugriffsberechtigte Organisationseinheit (0,*) | Cube DimensionsElement Kennzahl | Organisations- einheit | zugriffsberechtigtAuf (0,*) |
| hatDatenproduzent (0,*) | DimensionsElement Kennzahl | Datenproduzent | produziertDatenFuer (1,*) |
| hatDatenkonsument (0,*) | DimensionsElement Kennzahl | Datenkonsument | nutztDatenVon (1,*) |
| hatBerichts Verantwortlichen (0,*) | Cube | Berichts Verantwortlicher | verantwortlichFuer Berichtswesen (1,*) |
| hatTechnischen Ansprechpartner (0,*) | System | Technischer Ansprechpartner | technischZustaendigFuer (1,*) |
| hatFachVerantwortlichen (0,*) | System Cube DimensionsElement Kennzahl | Fach Verantwortlicher | fachlichZustaendigFuer (1,*) |

Tabelle 13: Rollen in der Kategorie DWH-Organisationselemente

DATA-WAREHOUSE-KONTEXTELEMENTE

Vergleichbar mit den Ausführungen der Kategorie DWH-Organisationselemente gilt auch die Verwendung der Klassen und Rollen dieser Kategorie als optional (0,*). Jedoch muss ein bestimmter Kontexttyp mindestens einem Objekt der Klassen *Cube*, *DimensionsElement* oder *Kennzahl* zugewiesen sein (inverse Rolle: 1,*). Die speziell zu den einzelnen Kontexttypen angelegten Rollen (z. B. *besitztAufgabenKontext*) sind als Untereigenschaft zu *besitztKontext* definiert. Analog ist dies auf die zugehörigen inversen Rollen zu übertragen. Die nachfolgende Tabelle veranschaulicht diesen Sachverhalt.

| Rolle | Domäne | Wertebereich | Inverse Rolle |
|---|---------------------------------------|-----------------|---------------------------------------|
| besitztKontext (0,*) | Cube DimensionsElement Kennzahl | Kontext | kontextZu (1,*) |
| besitztAufgabenKontext (0,*) - (besitztKontext) | Cube DimensionsElement Kennzahl | AufgabenKontext | aufgabenKontextZu (1,*) - (kontextZu) |
| besitztNutzerKontext (0,*) - (besitztKontext) | Cube DimensionsElement Kennzahl | NutzerKontext | nutzerKontextZu (1,*) - (kontextZu) |
| besitztDomaenenKontext (0,*) - (besitztKontext) | Cube DimensionsElement Kennzahl | DomaenenKontext | domaenenKontextZu (1,*) - (kontextZu) |
| besitztUmweltKontext (0,*) - (besitztKontext) | Cube DimensionsElement Kennzahl | UmweltKontext | umweltKontextZu (1,*) - (kontextZu) |

Tabelle 14: Rollen in der Kategorie DWH-Kontextelemente

DATA-WAREHOUSE-SYSTEMELEMENTE

Die Verwendung der Elemente dieser Metadaten-Kategorie ist einem Nutzer der DWH-OWL vollständig freigestellt. Rollen und inverse Rollen besitzen durchgängig die Kardinalität 0,* . Hierdurch ist ein hoher Freiheitsgrad bei der Modellierung system- oder datenbankspezifischer Merkmale geboten.

Die Rolle *basiertAufSystem* stellt die Superrolle für die spezifischen Rollen *basiertAufVorsystem* und *basiertAufDWHSsystem* dar. Diese Ausführung gilt entsprechend für die zugehörige inverse Rolle *systemFuer*. Des Weiteren sind die beiden Rollen *entstammtQuellattribut* und *wirdTransformiertAuf* als transitiv definiert.

Einen Überblick über die Rollen in dieser Kategorie bietet Tabelle 15.

| Rolle | Domäne | Wertebereich | Inverse Rolle |
|--|---------------------------------------|------------------------|--|
| basiertAufSystem (0,*) | Cube DimensionsElement Kennzahl | System | systemFuer (0,*) |
| basiertAufVorsystem (0,*) - (basiertAufSystem) | Cube DimensionsElement Kennzahl | Vorsystem | vorsystemFuer (0,*) - (systemFuer) |
| basiertAufDWHSystem (0,*) - (basiertAufSystem) | Cube DimensionsElement Kennzahl | DWHSystem | dwhsystemFuer (0,*) - (systemFuer) |
| basiertAufHardware (0,*) | System | LITERAL | % |
| eingesetzteSoftware (0,*) | System | Software | laeuftAufSystem (0,*) |
| angebotenDurch SoftwareHersteller (0,*) | Software | SoftwareHersteller | angeboteneSoftware (0,*) |
| erstelltSoftware Komponente (0,*) | SoftwareHersteller | Software Komponente | bereitgestelltDurch SoftwareHersteller (0,*) |
| hatSoftwareVersion (0,*) | SoftwareKomponente | LITERAL | % |
| verwaltetDatenbank (0,*) | Software | Datenbank | verwaltetDurchSoftware (0,*) |
| basiertAufDatenbank (0,*) | Cube DimensionsElement Kennzahl | Datenbank | datenbankFuer (0,*) |
| basiertAufTabelle (0,*) | Cube DimensionsElement Kennzahl | Tabelle | tabelleFuer (0,*) |
| basiertAuf Tabellenattribut (0,*) | Cube DimensionsElement Kennzahl | Tabellenattribut | tabellenattributFuer (0,*) |
| beinhaltetTabelle (0,*) | Datenbank | Tabelle | zugehoerigeDatenbank (0,*) |
| hatTabellenattribut (0,*) | Tabelle | Tabellenattribut | zugehoerigeTabelle (0,*) |
| entstammtQuellattribut (0,*) - <i>transitiv</i> | Tabellenattribut | Tabellenattribut | wirdTransformiertAuf (0,*) - <i>transitiv</i> |
| wirdTransformiert (0,*) | Tabellenattribut | LITERAL | % |

Tabelle 15: Rollen in der Kategorie DWH-Systemelemente

QUELLTEXTAUSZUG DER DWH-OWL

Das vollständige Vokabular der DWH-OWL ist über folgende URL abrufbar:

```
xmlns:dwhowl="http://www.semantic-metadata.com/dwhowl/dwhowl#"
```

Zur Veranschaulichung ist nachstehend ein Ausschnitt der DWH-OWL in OWL-RDF-Syntax abgedruckt:

```
<?xml version="1.0"?>
<!-- Namespace-Deklaration -->
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:dwhowl="http://www.semantic-metadata.com/dwhowl/dwhowl#">
<!-- Allgemeine Dokumentation der Ontologie -->
  <owl:Ontology rdf:about="">
    <dc:creator>Stefan Hartmann</dc:creator>
    <dc:title>Eine Ontologie zur semantisch reichhaltigen Metadaten-
      Beschreibung von Data-Warehouse-Systemen</dc:title>
    <owl:versionInfo>1.0</owl:versionInfo>
    <dc:date>29/03/2008</dc:date>
    <dc:description>Diese OWL-Ontologie stellt Klassen und Rollen
      fuer eine semantisch reichhaltige Metadaten-Beschreibung
      von DWH-Systemen bereit.
    </dc:description>
  </owl:Ontology>
<!-- Definition der Klassen -->
  <owl:Class rdf:about="#Cube">
    <rdfs:label>Cube</rdfs:label>
    <rdfs:comment>Ein Cube dient der Beschreibung von
      mehrdimensionalen Sichten auf ein DWH.</rdfs:comment>
    <owl:disjointWith rdf:resource="#Dimension"/>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="hatDimension"/>
        </owl:onProperty>
        <owl:someValuesFrom rdf:resource="#Dimension"/>
      </owl:Restriction>
    </rdfs:subClassOf>
  </owl:Class>
```

```

<owl:Class rdf:ID="Dimension">
  <rdfs:label>Dimension</rdfs:label>
  <rdfs:comment>Ein Cube wird durch Dimensionen aufgespannt.
    Mithilfe dieser Klasse wird eine Dimension beschrieben.
  </rdfs:comment>
  <owl:disjointWith rdf:resource="#Cube"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="zugehoerigerCube"/>
      </owl:onProperty>
      <owl:someValuesFrom rdf:resource="#Cube"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
[...]
```

```

<!-- Definition der Rollen -->
<owl:ObjectProperty rdf:about="#hatDimension">
  <rdfs:label>hat Dimension</rdfs:label>
  <rdfs:comment>Eine Verknuepfung von Cube und Dimension
    wird durch diesen Eigenschaftstyp definiert.
  </rdfs:comment>
  <owl:inverseOf rdf:resource="#zugehoerigerCube"/>
  <rdfs:domain rdf:resource="#Cube"/>
  <rdfs:range rdf:resource="#Dimension"/>
</owl:ObjectProperty>
[...]
```

```

</rdf:RDF>
```

Quellcode 4: OWL-Ontologie für DWH-Metadaten (Ausschnitt)

Anzumerken ist, dass aufgrund der Analogie zum DWH-RDFS an dieser Stelle von einem konkreten Anwendungsbeispiel abgesehen wurde und stattdessen auf die Fallstudie in Kapitel 9 verwiesen wird.

7.3.4 Zusammenfassung

Die Ausführungen dieses Abschnitts widmeten sich der Ausarbeitung einer OWL-Ontologie zur semantisch reichhaltigen Beschreibung von DWH-Metadaten. Eingangs wurden hierzu die Potentiale von OWL zur semantisch reichhaltigen Beschreibung von DWH-Metadaten untersucht. Die gewonnen Erkenntnisse zeigten, dass die Klassen und Eigenschaftstypen des DWH-RDFS direkt in eine OWL-basierte Beschreibung überführt werden können. Basierend auf dieser Grundstruktur wurden anschließend diese Klassen und Rollen mit OWL-DL-spezifischen Sprachkonstrukten semantisch angereichert und detailliert vorgestellt. Die im Fazit des DWH-RDFS genannten Vorteile und realisierten Anforderungen gelten – den Ausführungen des vorliegenden Abschnitts logisch folgernd – in gleicher Weise für die DWH-OWL (z. B. eindeutige Erfassung der Semantik eines Metadaten-Elements, Erweiterbarkeit der DWH-OWL, Vergleichbarkeit von – in Bezug auf die DWH-OWL – validen OWL-Dokumenten). Eine umfassende Werkzeugunterstützung zur Modellierung eines DWH-Systems auf Basis der DWH-OWL besteht ebenfalls (vgl. Abschnitt 6.3.3).

Bei der Entwicklung der DWH-OWL wurde darauf geachtet, die grundlegenden Entwicklungskriterien für Ontologien (vgl. Abschnitt 6.3.1) einzuhalten und gleichzeitig die Kriterien des Anforderungskatalogs (vgl. Abschnitt 5.5) zu erfüllen. Ersteres wurde z. B. dadurch beachtet, dass lediglich die zur Beschreibung eines DWH-Systems zwingend notwendigen Elemente über Kardinalitätsrestriktionen als verbindlich deklariert wurden (dies betrifft insbesondere Klassen und Rollen der Kategorie DWH-Kernelemente). Daneben wurde auf eine intuitiv erfassbare, sowie möglichst allgemeingültige Beschreibung der Klassen und Eigenschaftstypen Wert gelegt, um eine hohe Akzeptanz bei den Anwendern dieser Ontologie zu erzielen.

7.4 Zusammenfassung

Im vorliegenden Kapitel wurden zunächst die Potentiale von RDF/RDFS zur semantischen Beschreibung von DWH-Metadaten untersucht. Basierend auf diesen Ergebnissen erfolgte die Vorstellung eines RDFS für DWH-Metadaten. Da OWL im Vergleich zu RDF/RDFS eine semantisch höhere Ausdrucksstärke besitzt, wurde in Abschnitt 7.3.2 eine Möglichkeit zur Überführung des DWH-RDFS in eine OWL-basierte Beschreibung aufgezeigt. Auf dieser Grundlage erfolgte mittels OWL DL die Entwicklung einer

Ontologie für eine semantisch reichhaltige Beschreibung von DWH-Metadaten. Dabei kamen u. a. Sprachelemente für Kardinalitätsangaben sowie zur Auszeichnung disjunkter Klassen und inverser oder transitiver Rollen zum Einsatz. Graphbasierte Darstellungen, tabellarische Überblicke und Ausschnitte aus dem jeweiligen Quellcode dienten dabei zur Veranschaulichung der Ausführungen. Die Anwendbarkeit des DWH-RDFS wurde exemplarisch aufgezeigt.

Das einerseits in RDF verfügbare Schema für die semantische Beschreibung von DWH-Metadaten und andererseits die in OWL entwickelte Ontologie für eine semantisch reichhaltige Beschreibung von DWH-Metadaten sind zentral über die jeweilige URL ihres Namensraums zugreifbar. Das DWH-RDFS bzw. die DWH-OWL können somit für die Modellierung von DWH-Schemata beliebig referenziert werden. Die strukturelle und semantische Logik für die Beschreibung eines DWH-Systems ist dabei inhärent vorhanden und muss nicht – wie gegenwärtig üblich – erst von spezifischen Modellierungswerkzeugen implementiert und angeboten werden. Darüber hinaus arbeiten alle Nutzer des DWH-RDFS bzw. der DWH-OWL (an dieser Stelle sind insbesondere Entwickler und Administratoren angesprochen) mit den gleichen Begrifflichkeiten und unterliegen alle denselben Restriktionen für die Beschreibung eines DWH-Systems. Auf diese Weise ist eine strukturelle und semantische Homogenität bei der Modellierung von DWH-Systemen erreichbar. Spezifische Erweiterungen des DWH-RDFS bzw. der DWH-OWL sind durch eine neue Version des Schemas bzw. der Ontologie zu kennzeichnen. Durch die Angabe von Versionsinformationen und Beschreibung des Deltas zur Vorgängerversion ist eine Abstimmung zwischen Instanzen unterschiedlicher Versionen nachvollziehbar und Vergleichbarkeit nachhaltig gewährleistet.

Die semantisch reichhaltige Metadaten-Beschreibung von DWH-Systemen stellt eine Vorbedingung für die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen dar (vgl. Abschnitt 1.2.2). Durch den Einsatz der DWH-OWL ist diese Vorbedingung erfüllbar. Zugleich ist Maschinenlesbarkeit der Metadaten-Beschreibung inklusive ihrer semantischen Merkmale gegeben (vgl. Abschnitt 6.3). In welcher Weise die DWH-OWL als Grundbaustein zur Erreichung des Untersuchungsziels dieser Arbeit dienen kann, wird im nächsten Kapitel beleuchtet.

8 Konzeption einer semantischen Homogenisierungsschicht zur Überwindung der semantischen Heterogenität bei multiplen Data-Warehouse-Systemen

Den Kernpunkt dieses Kapitels bildet die Konzeption einer semantischen Homogenisierungsschicht (sHGS), als eine multiplen DWH-Systemen übergeordnete Schicht. Sie dient als Hilfsmittel zur Überwindung deren semantischer Heterogenität. Die Ausführungen umfassen einleitend die Vorstellung der Grundidee einer sHGS für DWH-Systeme. Hauptbestandteil der weiteren Untersuchung bildet die Erläuterung, auf welche Weise eine sHGS für DWH-Systeme in eine mögliche Gesamt-Architektur eingebunden werden kann. Parallel dazu wird dargelegt, wie durch deren Einsatz die semantische Heterogenität multipler DWH-Systeme überwunden werden kann. Abschließend erfolgt eine kritische Würdigung des erarbeiteten Lösungsvorschlags.

8.1 Aufbau einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme

Der vorliegende Abschnitt dient der Konzeption einer semantischen Homogenisierungsschicht (sHGS), die oberhalb multipler DWH-Systeme anzuordnen ist. Durch deren Einsatz kann das gegebene Untersuchungsziel – die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen – erreicht werden. Die in Abschnitt 7.3 erarbeitete DWH-OWL soll dabei als ein elementarer Baustein einer sHGS für DWH-Systeme (sHGS-DWH) fungieren.

Nachfolgend wird zunächst aus allgemeiner Sicht die Grundidee einer sHGS vorgestellt. In diesem Zusammenhang wird besonders der Einsatz von Ontologien hervorgehoben. Auf Basis dieser Ausführungen erfolgt anschließend eine Untersuchung hinsichtlich des Aufbaus einer sHGS-DWH. Zu dieser zählt auch, die Rolle der DWH-OWL in einer sHGS-DWH herauszustellen.

8.1.1 Grundidee einer semantischen Homogenisierungsschicht

Eine sHGS soll ein Hilfsmittel zur Überwindung der semantischen Heterogenität (z. B. bei multiplen DWH-Systemen) darstellen. Den Ausgangspunkt der weiteren Überlegungen bildet der Leitgedanke, Überschneidungsbereiche oder Abhängigkeiten zwischen den Schemata multipler DWH-Systeme zu identifizieren. Auf Basis der erkannten Korrespondenzen werden die jeweiligen DWH-Schemata in Beziehung gesetzt und somit eine Verknüpfung zwischen diesen hergestellt. Zur Realisierung einer sHGS kann grundsätzlich auf verschiedenartige Lösungsansätze, wie sie z. B. aus dem Bereich des Schema-Managements (vgl. Abschnitt 5.4) bekannt sind, zurückgegriffen werden. Um jedoch insbesondere die Semantik der Schema-Elemente exakt ermitteln und diese bei der Heterogenitätsüberwindung einbeziehen zu können, wird an dieser Stelle der Aufbau einer ontologiebasierten sHGS vorgeschlagen.

ONTOLOGIEBASIERTE KONZEPTION EINER SEMANTISCHEN HOMOGENISIERUNGSSCHICHT

Der Grundgedanke einer ontologiebasierten Konzeption einer sHGS beruht auf der Definition einer übergeordneten, allgemeingültigen Ontologie, die als Vorlage für die Entwicklung (fach-)spezifischer Ontologien dient. Derartige (fach-)spezifische Ontologien können sodann eingesetzt werden, um bspw. DWH-Schemata semantisch reichhaltig zu beschreiben. Durch ihre Abstammung von ein und derselben Ontologie ist eine Beziehung zwischen Elementen (fach-)spezifischer Ontologien grundsätzlich durch die Existenz eines gemeinsamen höheren Konzepts möglich. Liegt eine mehrstufige Abhängigkeit vor, d. h. mehrere Ontologien bauen jeweils aufeinander auf, ist eine Verknüpfung auf der jeweils niedrigsten und somit semantisch reichhaltigsten gemeinsamen Stufe anzustreben.

Eine ontologiebasierte sHGS kann im Allgemeinen als ein für den Einsatz in dezentralen Umgebungen ausgelegtes Konstrukt bezeichnet werden. Die für den jeweiligen Anwendungsfall innerhalb einer sHGS benötigten Ontologien können an unterschiedlichen Orten, d. h. *dezentral* bereitgestellt sein. Sie müssen lediglich über ihren URI referenzierbar sein und auf diese Weise in eine sHGS eingebunden werden können (vgl. Import von OWL-Ontologien; Abschnitt 6.3.3).

8.1.2 Ontologiebasierte Ausgestaltung einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme

Für die Konzeption einer sHGS für DWH-Systeme (sHGS-DWH) wird der ontologiebasierte Ansatz gewählt. (Zur Begründung dieser Entscheidung wird auf die direkt vorgegangen Ausführungen verwiesen.) Um einerseits einen flexiblen Einsatz einer sHGS-DWH zu gewährleisten und andererseits eine semantisch reichhaltige Beschreibung des jeweiligen (fachspezifischen) Untersuchungsbereichs zu erzielen, wird bei der Konzeption einer sHGS-DWH auf die Ontologie-Klassifizierung nach GUARINO zurückgegriffen (vgl. Abschnitt 6.3.1). Die Gesamt-Ontologie in einer sHGS-DWH wird gemäß dieser Ontologie-Klassifizierung in *Top-Level*-, *Domain*- und *Applikationsontologie* unterteilt. Welche Aufgaben Ontologien der unterschiedlichen Kategorien innerhalb einer sHGS-DWH besitzen und in welcher Beziehung diese zueinander stehen, wird im Folgenden erläutert. Dabei wird von einem durchgängigen Einsatz der Ontologiesprache OWL ausgegangen. Zu betonen ist, dass aufgrund des Namensraumkonzepts (vgl. Abschnitt 6.2.2) generell eine Identifikation und somit Unterscheidung der Konzepte der verschiedenartigen Ontologien sichergestellt ist.

TOP-LEVEL-ONTOLOGIE

In Abschnitt 7.3 wurde mit der DWH-OWL eine Top-Level-Ontologie zur semantisch reichhaltigen Beschreibung von DWH-Systemen eingeführt. Die dabei beschriebenen Konzepte sind – gemäß der Bestimmung einer Top-Level-Ontologie – allgemeingültig und somit unabhängig von einem (fach-)spezifischen Anwendungsbereich definiert. Die DWH-OWL soll aus diesem Grund die Top-Level-Ontologie einer sHGS-DWH darstellen. Sie bildet zugleich die oberste Ebene einer sHGS-DWH und dient als Ausgangsbasis für die Ableitung (fach-)spezifischer DWH-Ontologien.

DOMAIN-ONTOLOGIE

Eine Domain-Ontologie verkörpert eine Spezialisierung der Konzepte einer Top-Level-Ontologie in Bezug auf eine bestimmte Domäne oder Aufgabe. Demzufolge werden auf dieser Ebene einer sHGS-DWH – abgeleitet von den Konzepten der DWH-OWL – die Schemaelemente eines DWH-Systems zu einem fest umrissenen Anwendungsbereich beschrieben (z. B. DWH-System zu einer Domäne *Studentendaten*). Domain-Ontologien innerhalb einer sHGS-DWH dienen als Referenzmodelle zur Repräsentation und

Wiederverwendung von Wissen über die Ausgestaltung eines DWH-Systems in einer gegebenen Domäne (vgl. Anwendungsfelder von Ontologien; Abschnitt 6.3.1).¹⁰⁵

Die in Top-Level- und Domain-Ontologien erfassten Konzepte stehen innerhalb einer sHGS-DWH in einer Unterklasse- und Untereigenschaftsbeziehung. Dies bedeutet, dass die (semantische) Konkretisierung einer Klasse der Top-Level-Ontologie als eine zugehörige Unterklasse auf Ebene der Domain-Ontologie modelliert ist. Dieses Vorgehen ist analog bei der Spezifikation von Rollen in einer Domain-Ontologie anzuwenden (Untereigenschaftsbeziehung). Die semantische Detaillierung kann dabei fallspezifisch und bedarfsgerichtet festgesetzt werden. Auch eine beliebige Kombination bei der Nutzung der Unterklasse- und Untereigenschaftsbeziehung ist denkbar. So können z. B. Klassen semantisch verfeinert und gleichzeitig die von der Top-Level-Ontologie offerierten Rollen unverändert in einer Domain-Ontologie referenziert werden.

Ein alternativer Ansatz besteht darin, eine Domain-Ontologie als Instanz der Top-Level-Ontologie abzubilden. Da eine Domain-Ontologie einer sHGS-DWH jedoch als Referenzmodell für die Ableitung weiterer Ontologien dienen soll, wäre bei dieser Lösung der Einsatz von OWL Full notwendig (Individuen der Domain-Ontologie wären gleichzeitig Klassen abhängiger Ontologien; vgl. Abschnitt 6.3.1). Durch die Nutzung der erläuterten Unterklasse- und Untereigenschaftsbeziehung zwischen Top-Level- und Domain-Ontologie ist jedoch ein Rückgriff auf die Instanzebene nicht erforderlich. Eine Beschränkung auf OWL DL ist hierdurch weiterhin möglich und Entscheidbarkeit innerhalb einer sHGS-DWH ist gewährleistet.

APPLIKATIONSONTOLOGIE

Applikationsontologien stellen in einer sHGS-DWH Konzepte zur Beschreibung von konkreten DWH-Systemen bestimmter Einsatzgebiete dar (z. B. DWH-System zu einer Domäne *Studentendaten* an Hochschule *A*). Sie sind eine Spezialisierung einer Domain-Ontologie. Analog zur Beziehung zwischen Top-Level- und Domain-Ontologie wird für

¹⁰⁵ Einen fundierten Überblick über Referenzmodellierung im Kontext von DWH-Systemen bieten u. a. [BeKn04] und [Knac06, S. 33 ff.].

die Ableitung einer Applikationsontologie von einer Domain-Ontologie das Konzept der Unterklasse- und Untereignschaftsbeziehung herangezogen.¹⁰⁶

Applikationsontologien sind der untersten Ebene einer sHGS-DWH zugeordnet und beinhalten somit die semantisch reichhaltigste Beschreibung von Konzepten innerhalb einer sHGS-DWH. Die Überwindung der semantischen Heterogenität zwischen DWH-Systemen wird vornehmlich auf dieser Ebene angestrebt. Eine ausführliche Erläuterung hierzu folgt in Abschnitt 8.2.

GESAMTSICHT AUF DIE GRUNDSTRUKTUR EINER SEMANTISCHEN HOMOGENISIERUNGSSCHICHT FÜR DATA-WAREHOUSE-SYSTEME

Zur Veranschaulichung der Ausführungen dieses Abschnitts bietet Abbildung 8-1 einen Überblick über die Grundstruktur einer sHGS-DWH. Den drei vorgestellten Ontologie-Kategorien ist dabei jeweils eine Ebene zugewiesen. In der Abbildung ist auf oberster Ebene (Top-Level-Ontologie) exemplarisch ein Ausschnitt der DWH-OWL dargestellt. Die mittlere Ebene (Domain-Ontologie) umfasst einen Teilbereich einer Domain-Ontologie zu einem DWH-System mit Studentendaten. Der untersten Ebene sind Ausschnitte von Applikationsontologien zu den DWH-Systemen (Studentendaten) zweier Hochschulen zugeordnet. Die in der Einleitung dieser Arbeit betrachteten konzeptuellen DWH-Schemata dienen dabei als Vorlage.¹⁰⁷

Die Unterklassebeziehungen zwischen den Ebenen sind plakativ durch dicke, vertikale Pfeile hervorgehoben. Zur Wahrung der Übersichtlichkeit wurden dabei nicht alle Unterklassebeziehungen herausgestellt und auf die Erfassung von Untereignschaftsbeziehungen verzichtet. Darüber hinaus sind die Rollenbeschriftungen sowie inverse Rollen ausgeblendet.

¹⁰⁶ Da von einer Applikationsontologie in einer sHGS-DWH keine weiteren Ontologien abhängen, ist an dieser Stelle die Abbildung einer Applikationsontologie als Instanz einer Domain-Ontologie – bei gleichzeitiger Einbehaltung der Sprachstufe OWL DL – ebenfalls denkbar. Im Folgenden wird jedoch eine Anwendung der Unterklasse- und Untereignschaftsbeziehung bevorzugt.

¹⁰⁷ Eine weiterführende Untersuchung und Beschreibung dieser beiden konzeptuellen DWH-Schemata als Applikationsontologien findet sich in Kapitel 9.

Anhand obiger Abbildung ist zu erkennen, dass die DWH-Systeme der Hochschulen A und B jeweils einen Cube mit Studentendaten umfassen. Obgleich der unterschiedlichen Benennung (*Studenten* vs. *Studierende*) kann für diese – aufgrund ihrer Abhängigkeit von dem gemeinsamen Konzept *Studenten* der Domain-Ontologie – eine Verknüpfung hergestellt werden. Ihre semantische Zusammengehörigkeit ist aufgrund gemeinsamer Eigenschaften mit dem übergeordneten Konzept *Studenten* belegt. Dieses Paradigma ist entsprechend auf alle in den Applikationsontologien erfassten Konzepte übertragbar, die ein gleiches Konzept einer Domain-Ontologie teilen.

Nicht direkt vereinbar sind dagegen in dem gezeigten Beispiel die beiden gefilterten Kennzahlen *Studenten im 1. FS* (Hochschule A) und *Studienanfänger* (Hochschule B). Anhand ihrer Beziehungen in Abbildung 8-1 ist zu erkennen, dass letztere auf Basis des Hochschulseesters bestimmt wird und nicht wie bei Hochschule A über das Fachsemester (*FS*) abgeleitet wird. Für diese Konzepte existiert somit keine direkt übergeordnete gemeinsame Entsprechung. Zwischen den beiden Konzepten besteht daher keine semantische Gleichheit. Jedoch kann z. B. semantische Ähnlichkeit gegeben sein. In den nachfolgenden Abschnitten wird dieser Aspekt vertieft.

8.1.3 Zusammenfassung

Im vorliegenden Abschnitt wurde zunächst die Grundidee einer sHGS skizziert sowie das Konzept einer ontologiebasierten sHGS vorgestellt. Dem folgte eine Erörterung des Aufbaus einer ontologiebasierten sHGS-DWH. Jeder Ebene einer sHGS-DWH wurde in diesem Zusammenhang – gemäß der Klassifizierung von GUARINO – eine bestimmte Ontologiekategorie zugeordnet. Die Bestimmung und Erklärung der jeweiligen Aufgaben einer Ontologiekategorie sowie deren ebenenübergreifenden Beziehungen bildeten den Schwerpunkt dieses Abschnitts. Zur Unterstützung der Ausführungen wurde die Grundstruktur einer sHGS-DWH in einer Gesamtsicht abgebildet.

Die Partitionierung der Gesamt-Ontologie einer sHGS-DWH erlaubt, die Teil-Ontologien gemäß den jeweils geforderten Restriktionen und Anforderungen an deren semantische Reichhaltigkeit auszugestalten. Dabei ist abwärts betrachtet über die Ebenen einer sHGS-DWH eine Zunahme der semantischen Reichhaltigkeit der jeweils beschriebenen Konzepte zu verzeichnen. Die Mächtigkeit der semantischen Auszeichnung eines Konzepts beeinflusst signifikant die Möglichkeiten, semantische Beziehungen zwischen Domain- oder Applikationsontologien aufzudecken und explizit zu erfassen.

Als grundsätzlich gegeben kann innerhalb einer sHGS-DWH die Verknüpfungsmöglichkeit der Applikationsontologien durch eine gemeinsame übergeordnete Klasse auf Ebene der Domain- oder Top-Level-Ontologie angesehen werden. Dies liegt in der Ableitung von Domain- und Applikationsontologien über Unterklasse- und Untereigenchaftsbeziehungen begründet. Die semantischen Merkmale der jeweils höheren Konzepte werden dabei übernommen und ggf. detailliert. Die somit erschlossene semantische Relation zwischen Applikationsontologien bildet den Ausgangspunkt für die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen.

8.2 Gesamt-Architektur einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme

An die Betrachtung der Grundstruktur einer sHGS-DWH schließt sich in diesem Abschnitt deren Einbindung in eine Gesamt-Architektur an. Anhand einer möglichen Gesamt-Architektur soll aufgezeigt werden, wie das Untersuchungsziel dieser Arbeit, die Überwindung semantischer Heterogenität bei multiplen DWH-Systemen, erreicht werden kann.

Einleitend wird zunächst eine Gesamt-Architektur zu einer sHGS-DWH vorgestellt und deren Aufbau erläutert. Die weiteren Ausführungen dienen der Herausstellung wesentlicher Bestandteile und Konzepte der vorgeschlagenen Gesamt-Architektur. In diesem Zusammenhang sind zunächst Kopplungskomponenten zur Anbindung von DWH-Systemen an eine sHGS-DWH zu nennen. Danach werden Aspekte der Versionierung für Ontologien in einer sHGS-DWH untersucht. Diese beziehen auch die Thematik der Evolution von DWH-Schemata mit ein. Eine Diskussion über elementare Funktionen einer sHGS-DWH rundet diesen Abschnitt ab.

8.2.1 Vorstellung einer Gesamt-Architektur

Als Hilfsmittel zur Erreichung des gesetzten Untersuchungsziels wird folgende Gesamt-Architektur einer sHGS-DWH vorgeschlagen (Abbildung 8-2). Die Architektur gliedert sich in den Bereich bestehender (multipler) DWH-Systeme, in eine Kopplungsschicht sowie in eine sHGS-DWH.

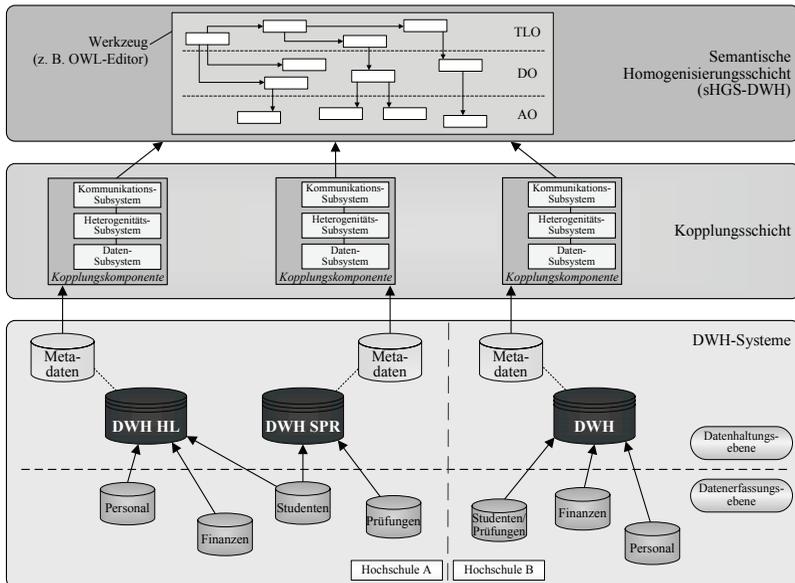


Abbildung 8-2: Gesamt-Architektur einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme

Ausgangspunkt der nachstehenden Erläuterungen bilden die in Abschnitt 1.3 illustrierten multiplen DWH-Systeme der Hochschulen A und B.

KOPPLUNG VON DATA-WAREHOUSE-SYSTEMEN

Um ein DWH-System an eine sHGS-DWH anbinden zu können, kommen so genannte Kopplungskomponenten zum Einsatz. Diese sind der *Kopplungsschicht* zugeordnet. Kopplungskomponenten unterstützen in diesem Anwendungsfall die Aufgabe, die jeweiligen Metadaten – und somit die durch diese beschriebenen DWH-Schemata – in eine ontologiebasierte Beschreibung, d. h. in eine Applikationsontologie, zu überführen. Ein Schema wird dabei auf genau eine Applikationsontologie abgebildet.

Wie im vorherigen Abschnitt dargelegt, erfolgt die Konstruktion einer Applikationsontologie auf Basis einer Domain- und der Top-Level-Ontologie DWH-OWL. Jedes Element eines DWH-Schemas wird dabei als (semantische) Konkretisierung eines zugehörigen Konzepts einer fachlich dem jeweiligen Ausschnitt eines DWH-Systems entsprechenden Domain-Ontologie erfasst. Zu diesem Zweck sind Unterklasse- und Untereignschaftsbeziehungen zu nutzen. Der Sonderfall nicht zuordenbarer Elemente eines DWH-Schemas aufgrund fehlender Konzepte in einer bestimmten Domain- oder der Top-Level-Ontologie wird im Rahmen der Diskussion zur Versionierung von Ontologien erörtert und gelöst (vgl. Abschnitt 8.2.3). Die Applikationsontologien werden abschließend mithilfe der OWL-Import-Anweisung (*owl:imports*) innerhalb einer sHGS-DWH zusammengeführt.

IDENTIFIKATION UND KLASSIFIZIERUNG SEMANTISCHER BEZIEHUNGEN

Überschneidungsbereiche und Abhängigkeiten zwischen Applikationsontologien werden bei Existenz gemeinsamer, übergeordneter Konzepte in einer sHGS-DWH automatisch (bzw. gestützt durch einen Reasoner) erkannt und je nach gewählter Darstellungsform entsprechend ausgewiesen. Anzumerken ist, dass bei einer optionalen Nutzung von Inferenzprozessen zur Identifikation von Relationen zwischen Schemaelementen deren Ergebnis von der semantischen Reichhaltigkeit der jeweiligen Begriffsbeschreibungen abhängt. Dies betrifft vor allem Domain- und Applikationsontologien. Je detaillierter die Konzepte auf diesen Ebenen erfasst sind, umso besser können die Möglichkeiten der Inferenz ausgeschöpft werden. Demgegenüber wird jedoch deren Verwendungsbereich zugleich restringiert.

Semantische Heterogenität gilt als überwunden, wenn unter den verwendeten Begriffsworten für Schemaelemente auch jeweils dasselbe verstanden wird, d. h. den Schemaelementen liegt jeweils die identische Bedeutung zugrunde (vgl. Abschnitt 5.3.4). Die Bedeutung eines Schemaelements (hier repräsentiert durch ein Konzept in einer Applikationsontologie) kann in einer sHGS-DWH zum einen anhand der geerbten Merkmale von einem Konzept einer übergeordneten Ontologie und zum anderen durch seine ergänzende, semantisch reichhaltige Auszeichnung präzise erfasst werden. Auf dieser Grundlage können bspw. unterschiedlich bezeichnete Schemaelemente als semantisch identisch erkannt werden (*Unique-Names-Assumption*, [AnHa04b, S. 145]). Mithilfe spezieller Sprachausdrücke (z. B. *rdfs:seeAlso*; Abschnitt 7.1.1) besteht anschließend die Möglichkeit, eine solche Beziehung explizit zu modellieren.

Darüber hinaus kann mithilfe einer sHGS-DWH eine semantische Beziehung zwischen zwei Konzepten unterschiedlicher (Applikations-)Ontologien klassifiziert werden. Die Zuhilfenahme eines semantischen Ähnlichkeitsmaßes bietet hierfür zum einen eine entsprechende Unterstützung und zum anderen eine adäquate Detaillierung des Ähnlichkeitsgrads (vgl. Abschnitt 5.3.5). Abhängig vom erkannten Ähnlichkeitsgrad können in einer sHGS-DWH zwischen verschiedenen Konzepten entsprechende Rollen zur Kategorisierung der jeweiligen semantischen Relation eingesetzt werden. OWL bietet jedoch nur für die Kennzeichnung semantischer Äquivalenz eine (rudimentäre) Unterstützung (z. B. *owl:equivalentClass*, *owl:equivalentProperty*). Für eine detaillierte Auszeichnung semantischer Ähnlichkeitsgrade wird daher eine OWL-Ontologie bereitgestellt, welche die in Abschnitt 5.3.5 definierten Ähnlichkeitsgrade als unabhängige Rollen, d. h. ohne Domänen- und Wertebereichseinschränkung, bereitstellt (z. B. *semPro:semanticResemblance*, *semPro:semanticRelationship*). Diese Ontologie kann über den nachstehenden Namensraum ergänzend in eine sHGS-DWH eingebunden werden.

```
xmlns:semPro="http://www.semantic-metadata.com/dwhowl/semanticProximity#"

```

Der jeweilige Ähnlichkeitsgrad zwischen zwei Konzepten kann somit durch den Einsatz der entsprechenden Rolle formal erfasst werden. Für die Überwindung deren semantischer Heterogenität bildet diese Relation einen wertvollen Anknüpfungspunkt. In der nachstehenden simplifizierten Abbildung ist exemplarisch für die gefilterten Kennzahlen *Studenten im 1. FS* und *Studienanfänger (HS=1)* eine Rolle zur Kennzeichnung eines hier nicht näher spezifizierten Ähnlichkeitsgrads herausgestellt (horizontaler, gestrichelter Pfeil).

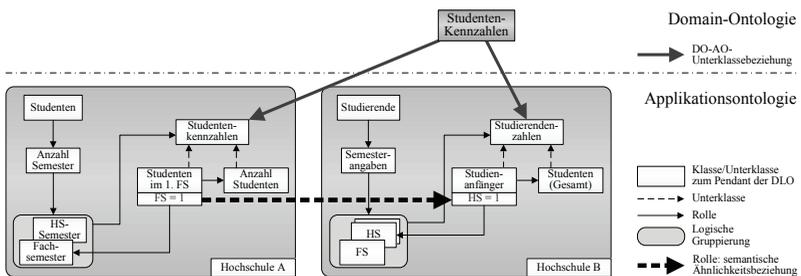


Abbildung 8-3: Semantische Ähnlichkeitsbeziehung zwischen Applikationsontologien

ÜBERWINDUNG SEMANTISCHER HETEROGENITÄT

Auf diesen Ergebnissen aufbauend können zum einen die nicht als unvereinbar erkannten Konzepte entsprechend ihrer semantischen Überschneidung und ggf. unter Mitwirkung eines Fachexperten aufeinander abgestimmt oder angeglichen werden. Semantische Heterogenität kann hier quasi auch aufgelöst werden. Hilfreich wirken sich in diesem Zusammenhang vor allem die vorgestellten Ansätze zur Behandlung sprachlicher Defekte (vgl. Abschnitt 5.2.2 und Abschnitt 5.2.3) sowie die Methode zur Rekonstruktion von Fachbegriffen aus (vgl. Abschnitt 5.2.4). Resultierend aus deren Einsatz kann zum anderen die exakte Spezifikation einer semantischen Beziehung zwischen zwei nicht unvereinbaren Konzepten durch die Definition einer genau auf den Kontext ausgerichteten Rolle erfasst werden. Eine solche Rolle ist somit als eine Art Transformationsvorschrift zu interpretieren, um von einem Konzept auf ein anderes semantisch vollständig zu schließen. Semantische Heterogenität bleibt in diesem Fall bestehen. Jedoch erlaubt das formal beschriebene Wissen über die semantische Relation zweier Konzepte, deren semantische Heterogenität zu überwinden.

Als verschiedenartig eingestufte Konzepte werden im Allgemeinen auch als solche behandelt und bleiben in ihrer jeweiligen Ausprägung bestehen.

RESÜMEE

In ihrer Summation bewirken die vorgestellten Komponenten und Konzepte einer sHGS-DWH-Gesamt-Architektur die Erreichung des Untersuchungsziels. Um dieses zu unterstreichen, wird nachfolgend auf wesentliche Merkmale der Gesamt-Architektur detailliert eingegangen.

8.2.2 Kopplungskomponenten

Zur Beschreibung der Kopplungskomponenten, welche zugleich die Kopplungsschicht bilden, wird auf Forschungsergebnisse aus dem Bereich der inner- und überbetrieblichen Integration von Anwendungssystemen (AwS) zurückgegriffen.¹⁰⁸ Die Integration von AwS erfordert deren Kopplung gemäß einer Kopplungsarchitektur. Eine Kopplungsarchitektur stellt eine Projektion auf die AwS-Architekturen von zwei oder mehreren integrierten AwS dar. In Form eines Bauplans beschreibt sie alle für die Integration erforderlichen Elemente sowie deren Beziehungen zueinander [MES+02, S. 1].

Kopplungsarchitekturen können in unterschiedliche Klassen differenziert werden. Mit Fokus auf die Übertragung der Ergebnisse des genannten Forschungsbereichs auf eine sHGS-DWH-Gesamt-Architektur wird hier lediglich die datenorientierte Kopplungsarchitektur eingeführt. Diese dient der Manipulation gemeinsamer Daten mehrerer AwS auf Basis einer engen Kopplung der auf den Daten operierenden Funktionen [FeSi06, S. 224]. Dabei ist zwischen einer Kopplung auf Instanz- und einer Kopplung auf Typ-Ebene zu unterscheiden. Der vorgestellte Ansatz einer sHGS-DWH betrachtet ausschließlich die Schemaebene von DWH-Systemen. Demgemäß wird lediglich die Kopplung auf Typ-Ebene weiter beleuchtet. Die nachfolgend skizzierten Subsysteme einer datenorientierten Kopplungsarchitektur ([MES+02], [SMFS02, S. 462 f.]) werden zugleich auf ihren Einsatz und ihre Funktion zur Kopplung von DWH-Systemen an eine sHGS-DWH untersucht. Ein DWH-System ist dabei als Quell-AwS, eine sHGS-DWH als Ziel-AwS anzusehen.

SUBSYSTEME EINES QUELL-ANWENDUNGSSYSTEMS

Eine typische datenorientierte Kopplungsarchitektur sieht auf Ebene eines Quell-AwS ein *Ereignis-* und ein *Empfänger-Subsystem* vor. Da der Empfänger – eine Applikationsontologie in einer sHGS-DWH – eindeutig definiert ist, wird ein Empfänger-Subsystem im vorliegenden Anwendungsszenario nicht benötigt. Ebenso kann auf ein Ereignis-Subsystem verzichtet werden, da von einer manuell initiierten Abbildung der DWH-Schemata auf eine Applikationsontologie ausgegangen wird. Regelmäßige bzw.

¹⁰⁸ An dieser Stelle sind insbesondere Arbeiten und Lösungsansätze, die im Rahmen des *OASYS-Projekts* (<http://141.13.6.53:8080/forwin>, Abruf am 10.06.2008) entstanden, einzubeziehen.

in kurzen Zeitabständen stattfindende Aktualisierungen einer Applikationsontologie stellen bei einer sHGS-DWH die Ausnahme dar.

Des Weiteren umfasst der Vorschlag einer datenorientierten Kopplungsarchitektur ein *Daten-Subsystem*. Dessen Aufgabe liegt in der Ermittlung der zu übertragenden Daten. An dieser Stelle ist insbesondere die Gruppe der DWH-Entwickler gefordert, in einem ersten Schritt die relevanten Metadaten eines zu untersuchenden DWH-Systems zu identifizieren. Die Auswahl der zu extrahierenden Metadaten orientiert sich dabei am angestrebten Beschreibungsumfang in einer sHGS-DWH (z. B. in- oder exklusive der Kategorie DWH-Systemelemente). Die Ergebnisse sind in Form einer Schnittstellenbeschreibung festzuhalten. Die vielfältigen, proprietären Formate bei der Metadatenverwaltung im DWH-Umfeld (vgl. Abschnitt 3.5) bedingen es, jeweils den individuell ausgestalteten Strukturen der Metadaten angepasste Schnittstellen zu erzeugen. Der Grad der Weiterverwendbarkeit einer solchen Schnittstellendefinition in Bezug auf ihren Einsatz bei anderen DWH-Systemen wird daher als gering eingeschätzt.

Die Überwindung der fachlichen Heterogenität zwischen einem Quell- und einem Ziel-AwS erfolgt mithilfe eines *Heterogenitäts-Subsystems*. Bei der erstmaligen Anbindung eines DWH-Systems an eine sHGS-DWH sind die erforderlichen Abbildungsinformationen zwischen den im vorherigen Schritt erfassten Metadaten eines DWH-Systems und den Konzepten einer Domain- oder der Top-Level-Ontologie zunächst durch Fachexperten zu spezifizieren. Im Ergebnis entsteht auf diese Weise eine dem DWH-Schema entsprechende Applikationsontologie. Die Vorschriften zur Abbildung eines DWH-Schemas in eine Applikationsontologie können mithilfe spezifischer Werkzeuge für eine Wiederverwendung gespeichert werden. Gemäß dem Vorbild des Anwendungsprogramms *Mindswap Convert To RDF*¹⁰⁹, das Daten einer flachen Datei auf Basis von Abbildungsvorschriften zu einem gegebenen RDFS in ein gültiges RDF-Dokument transformiert, ist hier eine analoge Unterstützung auf Basis der OWL denkbar.

¹⁰⁹ Das Anwendungsprogramm *Mindswap Convert To RDF Tool* ist unter <http://www.mindswap.org/~mhgrove/convert/> (Abruf am 10.06.2008) beschrieben und steht unter selbiger URL frei zur Verfügung. Eine Aufstellung von Converterprogrammen, die Daten unterschiedlicher Formate in jeweils korrekte RDF-Dokumente transformieren, bietet die Webseite <http://esw.w3.org/topic/ConverterToRdf> (Abruf am 10.06.2008).

Das *Kommunikations-Subsystem* bildet das letzte Element der datenorientierten Kopplungsarchitektur auf Seiten eines Quell-AwS. Es dient der Überwindung der technischen Heterogenität und übernimmt die Übertragung der Daten von einem Quell- zu einem Ziel-AwS. Im vorliegenden Anwendungsfall ist die Aufgabe des Kommunikations-Subsystems mit der Bereitstellung der generierten Applikationsontologie an einer eindeutig referenzierbaren Stelle erfüllt (z. B. Webadresse, Verzeichnis in einem lokalen Netzwerk).

SUBSYSTEME EINES ZIEL-ANWENDUNGSSYSTEMS

Die zu einem Ziel-AwS vorgeschlagenen Subsysteme können für eine sHGS-DWH wie folgt interpretiert werden: Ein Kommunikations-Subsystem ist in einer sHGS-DWH durch die Referenzierung einer Applikationsontologie über ihren URI zu ersetzen. Zudem ist ein Heterogenitäts-Subsystem entbehrlich, da eine Applikationsontologie bereits über alle erforderlichen Beziehungsinformationen zu Konzepten einer Domain- oder der Top-Level-Ontologie verfügt. Diese müssen lediglich – z. B. von einem Werkzeug, das für die Arbeit mit einer sHGS-DWH eingesetzt wird – korrekt interpretiert werden (vgl. Abschnitt 8.2.4). Das bei einer typischen datenorientierten Kopplungsarchitektur vorgeschlagene Aktualisierungs-Subsystem zur Aktualisierung der Daten in einem Ziel-AwS ist im vorliegenden Einsatzbereich ebenfalls nicht notwendig. Ontologien werden jeweils vollständig in eine sHGS-DWH eingebunden, d. h. eine so genannte Delta-Aktualisierung ist nicht vorgesehen. Darüber hinaus sind Änderungen an DWH-Schemata stets als neue Version der entsprechenden Applikationsontologie zu erfassen (vgl. Abschnitt 8.2.3).

Die Subsysteme eines Ziel-AwS sind aufgrund ihrer – in diesem Szenario gegebenen – Ersetzbarkeit in Abbildung 8-2 nicht aufgeführt.

BINDUNGRICHTUNG BEI EINER SEMANTISCHEN HOMOGENISIERUNGSSCHICHT FÜR DATA-WAREHOUSE-SYSTEME

Die Abbildung von DWH-Schemata wird in dieser Arbeit nur in Richtung einer sHGS-DWH beleuchtet, da sich vor allem das Rückschreiben in Metadaten-Bestände kommerzieller Produkte problematisch gestalten kann (z. B. Wahrung der Konsistenz der Metadaten, fehlende Schreibberechtigung auf Metadaten). Zudem bietet nach derzeitigem Entwicklungsstand kein Anbieter die Möglichkeit, semantische Angaben zu den Metadaten eines DWH-Systems in der vorgestellten Reichhaltigkeit zu erfassen.

Für das Untersuchungsziel, die semantische Heterogenität bei multiplen DWH-Systemen zu überwinden, ist die eben erläuterte und hier nur unidirektional betriebene Kopplung von DWH-Systemen an eine sHGS-DWH jedoch hinreichend.

8.2.3 Versionierung

Im Rahmen der Einführung einer sHGS-DWH-Gesamt-Architektur wurde bei der Überführung eines DWH-Schemas in eine Applikationsontologie der Sonderfall nicht zuordenbarer Schemaelemente erkannt (vgl. Abschnitt 8.2.1). Wie dieser Sonderfall durch den Einsatz eines Versionierungskonzepts behandelt und gelöst werden kann, wird nachfolgend dargelegt. Hierzu wird zunächst auf die Thematik der Schemaversionierung bei DWH-Systemen eingegangen.

SCHEMAVERSIONIERUNG BEI DATA-WAREHOUSE-SYSTEMEN

Im Gegensatz zu einem herkömmlichen Softwareentwicklungsprojekt, bei dem im Allgemeinen anhand von Systemversionen die Phasen Entwicklung und Betrieb voneinander abgrenzbar sind, ist bei einem DWH-System von einer kontinuierlichen Weiterentwicklung im Rahmen der Betriebsphase auszugehen [SUPH07b, S. 22]. Bei einem DWH-System kann daher nicht von einem konstant bleibenden Schema ausgegangen werden. Inhaltliche und technische Änderungen an einem DWH-System mit Auswirkung auf die Ausgestaltung des zugehörigen Schemas können z. B. aufgrund von Modifikationen an den Quellsystemen, Reorganisation von Klassifikationshierarchien oder neuen Anforderungen der Endanwender entstehen.

Für die Reproduzierbarkeit von Änderungen an Schemata von DWH-Systemen im Zeitverlauf kann bspw. auf das Konzept der *Schemaversionierung* zurückgegriffen werden. Schemaänderungen werden dabei jeweils durch eine neue Version des betroffenen Schemas erfasst. Durch die Existenz chronologisch generierter Versionen kann der Entwicklungsgang eines Schemas jederzeit nachvollzogen werden. Detaillierte Informationen zu diesem Themengebiet sind u. a. den Literaturquellen [BISH99], [Blas00] und [BaGü04, S. 188 ff.] zu entnehmen.

VERSIONIERUNG IN EINER SEMANTISCHEN HOMOGENISIERUNGSSCHICHT FÜR DATA-WAREHOUSE-SYSTEME

Änderungen an den Schemata von DWH-Systemen sind in einer sHGS-DWH mithilfe eines Versionierungskonzepts zu erfassen. Eine Modifikation an einem DWH-Schema bewirkt die Generierung einer neuen Version der zugehörigen Applikationsontologie, in der die Umgestaltung entsprechend abgebildet und dokumentiert wird. Das Konzept der Versionierung ist für jede Ontologie der sHGS-DWH – unabhängig von ihrer Klassifizierung – einzusetzen.

Eine Applikationsontologie in einer sHGS-DWH bezieht sich gemäß oben eingeführter Definition stets auf eine übergeordnete Domain-Ontologie sowie transitiv auf die Top-Level-Ontologie DWH-OWL. Durch das Konzept der Versionierung ist diese Verknüpfung nachhaltig garantiert, da bestehende Ontologien weder gelöscht noch in irgendeiner Form modifiziert werden. Zudem besitzen unterschiedliche Versionen einer OWL-Ontologie jeweils einen eigenen URI und können somit exakt bestimmt werden. Auf diese Weise kann jederzeit ein Bezug zu früheren Versionen einer spezifischen OWL-Ontologie hergestellt werden. Für die regelgerechte Realisierung einer solchen Versionierung in einer sHGS-DWH sind die von OWL zu diesem Zweck offerierten Sprachelemente einzusetzen. Diese bieten eine ausreichende Ausdrucksstärke, um Schemaänderungen eingehend beschreiben zu können (vgl. Abschnitt 6.3.3).

Bei Bedarf kann eine bestehende Zuordnung einer Ontologie zu einer bestimmten Version einer ihr übergeordneten Ontologie in eine Zuordnung zu einer neueren Version dieser höheren Ontologie überführt werden. Zur Unterstützung dieses Prozesses stehen die bei der Versionierung erfassten Beschreibungen über die vorgenommenen Modifikationen, die Hinweise zur Kompatibilität mit früheren Versionen sowie die Nennung der nicht weiter unterstützten Klassen und Rollen zur Verfügung. Demgemäß kann eine Übernahme der bestehenden Zuordnungen oder eine Neuzuweisung zwischen den Konzepten dieser beiden Ontologien erfolgen.

Anzumerken ist, dass die Definition einer neuen Version einer Ontologie innerhalb der sHGS-DWH in bestimmten Fällen auch von unten nach oben (Bottom-Up) ausgelöst werden kann. So schlägt sich eine Modifikation an einem DWH-Schema zunächst in der zugehörigen Applikationsontologie nieder. Betrifft diese Modifikation bspw. mehrere DWH-Systeme, die von derselben Domain-Ontologie abgeleitet sind, kann diese Änderung eine Anpassung der Domain-Ontologie initiieren. Selbiges Verhalten tritt auf,

wenn in einer Applikationsontologie Konzepte erfasst sind, die sich nicht (eindeutig) Konzepten der übergeordneten Domain-Ontologie zuordnen lassen. Im Einzelfall ist zu entscheiden, ob die betroffene Domain-Ontologie erweitert wird, oder die Applikationsontologie sich des Vokabulars der Top-Level-Ontologie bedient, um diesen Sonderfall zu beschreiben.

Darüber hinaus können spezielle Varianten einer DWH-Architektur auch eine Anpassung oder Erweiterung der Top-Level-Ontologie DWH-OWL erfordern. Für diese ist ergo eine neue Version anzulegen.

8.2.4 Funktionen

Nach den Erläuterungen zur Kopplung von DWH-Systemen an eine sHGS-DWH sowie zur Versionierung von Ontologien steht nachfolgend die Erstellung, Bearbeitung und Nutzung von Ontologien in einer sHGS-DWH im Fokus. Diesbezüglich wird die Unterstützung der anschließend genannten Funktionen – insbesondere im Hinblick auf die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen – durch eine sHGS-DWH gefordert.

- Eine elementare Anforderung an eine sHGS-DWH besteht in der *Beschreibung von DWH-Schemata* auf Basis der Top-Level- oder einer Domain-Ontologie und somit in der Erstellung adäquater Applikationsontologien. Eine sHGS-DWH sollte zu diesem Zweck Möglichkeiten besitzen, die für die Modellierung relevanten (Referenz-)Ontologien einzulesen und deren Konzepte zur weiteren Verwendung oder Spezialisierung bereitzustellen. Dabei sind einerseits – zur korrekten semantischen Erfassung – alle Beschreibungselemente zu einem Konzept auszuweisen und andererseits – gemäß den in den Ontologien definierten Restriktionen – nur die zu einem Konzept gültigen Beziehungen für die Modellierung anzubieten. Als obligatorisch für eine sHGS-DWH ist in diesem Zusammenhang eine generelle Hilfestellung bei der (semantischen) Detaillierung der Konzepte und Rollen einer höheren Ontologie zu betrachten (z. B. in Form einer graphisch gestützten Definition der Unterklasse- und Untereigenschaftsbeziehungen).

- Ontologien müssen innerhalb einer sHGS-DWH *bearbeitbar* sein. Neben der soeben skizzierten Ableitung einer neuen Ontologie wird an dieser Stelle insbesondere auf die Erweiterung und semantische Verfeinerung bestehender Ontologien gezielt. Für diese Aufgaben sollte eine entsprechende ggf. graphische Unterstützung bei der Bearbeitung einer Ontologie geboten werden (z. B. Bereitstellung von Textfeldern für die Erfassung beschreibender Elemente, automatische Gültigkeitsprüfung bei Formel-Definitionen, intuitiv erfassbare Ausdrücke für die Nutzung OWL-spezifischer Sprachkonstrukte). Eine modifizierte Ontologie ist anschließend als neue Version zu kennzeichnen (vgl. Abschnitt 8.2.3).
- Eine sHGS-DWH sollte des Weiteren die *Speicherung* einer Ontologie an einem zentralen Ort (z. B. Webadresse, Verzeichnis in einem lokalen Netzwerk) ermöglichen. Hierbei ist vorzugsweise die OWL-RDF-Syntax zu verwenden.
- Eine graphbasierte Darstellung einer Ontologie fördert die leichte Erfassung sowie das korrekte Verständnis der semantischen Beziehungen in einer Ontologie. Verschiedenartige Formen der *Visualisierung* einer Ontologie (z. B. graphbasierte Darstellung, Sichtenbildung auf Ebene von Oberklassen) sowie der *semantischen Navigation* innerhalb einer Ontologie sollten daher durch eine sHGS-DWH offeriert werden.
- Funktionen zur *Suche* (z. B. nach einem bestimmten Konzept innerhalb einer Ontologie) sind von einer sHGS-DWH ebenfalls zur Verfügung zu stellen. Dabei sollten bekannte Anfragesprachen wie z. B. OWL-QL (vgl. Abschnitt 6.3.3) oder SPARQL (vgl. Abschnitt 6.2.3) unterstützt werden.
- Eine elementare Voraussetzung für die Überwindung der semantischen Heterogenität multipler DWH-Systeme besteht in der *automatischen Zusammenführung* der Applikationsontologien in eine Gesamt-Ontologie. Überschneidungsbereiche und Abhängigkeiten zwischen den Applikationsontologien sind von einer sHGS-DWH – ggf. mit Unterstützung eines Reasoners (vgl. Abschnitt 6.3.1) – innerhalb der Gesamt-Ontologie zu ermitteln und offen zu legen.

Die soeben skizzierten Funktionen sind u. a. zur Erfüllung der in Abschnitt 5.5 aufgestellten Formalziele erforderlich. Für die Realisierung einer sHGS-DWH ist daher ein Werkzeug zu wählen, das die soeben dargelegten Funktionen bereitstellt. Von den – im Rahmen der Erläuterungen zur OWL (vgl. Abschnitt 6.3.3) – vorgestellten Werkzeugen werden diese funktionalen Anforderungen bspw. von *Protégé* vollständig erfüllt.

8.2.5 Zusammenfassung

Die Ausführungen dieses Abschnitts widmeten sich der Einbindung einer sHGS-DWH in eine Gesamt-Architektur. Anhand des Aufbaus einer möglichen Gesamt-Architektur wurde erläutert, wie die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen erreicht werden kann. Die hierfür erforderlichen Komponenten und Konzepte wurden vorgestellt und detailliert beschrieben. Eine Zusammenfassung der wesentlichen Charakteristika ist nachstehend gegeben.

- *Kopplungskomponenten* dienen der Verknüpfung der Metadaten bestehender DWH-Systeme mit einer sHGS-DWH. Subsysteme unterstützen dabei die Schritte Datenextraktion (Daten-Subsystem), Überführung der DWH-Metadaten in eine Applikationsontologie (Heterogenitäts-Subsystem) und Bereitstellung der Applikationsontologie an einem zentralen Ort (Kommunikations-Subsystem).
- Mithilfe des erörterten Konzepts zur *Versionierung* der Ontologien einer sHGS-DWH sind Änderungen an einer Ontologie zu dokumentieren. Einen zentralen Punkt stellt dabei die nachhaltige Sicherstellung bestehender Verknüpfungen zwischen Ontologien unterschiedlicher Ebenen dar. Für die Problematik nicht zuordenbarer Elemente bspw. bei der Abbildung eines DWH-Schemas als Applikationsontologie konnte in diesem Zusammenhang ein Lösungsweg nachgewiesen.
- Die erörterten *funktionalen Anforderungen*, die in einer sHGS-DWH grundsätzlich erfüllt sein sollten, bilden die Vorgabe an eine mögliche Werkzeugunterstützung.

Die Konzeption einer sHGS-DWH kann mit Abschluss obiger Ausführungen als hinreichend im Hinblick auf die Überwindung semantischer Heterogenität multipler DWH-Systeme betrachtet werden. Zu dieser wird im nächsten Abschnitt eine kritische Würdigung vollzogen.

8.3 Kritische Würdigung einer semantischen

Homogenisierungsschicht für Data-Warehouse-Systeme

Zu dem in den zurückliegenden Abschnitten erarbeiteten Konzept einer sHGS-DWH erfolgt nachstehend eine kritische Würdigung. Einleitend wird der Aspekt der Modellierung von DWH-Schemata mithilfe einer sHGS-DWH beleuchtet sowie die dabei erzielbaren Nutzenpotentiale herausgestellt. Dem folgt ein Resümee zur Überwindung semantischer Heterogenität bei multiplen DWH-Systemen unter Einsatz einer sHGS-DWH. Die Ausführungen umfassen zudem jeweils eine Erörterung der Zweckdienlichkeit einer sHGS-DWH für die Personengruppen Entwickler und Endanwender eines DWH-Systems. Eine Untersuchung hinsichtlich der Erfüllung der in Abschnitt 1.3 aufgestellten Formalziele an den Lösungsansatz rundet diesen Abschnitt ab.

8.3.1 Modellierung von Data-Warehouse-Schemata

Die Ausführungen dieses Kapitels haben gezeigt, dass eine sHGS-DWH zur Modellierung und semantischen Beschreibung eines (neuen) DWH-Systems eingesetzt werden kann. Dabei können sowohl Domain-Ontologien – als Referenzmodelle für fachspezifische Ausprägungen eines DWH-Systems – definiert als auch Applikationsontologien – zur Ausgestaltung und semantisch reichhaltigen Beschreibung eines konkreten DWH-Systems – erstellt werden. Der Einsatz einer sHGS-DWH ermöglicht und fördert in diesem Zusammenhang die Modellierung *konsistenter* sowie *strukturell und semantisch homogener DWH-Schemata*. Die beiden letztgenannten Merkmale sind aufgrund der Ableitung eines DWH-Schemas von einer fachspezifischen Domain- bzw. der domänenunabhängigen Top-Level-Ontologie gewährleistet. Die Konsistenz einer Ontologie – und somit einer Beschreibung eines DWH-Systems – kann bspw. mithilfe eines Reasoners überprüft werden. Die in Abschnitt 5.5 geforderte konsistente Metadaten-Beschreibung eines DWH-Systems ist demzufolge gewährleistet.

Des Weiteren kann – den bisherigen Ausführungen folgend – eine Applikationsontologie als eine *semantisch reichhaltige Dokumentation eines DWH-Systems* angesehen werden (vgl. passive Metadaten; Abschnitt 3.3). Die semantische Ausdrucksstärke von OWL ermöglicht es, das über die Modellierung eines spezifischen DWH-Systems vorhandene Wissen (z. B. eines Entwicklers) in einer Ontologie formal und semantisch reichhaltig zu erfassen. Die Bedeutung eines Schemaelements, d. h. die Intension des repräsentierten Begriffs (vgl. Abschnitt 5.2.1), kann dabei eindeutig beschrieben werden. Bisher nicht beachtete oder nicht formalisierbare semantische Merkmale können dabei ebenfalls mit einbezogen werden. Auf diese Weise bleibt das soeben angesprochene Modellierungswissen in einer Organisation nachhaltig bestehen (vgl. Wissensmanagement; Abschnitt 4.1.3). Zudem kann dieses Wissen weiteren Personen auf einfache Weise zur Verfügung gestellt sowie von diesen angeeignet und ausgebaut werden (Förderung des Wissenstransfers). Die Maschinenlesbarkeit einer Ontologie hilft dabei, den Prozess der Erfassung, Weitergabe und Wiederverwendung dieses Wissens durchgängig zu unterstützen (vgl. Abschnitt 4.1.1 und Abschnitt 4.1.2).

REFLEXION DES NUTZENS FÜR ENTWICKLER UND ENDANWENDER EINES DATA-WAREHOUSE-SYSTEMS

Die hier vorgestellten Ergebnisse bilden einerseits für die Gruppe der *Entwickler* und andererseits für die *Endanwender* eines DWH-Systems einen expliziten Mehrwert. Letztere erhalten durch die semantisch reichhaltige Dokumentation eines DWH-Systems ein wertvolles Hilfsmittel an die Hand, um Zusammenhänge sowie die Bedeutung der Daten unmittelbar und korrekt erfassen zu können. Die in Abschnitt 8.2.4 geforderten Funktionen hinsichtlich Visualisierung, Suche und semantischer Navigation innerhalb einer sHGS-DWH tragen hierzu im Besonderen bei.

Der Personenkreis der Entwickler zieht darüber hinaus einen Nutzen aus der Unterstützung bei der Konzeption eines DWH-Systems, indem auf die frei verfügbare Top-Level- oder eine für den gewünschten Einsatzbereich vordefinierte Domain-Ontologie Bezug genommen werden kann. Die Wiederverwendung bestehender Modellierungsbausteine wirkt sich dabei besonders förderlich aus. Zudem kann die Konsistenz eines Modellierungsergebnisses sowohl hinsichtlich struktureller als auch semantischer Merkmale sichergestellt werden (z. B. Konsistenzprüfung mithilfe eines Reasoners).

8.3.2 Überwindung semantischer Heterogenität

Mithilfe einer sHGS-DWH kann semantische Heterogenität bei multiplen DWH-Systemen überwunden werden. In einem ersten Schritt sind hierfür bestehende DWH-Systeme durch jeweils eine Applikationsontologie zu beschreiben (vgl. Abschnitt 8.1.2). Applikationsontologien werden generell von einer fachspezifischen Domain- sowie transitiv von der Top-Level-Ontologie abgeleitet. Deren Existenz und Verfügbarkeit wird vorausgesetzt. Für den Prozess der Erstellung einer Applikationsontologie bieten Kopplungskomponenten eine (semi-)automatische Unterstützung (vgl. Abschnitt 8.2.2). Anschließend werden die Applikationsontologien in einer sHGS-DWH zu einer Gesamt-Ontologie zusammengeführt. Überschneidungsbereiche oder Beziehungen zwischen diesen können bspw. aufgrund gemeinsamer Konzepte in einer übergeordneten Domain-Ontologie identifiziert werden. Der Einsatz so genannter Reasoner erlaubt zudem, nicht explizit modellierte Zusammenhänge aufzudecken. Eine Voraussetzung hierfür bildet die semantisch reichhaltige und konsistente Beschreibung der Konzepte in den beteiligten Ontologien.

Die formale Erfassung und Einbeziehung semantischer Merkmale bei der Heterogenitätsüberwindung grenzen den hier vorgestellten Lösungsansatz deutlich von den Konzepten des Schema- bzw. Model-Managements ab (vgl. Abschnitt 5.4). Bei diesen wird das benötigte Wissen zur Überwindung der (semantischen) Heterogenität zwischen verschiedenen (DWH-)Schemata in Form von Korrespondenzzusicherungen erfasst. Die genaue Bedeutung der Schema-Elemente und ihrer Beziehungen wird jedoch von den Matching-Algorithmen ignoriert [LeNa07a, S. 267]. Dessen ungeachtet wird an dieser Stelle versucht, eine sHGS-DWH in die bekannte Klassifikation von Schema-Matching-Ansätzen (Abbildung 5-8) einzuordnen. Im Bereich der *Individuellen Ansätze* ist sie unter dem Knoten *Schemabasiert* sowohl der Kategorie *Element-* als auch *Strukturebene* zuzuweisen. Die beiden Kategorien sind in diesem Fall als nicht überschneidungsfrei anzusehen.

REFLEXION DES NUTZENS FÜR ENTWICKLER UND ENDANWENDER EINES DATA-WAREHOUSE-SYSTEMS

Bei der Betrachtung des Nutzens einer sHGS-DWH wird erneut zwischen den Personengruppen *Entwickler* und *Endanwender* eines DWH-Systems unterschieden. Der im vorherigen Abschnitt aufgezeigte Mehrwert einer sHGS-DWH für die Gruppe der End-

anwender wird im hier diskutierten Anwendungsfall durch die deutliche Verbesserung des Informationsstands eines Endanwenders (z. B. Entscheidungsträgers, Führungsverantwortlichen) ausgeweitet (vgl. Abschnitt 4.2.1). Die Kombination mehrerer Applikationsontologien an einer zentralen Stelle (in einer sHGS-DWH) bildet einen umfassenden Überblick über – im Idealfall alle – bestehenden DWH-Systeme und deren Beziehungen. Somit ist das verfügbare Informationsangebot transparent dargelegt und relevante Daten können anschließend gezielt abgerufen werden. Differenzen hinsichtlich der Interpretation von Daten aus unterschiedlichen DWH-Systemen können dabei gemäß den obigen Erläuterungen ausgeglichen werden. In diesem Zusammenhang steigt ebenfalls die Unterstützung bei der Bearbeitung von Entscheidungsaufgaben (z. B. aufgrund geringerer Kosten bei der Informationsbeschaffung; vgl. Abschnitt 4.2.2). Der Einsatz einer sHGS-DWH mit dem in diesem Abschnitt diskutierten Ziel trägt zudem zur Verringerung von Informationspathologien bei (vgl. Abschnitt 4.2.3). Die erläuterten Argumente – z. B. umfassender Überblick über das bestehende Informationsangebot, semantisch reichhaltige Beschreibung der Konzepte und die hierdurch gestützte korrekte Interpretation der Daten – sind hierbei als Begründung anzuführen. Eine sHGS-DWH kann somit die Realisierung eines einheitlichen Lenkungssystems z. B. in verteilten Organisationen entscheidend stützen.

Entwickler erfahren einen umfangreichen Nutzen vor allem bei der Konsolidierung bestehender DWH-Systeme. Durch die Zusammenführung der entsprechenden Applikationsontologien in einer sHGS-DWH können u. a. redundant verwaltete Datenbestände auf einfache Weise identifiziert werden. Besonders förderlich wirkt sich in diesem Zusammenhang – neben der semantisch reichhaltigen Beschreibung der DWH-Kernelemente – die Verwendung von Konzepten der Kategorie DWH-Systemelemente aus. So besteht z. B. die Möglichkeit, Abstammung und Transformationsschritte der Daten unterschiedlicher DWH-Systeme vollständig zu erfassen, in Beziehung zu setzen und miteinander zu vergleichen. Zudem können bereits validierte Konzepte in den Applikationsontologien faktischer DWH-Systeme (z. B. Kennzahldefinitionen) zum Aufbau einer Applikationsontologie eines übergreifenden DWH-Systems übernommen werden. Abschließend ist zu konstatieren, dass die Nutzung einer sHGS-DWH nicht nur der Überwindung semantischer Heterogenität bei multiplen DWH-Systemen dient, sondern darüber hinaus Endanwender und Entwickler in vielfältiger Hinsicht bei der Erfüllung ihrer Aufgaben unterstützt.

8.3.3 Erfüllung von Formalzielen

Mit dem in dieser Arbeit vorgegebenen Untersuchungsziel wurden im Eingangskapitel einige Formalziele verbunden (vgl. Abschnitt 1.2.3). Deren Einhaltung soll zum Abschluss dieses Kapitels überprüft und erläutert werden.

Das erstgenannte und zugleich gewichtigste Formalziel ist die *Wahrung der Autonomie bestehender DWH-Systeme*. Um dieses erreichen zu können, wurde eine entsprechende Gesamt-Architektur vorgestellt. Diese umfasst u. a. Kopplungskomponenten, mit deren Hilfe auf die Metadaten von DWH-Systemen lesend zugegriffen werden kann. Ein Rückschreiben von einer sHGS-DWH in die Metadaten eines DWH-Systems wurde aus mehreren Gründen ausgeschlossen (vgl. Abschnitt 8.2.2). Erfordern die durch den Einsatz einer sHGS-DWH gewonnen Erkenntnisse eine Modifikation an einem DWH-System, so ist diese durch einen Entwickler bzw. einen Administrator am jeweiligen DWH-System direkt vorzunehmen.

Des Weiteren galt es, *Flexibilität* und *Erweiterbarkeit* des Lösungsansatzes zu gewährleisten. Aus diesem Grund wurde eine sHGS-DWH auf der Grundlage von Ontologien entworfen. Ontologien können zum einen für beliebige Anwendungsbereiche entworfen werden (Flexibilität), zum anderen sind sie ad libitum erweiterbar. Außerdem können Ontologien bei Bedarf miteinander in Beziehung gesetzt werden. Zu beachten ist lediglich, dass mit zunehmender Zahl der in einer Ontologie beschriebenen Konzepte auch der Zeitbedarf bei der Inferenzberechnung steigt.

Der ontologiebasierte Aufbau einer sHGS-DWH gestattet ebenfalls die Erfüllung der Formalziele *Plattform-* und *DBMS-Unabhängigkeit*. Erstere besteht, da einerseits Ontologien generell unabhängig von einer Plattform erstellt werden können und andererseits die Bildung einer sHGS-DWH an kein spezifisches Werkzeug oder an eine bestimmte Systemumgebung gebunden ist (vgl. Abschnitt 8.1.1). Die Speicherung einer Ontologie in einer sHGS-DWH kann generell unabhängig von einem DBMS erfolgen (vgl. Abschnitt 8.2.4). Zudem ermöglicht der Einsatz von Kopplungskomponenten, dass keine Beschränkung aufgrund des DBMS eines DWH-Systems bei dessen Anbindung an eine sHGS-DWH besteht.

Summa summarum erfüllt eine sHGS-DWH alle aufgestellten Formalziele.

8.4 Zusammenfassung

Den Kern dieses Kapitels bildete die Konzeption einer sHGS-DWH, mit deren Hilfe das gegebene Untersuchungsziel – die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen – erreicht wird. Hierfür wurde zunächst das Grundmotiv einer sHGS skizziert und eine Erweiterung zu einer ontologiebasierten Lösung für DWH-Systeme (sHGS-DWH) aufgezeigt. Die anschließenden Ausführungen konzentrierten sich auf die Einbindung einer sHGS-DWH in eine Gesamt-Architektur. Zum Schwerpunkt der Erläuterungen einer möglichen Gesamt-Architektur gehörte die explizite Darlegung, auf welche Weise semantische Heterogenität bei multiplen DWH-Systemen überwunden werden kann. Im Hinblick darauf wurden Kopplungskomponenten zur Anbindung bestehender DWH-Systeme, ein Versionierungskonzept sowie grundlegende Funktionen für eine sHGS-DWH eingeführt und detailliert beschrieben. Den Abschluss des Kapitels prägte eine kritische Würdigung einer sHGS-DWH. Dabei wurden insbesondere der Nutzen und die Unterstützung für Entwickler und Endanwender eines DWH-Systems herausgestellt.

9 Fallstudie zu einer semantischen Homogenisierungsschicht für Data-Warehouse-Systeme

Zentraler Gegenstand dieses Kapitels ist es, anhand des prototypischen Einsatzes einer sHGS-DWH für zwei aus dem Projekt CEUS^{HB} stammende Ausschnitte multipler DWH-Systeme die Erreichung des Untersuchungsziels zu belegen. Nach der Einführung der Fallstudie wird zunächst die Modellierung von DWH-Schemata beleuchtet. Anschließend erfolgt exemplarisch für die DWH-Systeme der Fallstudie die Überwindung deren semantischer Heterogenität.

9.1 Einführung der Fallstudie

Einleitend wird zunächst die Qualifikation der DWH-Systeme aus dem Projekt CEUS^{HB} als Repräsentant für den gegebenen Untersuchungsgegenstand aufgezeigt und anschließend die Fallstudie vorgestellt.

FÜHRUNGSMFORMATIONEN IM HOCHSCHULWESEN

Im Zuge ihres Reformprozesses haben die Hochschulen eine Stärkung ihrer Autonomie und Eigenverantwortung bei gleichzeitiger Reduzierung staatlicher Zuständigkeiten erfahren.¹¹⁰ Zielvereinbarungen nehmen in diesem Zusammenhang einen zunehmenden Stellenwert als zentrales Steuerungselement für die Hochschulen ein. Darüber hinaus befinden sich die Hochschulen verstärkt im (inter-)nationalen Wettbewerb um Studierende, Forschungsgelder und Reputation [SUPH07a].

Zur Bewältigung dieser Herausforderungen stellt die Verfügbarkeit valider und vergleichbarer Daten für das Hochschulmanagement eine elementare Voraussetzung dar [SUPH07a, S. 9]. Auf diese gestützt können – im Hinblick auf das Bestehen und den Erfolg in den genannten Wettbewerbsfeldern – fundierte Entscheidungen zur zielgerichteten Lenkung und Gestaltung der leistungserstellenden Prozesse einer Hochschule getroffen werden (vgl. Abschnitt 2.1.3).

¹¹⁰ Detaillierte Informationen zum Reformprozess der Hochschulen finden sich z. B. in der Begründung zum Entwurf des Bayerischen Hochschulgesetzes (BayHSchG) 2006 http://www.stmwfk.bayern.de/hs_hochschulgesetz.html (Abruf am 10.06.2008).

Daneben existieren auch im Hochschul Umfeld Ansätze, die mit dem Begriffspaar Mergers&Acquisitions in der Privatwirtschaft vergleichbar sind (vgl. Abschnitt 2.1.2). Das Streben nach einer höheren Effizienz und Qualität im Hochschulwesen kann auch bei Hochschulinstitutionen zu deren Zusammenschluss führen. Erfahrungen hierzu sind bspw. in [Skod99] beschrieben.

Hochschulen benötigen somit einerseits ein Hilfsmittel zum Aufbau eines homogenen sowie fundierten Berichtswesens (z. B. ein DWH-System). Andererseits besitzen sie – vergleichbar mit verteilten Organisationen – Strukturmerkmale, die den Aufbau einer zentralen, einheitlichen DWH-Lösung beeinträchtigen und das Entstehen einer heterogenen DWH-Landschaft sowohl innerhalb einer Hochschule als auch hochschulübergreifend begünstigen (vgl. zusätzlich Abschnitt 2.1.3). Letztere wirken sich insbesondere auf die DWH-Systeme aus, die im Rahmen des Projekts CEUS^{HB} verwaltet werden. Diese sind daher zu einem bestimmten Grad als multipel einzustufen (vgl. zusätzlich Abschnitt 1.3).

Die geschilderten Charakteristika des Hochschulwesens sowie die dargelegten (Architektur-)Merkmale der DWH-Systeme des Projekts CEUS^{HB} rechtfertigen, diese als Repräsentant für multiple DWH-Systeme in verteilten Organisationen und somit als empirische Grundlage für die Fallstudie einer sHGS-DWH einzusetzen.

VORSTELLUNG DER FALLSTUDIE

Mit Bezug zu den in Abschnitt 1.3 eingeführten, multiplen DWH-Systemen der *Hochschulen A* und *B* sowie den dort geschilderten Herausforderungen hinsichtlich der Vergleichbarkeit von Berichten, soll an dieser Stelle insbesondere die Vergleichbarkeit von Studienanfängerberichten dieser beiden Hochschulen als Untersuchungsgegenstand dienen. Eine Übereinstimmung dieser Berichtsdefinitionen kann bspw. im Hinblick auf ein einheitliches Lenkungssystem des StMWFK vonnöten sein.¹¹¹

Die in der Einleitung zu dieser Arbeit betrachteten DWH-Systeme von Hochschule A und Hochschule B verfügen beide über Kennzahlen, um die Anzahl der Studienanfänger zu ermitteln (vgl. Abbildung 1-3). Diese werden bei Hochschule A anhand der Fachse-

¹¹¹ Anzumerken ist, dass das StMWFK keinen Zugriff auf die DWH-Systeme der Hochschulen besitzt bzw. besitzen darf, die Hochschulen jedoch im beschränkten Rahmen Berichtsergebnisse aus ihren DWH-Systemen zur Entscheidungsunterstützung an das StMWFK melden.

mester ermittelt (Kennzahl: *Studenten im 1. FS*). Für Hochschule B wird dagegen von einer Definition auf Basis der Hochschulsemester ausgegangen (Kennzahl: *Studienanfänger*). Des Weiteren bestehen zwischen den genannten DWH-Systemen Unterschiede in der Benennung der Schemaelemente sowie in der Ausgestaltung der Dimensionen. Eine direkte Vergleichbarkeit der Studienanfängerberichte von Hochschule A und B ist demnach nicht gegeben.

Mithilfe einer sHGS-DWH ist die semantische Heterogenität dieser beiden DWH-Systeme in Bezug auf die skizzierte Problematik vergleichbarer Studienanfängerberichte zu überwinden (vgl. Abschnitt 9.3). Im Hinblick darauf wird im Folgenden zunächst die Modellierung von DWH-Schemata unter Einsatz einer sHGS-DWH beschrieben und eine Domain-Ontologie für ein DWH-System mit Studentendaten entwickelt.

Der Aufbau und die Nutzung einer sHGS-DWH für die gewählte Fallstudie erfolgt mit Unterstützung des Werkzeugs Protégé (vgl. Abschnitt 6.3.3).

9.2 Modellierung von Data-Warehouse-Schemata

Ziel des vorliegenden Abschnitts ist der Nachweis, dass mithilfe einer sHGS-DWH DWH-Schemata mit hoher struktureller und semantischer Verwandtschaft modelliert werden können. Zu diesem Zweck wird zunächst eine Domain-Ontologie für ein DWH-System mit Studentendaten eingeführt, welche anschließend als Referenzmodell zur Modellierung konkreter DWH-Schemata in Form von Applikationsontologien fungiert.

VORBEREITUNG EINER SEMANTISCHEN HOMOGENISIERUNGSSCHICHT FÜR DATA-WAREHOUSE-SYSTEME

Für die Erstellung einer Domain-Ontologie innerhalb einer sHGS-DWH bildet die Top-Level-Ontologie DWH-OWL die Vorlage (vgl. Abschnitt 8.1.2). Die DWH-OWL wird daher in einem ersten Schritt über ihren URI in eine sHGS-DWH importiert. Im vorliegenden Fall erfolgt dies unter Nutzung des Werkzeugs Protégé.

Der mit dem Buchstaben *A* gekennzeichnete Bereich (Abbildung 9-1) listet die in der DWH-OWL definierten Klassen in Baumstruktur auf. Deren zugehörigen Beschreibungen sind im Bildschirmausschnitt *B* und die jeweils hinterlegten Restriktionen in den Bereichen *C* und *D* dargestellt. In der nachfolgend gewählten Ansicht sind die Einschränkungen bezüglich der Kombinierbarkeit der Klassen aufgezeigt.

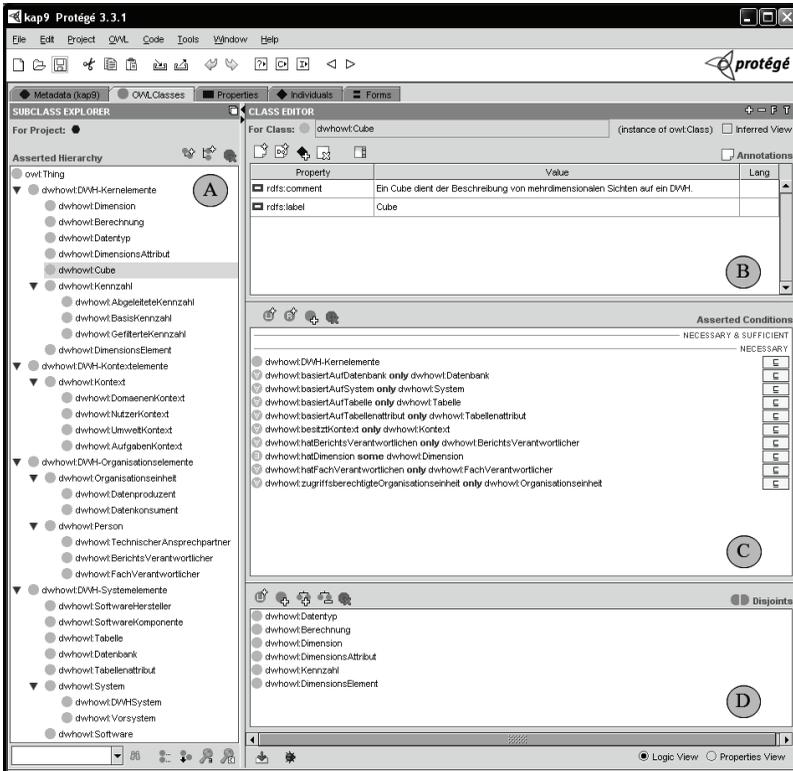


Abbildung 9-1: Darstellung der DWH-OWL mithilfe des Editors Protégé

ENTWICKLUNG EINER DOMAIN-ONTOLOGIE FÜR STUDENTENDATEN

Die in einer Domain-Ontologie für Studentendaten benötigten Klassen werden in einer sHGS-DWH jeweils als Unterklassen zur DWH-OWL definiert und semantisch konkretisiert. Zur Beschränkung der Komplexität der Fallstudie werden die von der DWH-OWL bereitgestellten Rollen unverändert übernommen, um die Relationen zwischen den Klassen der Domain-Ontologie auszuzeichnen. Selbiges gilt für die Kardinalitätsbedingungen (nachfolgend nicht visualisiert). Abbildung 9-2 stellt einen Ausschnitt der Domain-Ontologie für Studentendaten dar.

Die Dimensionselemente der Domain-Ontologie für Studentendaten sind grundsätzlich mit Kennzahlen der Klasse *Studentenkennzahlen* kombinierbar. Lediglich bei den Elementen der Dimension *Zeit* besteht eine Einschränkung hinsichtlich ihrer Aggregierbarkeit mit *Studentenkennzahlen*. Diese Besonderheit ist durch den Eigenschaftstyp *istSemiaggregierbar* gekennzeichnet. Des Weiteren sind Kennzahlen und Dimensionsattribute entsprechenden Datentypen zugewiesen (z. B. *Zeichenkette*, *Ganzzahl*).

Die Klasse *Anzahl Studenten* verkörpert eine elementare Kennzahl (Basiskennzahl). Auf dieser basieren zwei gefilterte Kennzahlen zur Berechnung von Studienanfängerkennzahlen (*Studienanfänger (1 FS)* und *Studienanfänger (1 HS)*). Deren qualitative Einschränkung ist durch eine Verknüpfung mit dem Dimensionselement *Fachsemester* bzw. *Hochschulsemester* charakterisiert. Der Filterwert ist als Literal hinterlegt (Fachsemester = 1, Hochschulsemester = 1). Exemplarisch ist zudem in der Domain-Ontologie eine Anteilsberechnung für Studienanfänger (*Anteil Studienanfänger (1FS)*) beschrieben. Die zugehörige Berechnung ist als Division auf Basis der gefilterten Kennzahl *Studienanfänger (1 FS)* und der Basiskennzahl *Anzahl Studenten* definiert.

Die Domain-Ontologie für Studentendaten bedient sich außerdem der in Abschnitt 8.2.1 eingeführten Rollen zur Auszeichnung von Ähnlichkeit. Zusätzlich zur DWH-OWL wird daher die für diesen Zweck erforderliche Ontologie (*semPro*) importiert. Zwischen den Kennzahlen *Studienanfänger (1 FS)* und *Studienanfänger (1 HS)* ist geringstenfalls eine so genannte *semantische Ähnlichkeit* (resemblance) zu identifizieren, da die beiden Kennzahlen dieselbe Rolle in ihrem jeweiligen Kontext besitzen (vgl. Abschnitt 5.3.5). Repräsentativ für weitere Ähnlichkeitsrelationen in der beschriebenen Domain-Ontologie wird diese Relation mit der ihrem semantischen Ähnlichkeitsgrad entsprechenden Rolle (*semPro:semanticResemblance*) explizit dokumentiert (in Abbildung 9-2 nicht visualisiert).

Die in der Domain-Ontologie spezifizierten Klassen verfügen über jeweils eigene Beschreibungselemente (z. B. *rdfs:label*, *rdfs:comment*), welche die in der DWH-OWL spezifizierten Konzepte semantisch detaillieren. Des Weiteren sind die Restriktionen hinsichtlich der Kombinierbarkeit der Klassen auf Basis der von der DWH-OWL bereitgestellten Rollen zusätzlich konkretisiert. Für die Dimension *Zeit* sind bspw. nur Dimensionselemente der Klassen *Semester*, *Semesterart* und *Studienjahr* zulässig.

QUELLTEXT DER DOMAIN-ONTOLOGIE FÜR STUDENTENDATEN

Der folgende Quellcode stellt einen Ausschnitt der Domain-Ontologie für Studentendaten in OWL-RDF-Syntax dar. Am Beispiel der Klasse Studiengang werden dabei deren semantische Auszeichnung, die Deklaration als Unterklasse zur Klasse *DimensionsElement*, die gültigen Aggregationsbeziehungen und Kennzahlenverknüpfungen, die Zuweisung von Dimensionsattributen sowie die Kennzeichnung von disjunkten Klassen aufgezeigt.

```
<?xml version="1.0"?>
<rdf:RDF NAMESPACE-Deklaration >
  <!-- Import der DWH-OWL -->
  <owl:Ontology rdf:about="">
    <owl:imports rdf:resource="http://www.semantic-metadata.com/
      dwhowl/dwhowl#"/>
  </owl:Ontology>
  <!-- Klassendefinition -->
  <owl:Class rdf:about="#Studiengang">
    <!-- Semantische Auszeichnung -->
    <rdfs:label>Studiengang</rdfs:label>
    <rdfs:comment>Als Studiengang wird eine abgeschlossene, in
      Studien- und Prüfungsordnungen vorgesehene berufsquali-
      fizierende oder berufsbezogene Hochschulausbildung
      bezeichnet. [...]
    </rdfs:comment>
    <!-- Kennzeichnung als Unterklasse zu DimensionsElement -->
    <rdfs:subClassOf rdf:resource="http://www.semantic-
      metadata.com/dwhowl/dwhowl#DimensionsElement"/>
    <!-- Festlegung gültiger Aggregationsbeziehungen -->
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="http://www.semantic-
          metadata.com/dwhowl/dwhowl#verdichtetZu"/>
        <owl:allValuesFrom>
          <owl:Class>
            <owl:unionOf rdf:parseType="Collection">
              <owl:Class rdf:about="#Abschluss"/>
              <owl:Class rdf:about="#Studienfach"/>
            </owl:unionOf>
          </owl:Class>
        </owl:allValuesFrom>
      </owl:Restriction>
    </rdfs:subClassOf>
```

```

<!-- Bestimmung gültiger Kennzahlen-Verknüpfungen -->
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="http://www.semantic-
      metadata.com/dwhowl/dwhowl#kombinierbarMitKennzahl"/>
    <owl:allValuesFrom>
      <owl:Class rdf:about="#StudentenKennzahlen"/>
    </owl:allValuesFrom>
  </owl:Restriction>
</rdfs:subClassOf>
<!-- Referenzierung der zugehörigen Dimensionsattribute -->
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="http://www.semantic-
      metadata.com/dwhowl/dwhowl#hatDimensionsAttribut"/>
    <owl:allValuesFrom>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Kennzeichen"/>
          <owl:Class rdf:about="#Beschreibung"/>
        </owl:unionOf>
      </owl:Class>
    </owl:allValuesFrom>
  </owl:Restriction>
</rdfs:subClassOf>
<!-- Definition disjunkter Klassen -->
<owl:disjointWith rdf:resource="#Abschluss"/>
<owl:disjointWith rdf:resource="#Studienfach"/>
<owl:disjointWith rdf:resource="#Lehreinheit"/>
<owl:disjointWith rdf:resource="#Fakultaet"/>
[...]
</owl:Class>
</rdf:RDF>

```

Quellcode 5: Domain-Ontologie für Studentendaten (Ausschnitt)

MODELLIERUNG VON DATA-WAREHOUSE-SCHEMATA

Gestützt auf die Domain-Ontologie für Studentendaten können durch den Einsatz einer SHGS-DWH konkrete DWH-Schemata mit hoher struktureller und semantischer Verwandtschaft erstellt werden. Diese werden in Form einer Applikationsontologie entweder mittels Unterklassebeziehungen oder als Ausprägung (Individuen) der Domain-Ontologie beschrieben (vgl. Abschnitt 8.1.2).

Die (semantisch) korrekte Verwendung der Elemente der Domain-Ontologie bzw. der DWH-OWL wird dabei aufgrund der in den jeweiligen Ontologien formal erfassten und semantisch annotierten Restriktionen gewährleistet. Exemplarisch sei hier die oben genannte Verknüpfungsbeschränkung zur Dimension *Zeit* genannt (*dwhowl:hatDimensionsElement*), die nur Dimensionselemente der Klassen *Semester*, *Semesterart* und *Studienjahr* umfassen darf. Eine diesbezügliche Prüfung einer (Applikations-)Ontologie auf Konsistenz kann Modellierungsfehler aufdecken (z. B. falsche (semantische) Verwendung der definierten Rollen, Nicht-Einhaltung der Kardinalitätsrestriktionen).

Die geforderte hohe semantische Verwandtschaft der mithilfe einer sHGS-DWH definierten DWH-Schemata stützt sich insbesondere auf die semantisch reichhaltige Auszeichnung der Konzepte der Domain-Ontologie. Durch diese ist die Bedeutung der Schemaelemente genau erfasst und einem Modellierer bei der Ausgestaltung eines konkreten DWH-Schemas, ergo einer Applikationsontologie, offen gelegt. Den Dimensionselementen und Kennzahlen der beschriebenen Domain-Ontologie sind bspw. die offiziellen Definitionen für die Studenten- und Prüfungsstatistik¹¹³ des Statistischen Bundesamts zugrunde gelegt (vgl. Quellcode 5). Die Spezialisierung eines Schemaelements in einer Applikationsontologie bezieht stets dessen übergeordnete Beschreibung mit ein. DWH-Schemata, die von derselben Domain-Ontologie abgeleitet wurden, besitzen somit nicht nur strukturell sondern auch semantisch eine hohe Verwandtschaft. Dies gilt für die Modellierung sowohl unter Nutzung der Unterklassebeziehung als auch der Ausprägungsebene (Individuen) der Domain-Ontologie.

Abbildung 9-3 visualisiert einen Modellierungsschritt innerhalb einer sHGS-DWH. In dem mit dem Buchstaben *A* beschrifteten Bereich sind die Klassen der DWH-OWL (*dwhowl*) sowie der Domain-Ontologie für Studentendaten (*do_stud*) entsprechend ihrer taxonomischen Abhängigkeiten dargestellt. Die rechts daneben angeordnete Spalte *B* listet die Individuen einer Klasse auf (hier: *ao_hsx:Studienanfänger* als Ausprägung der Klasse *do_stud:AnzahlStudienanfänger_FS*). Individuen repräsentieren bei der hier gewählten Modellierungsvariante die Elemente eines konkreten DWH-Schemas (Ausprägung einer Domain-Ontologie). Bildschirmsegment *C* beinhaltet die Auszeichnungselemente zur semantischen Spezifikation eines Schemaelements.

¹¹³ Vgl. http://www.statistik.bayern.de/imperia/md/content/lfstad/definition_teil_1.pdf (Abruf am 10.06.2008).

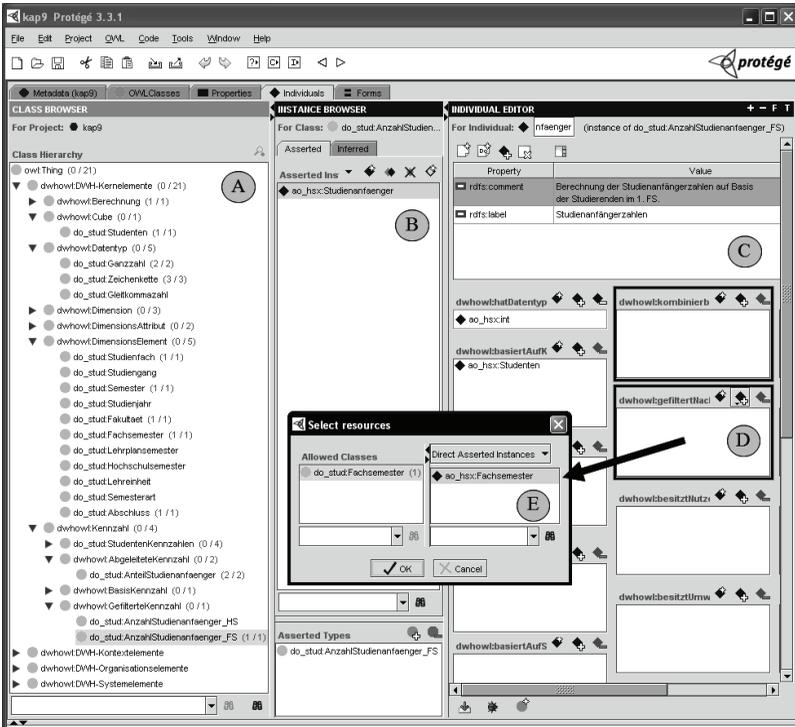


Abbildung 9-3: Bildschirmkopie eines Modellierungsschritts innerhalb einer shGS-DWH

Durch die exklusive Bereitstellung nur der für ein spezifisches Schemaelement gültigen Beziehungen kann der Modellierungsprozess zusätzlich gestützt werden. Elementare Fehler bei der Modellierung werden somit bereits im Vorfeld unterbunden. Die einzelnen Felder im Bereich *D* zeigen die erlaubten Verknüpfungen (Rollen) für ein gewähltes Schemaelement. Obligatorische Rollen sind darüber hinaus durch eine zusätzliche Umrandung hervorgehoben.

Zur Definition einer konkreten Beziehung zwischen zwei Schemaelementen werden in einem eigenen Auswahlfenster (hier mit dem Buchstaben *E* gekennzeichnet) nur die für eine Rolle zulässigen Klassen und Individuen zur Auswahl angeboten. Diese Eigenschaft konstituiert ein weiteres Kriterium zur Vermeidung von Modellierungsfehlern. Im Beispiel ist dieser Sachverhalt anhand der Rolle *dwhowl:gefiltertNach* aufgezeigt. Die vorliegende Domain-Ontologie fordert für die Definition einer Studienanfänger-kennzahl vom Typ *do_stud:AnzahlStudienanfenger_FS* (gefilterte Kennzahl) die An-

gabe des Schemaelements, welches die Fachsemester repräsentiert und somit das qualitative Merkmal für die Filterung darstellt. Entsprechend weist das Auswahlfenster *E* nur die Klasse *do_stud:Fachsemester* mit dem zugehörigen Individuum auf.

Eine abschließende Konsistenzprüfung der modellierten Ontologie – unabhängig von der gewählten Modellierungsvariante – gibt Aufschluss über die hinsichtlich struktureller und semantischer Merkmale korrekte sowie hinsichtlich der als zwingend deklarierten Rollen vollständige Definition eines DWH-Schemas.

VISUALISIERUNG EINER ONTOLOGIE

Nachfolgende Abbildung visualisiert den Ausschnitt einer möglichen Applikationsontologie. Zur Darstellung wurde die Erweiterung *TGVizTab*¹¹⁴ für den Editor Protégé gewählt. Die Typen der gezeigten Klassen oder Individuen und ihrer Beziehungen sind dabei durch unterschiedliche Farben gekennzeichnet. Neben zahlreichen Möglichkeiten, die Illustration gemäß den eigenen Ansprüchen anzupassen (z. B. Zoom, Rotation, Wahl der dargestellten Verknüpfungstiefe, farbliche Hervorhebung), bietet TGVizTab auch Unterstützung für eine semantische Navigation innerhalb einer Ontologie. Durch Auswahl eines bestimmten Elements rückt dieses ins Zentrum der Darstellung. Gemäß der gewählten Darstellungstiefe werden über die jeweiligen Rollen zugleich die abhängigen Elemente ausgewiesen. Von diesen kann wiederum zu weiteren, abhängigen Elementen navigiert werden.

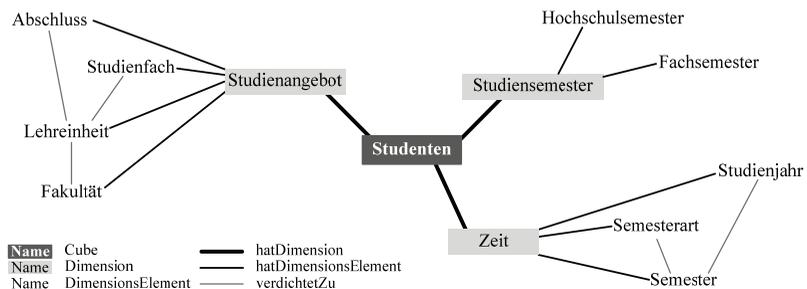


Abbildung 9-4: Ontologiedarstellung in Protégé mithilfe der Erweiterung TGVizTab

¹¹⁴ Vgl. <http://users.ecs.soton.ac.uk/ha/TGVizTab/> (Abruf am 10.06.2008).

9.3 Überwindung der semantischen Heterogenität

Im vorausgegangenen Abschnitt wurde die Modellierung eines DWH-Schemas mithilfe einer sHGS-DWH aufgezeigt und ein Ausschnitt einer Domain-Ontologie für Studentendaten vorgestellt. Diese Domain-Ontologie bildet im Folgenden einen elementaren Baustein einer sHGS-DWH, um die semantische Heterogenität der multiplen DWH-Systeme der Fallstudie zu überwinden. Einleitend wird hierfür eine Kopplung dieser DWH-Systeme an eine sHGS-DWH skizziert.

9.3.1 Kopplung der Data-Warehouse-Systeme der Fallstudie an eine semantische Homogenisierungsschicht

Zur Kopplung der multiplen DWH-Systeme der Hochschulen A und B an eine sHGS-DWH wird auf die in Abschnitt 8.2.2 erläuterten Kopplungskomponenten zurückgegriffen. Vereinfachend wird angenommen, dass im Vorfeld bereits die zur Beschreibung der DWH-Systeme relevanten Metadaten ermittelt wurden.

Exemplarisch sind nachstehend Metadaten zu den Dimensionen des DWH-Systems von Hochschule A abgebildet. Zur Realisierung ihres DWH-Systems nutzt Hochschule A Komponenten der *Java-basierten DWH-Toolsuite*¹¹⁵ (JDW-Toolsuite). Die Struktur der folgenden Tabelle ist demnach durch die JDW-Toolsuite vorgegeben.

| DimID | Dimensionsname | SemiAgg | Beschreibung | Reihenfolge | ModellID |
|-------|-------------------|---------|----------------------|-------------|----------|
| 1 | Semesterdimension | Ja | Zeitmerkmale | 2 | 1 |
| 2 | Studium | Nein | Merkmale zum Studium | 1 | 1 |
| 3 | Anzahl Semester | Nein | Semesterangaben | 3 | 1 |
| ▶ | | | | | |

Abbildung 9-5: Metadaten zu den Dimensionen eines DWH-Systems (JDW-Toolsuite)

Die Metadaten sind zunächst in eine flache Datei, vornehmlich im Format CSV¹¹⁶ (*Comma-Separated Values*), zu exportieren (Daten-Subsystem).

¹¹⁵ Vgl. <http://ceus.uni-bamberg.de/jdwtoolsuite/> (Abruf am 10.06.2008).

¹¹⁶ Detaillierte Informationen zum CSV-Format finden sich unter <http://rfc.net/rfc4180.html> (Abruf am 10.06.2008).

Als Heterogenitäts-Subsystem wurde in Abschnitt 8.2.2 ein Vorgehen analog zur Unterstützung durch das Werkzeug *Convert To RDF* vorgeschlagen. *Convert To RDF* ist in erster Linie für die Arbeit mit RDF-Dokumenten konzipiert. Es soll jedoch an dieser Stelle vornehmlich der Veranschaulichung dienen, wie Metadaten eines DWH-Systems in eine (Applikations-)Ontologie überführt werden können.

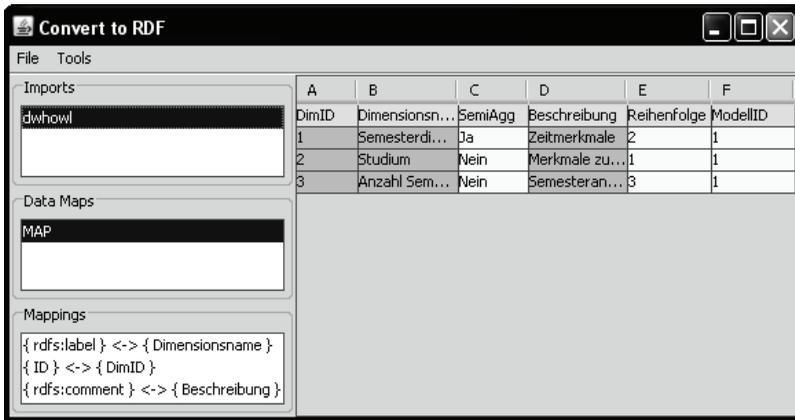


Abbildung 9-6: Bildschirmkopie zum Werkzeug Convert To RDF

In der gezeigten Bildschirmkopie (Abbildung 9-6) ist im Feld oben links das importierte RDF-Schema bzw. die importierte Ontologie ausgewiesen. Unterhalb dessen sind entsprechende Abbildungsvorschriften genannt sowie die konkreten Zuweisungen zwischen den Feldnamen in der CSV-Datei und den Elementen in einem RDF-Schema bzw. in einer Ontologie dargestellt (das Feld *Dimensionsname* wird z. B. durch *rdfs:label* erfasst). Durch die Festlegung dieser Allokationen wird ein entscheidender Grundstein für die spätere Überwindung der semantischen Heterogenität gelegt. Der rechte Fensterbereich stellt die Feldnamen und zugehörigen Daten der CSV-Datei dar.

Die Generierung eines RDF-Dokuments erfolgt gemäß den definierten Zuordnungen. Allerdings unterstützt das hier vorgestellte Werkzeug jeweils nur die Erstellung von RDF-Dokumenten zu einer bestimmten Klasse (hier: Dimension). Die einzeln erzeugten RDF-Dokumente sind daher in einem gemeinsamen Dokument zusammenzuführen. Zugleich können die (semantischen) Auszeichnungen der Schemaelemente z. B. durch einen Fachexperten manuell angereichert werden. Das resultierende Dokument ist abschließend mit einem URI zu kennzeichnen (Kommunikations-Subsystem), über den es

in eine sHGS-DWH importiert werden kann (vgl. Abschnitt 8.2.2). Das skizzierte Vorgehen ist mit entsprechender Werkzeugunterstützung analog für Applikationsontologien realisierbar.

9.3.2 Ermittlung semantischer Beziehungen zwischen den Data-Warehouse-Systemen der Fallstudie

Die Kopplung der DWH-Systeme der Fallstudie an eine sHGS-DWH wurde im vorherigen Abschnitt in Eckpunkten vorgestellt. Im nächsten Schritt werden die zu den DWH-Systemen der Hochschulen A und B generierten Applikationsontologien mithilfe der OWL-Importanweisung (*owl:imports*) in einer sHGS-DWH zusammengeführt.¹¹⁷ Jede Applikationsontologie sollte für dieses Ziel nach der gleichen Modellierungsvariante (Unterklassbeziehungen vs. Ausprägungen zu einer Domain-Ontologie) erstellt worden sein. Über ihren URI können die Applikationsontologien innerhalb einer sHGS-DWH identifiziert und somit einer Hochschule genau zugeordnet werden.

Die beiden folgenden Darstellungen visualisieren die zu den DWH-Systemen der Hochschulen A und B zugehörigen Applikationsontologien. Deren Inhalte entsprechen den im ersten Kapitel vorgestellten konzeptuellen DWH-Schemata (vgl. Abbildung 1-3). Die Struktur der Graphiken folgt der der Domain-Ontologie (vgl. Abbildung 9-2).

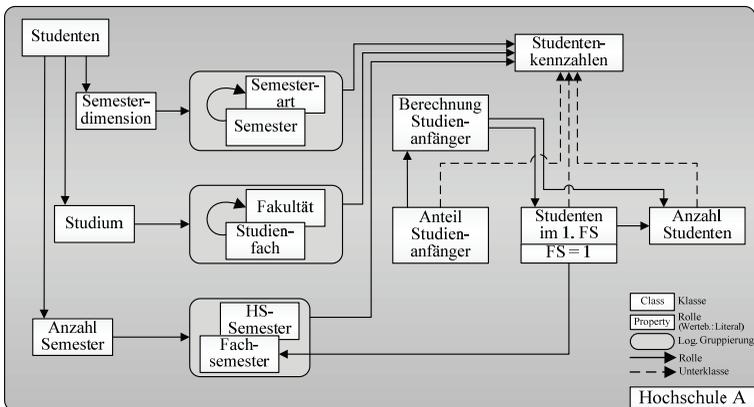


Abbildung 9-7: Applikationsontologie zum DWH-System der Hochschule A

¹¹⁷ Die in einer Applikationsontologie referenzierten Domain- oder Top-Level-Ontologien werden beim Importvorgang automatisch miteinbezogen.

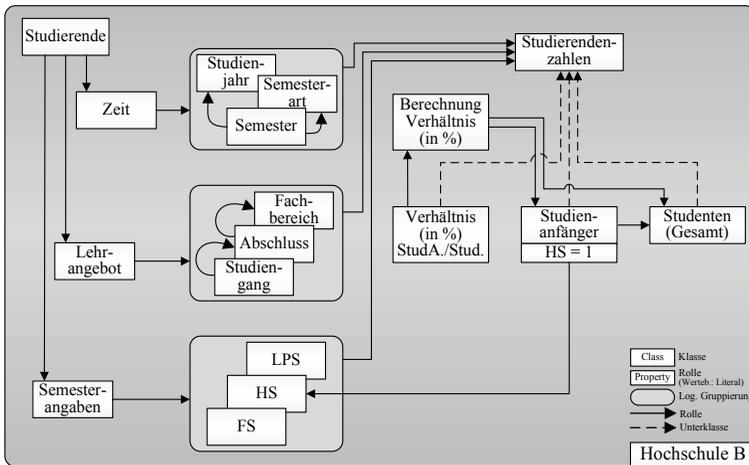


Abbildung 9-8: Applikationsontologie zum DWH-System der Hochschule B

Das Ergebnis der Importanweisung der beiden Applikationsontologien in eine sHGS-DWH zeigt die nachstehende Bildschirmkopie (Abbildung 9-9). Bei der Definition der Applikationsontologien wurde im vorliegenden Fall die Modellierung als Unterklasse zu einer Domain-Ontologie gewählt, d. h. ein Element eines DWH-Schemas ist durch eine Klasse in der Applikationsontologie repräsentiert.

Im Fensterbereich *A* sind anhand der Kürzel für die Namensräume die jeweiligen Ontologien sowie die taxonomischen Beziehungen zwischen den Klassen zu erkennen. Auf oberster Ebene befindet sich – gemäß der Definition einer sHGS-DWH – die DWH-OWL (*dwhowl*) als Vertreter für eine Top-Level-Ontologie. Von dieser abhängig ist die Domain-Ontologie für Studentendaten ausgewiesen (*do_stud*). Auf der untersten Ebene sind die beiden Applikationsontologien zu den multiplen DWH-Systemen von Hochschule A (*ao_hsa*) und B (*ao_hsb*) dargestellt. Die Veranschaulichung der Abhängigkeiten erfolgt durch das hier gewählte Werkzeug gemäß der in den Ontologien definierten Beziehungen und Restriktionen.

Bildschirmsegment *B* enthält die semantischen Erläuterungen zu einem Schemaelement. Im Bereich *C* sind alle direkt spezifizierten sowie geerbten Restriktionen hinsichtlich der gültigen und notwendigen Beziehungen (Rollen) zu anderen Schemaelementen (hier: Klassen) dargelegt. Die in der folgenden Abbildung exemplarisch markierte, gefilterte Kennzahl *ao_hsb:Studienanfänger* erbt von der Domain-Ontologie (*do_stud*) z. B. die Filterung auf den Wert 1.

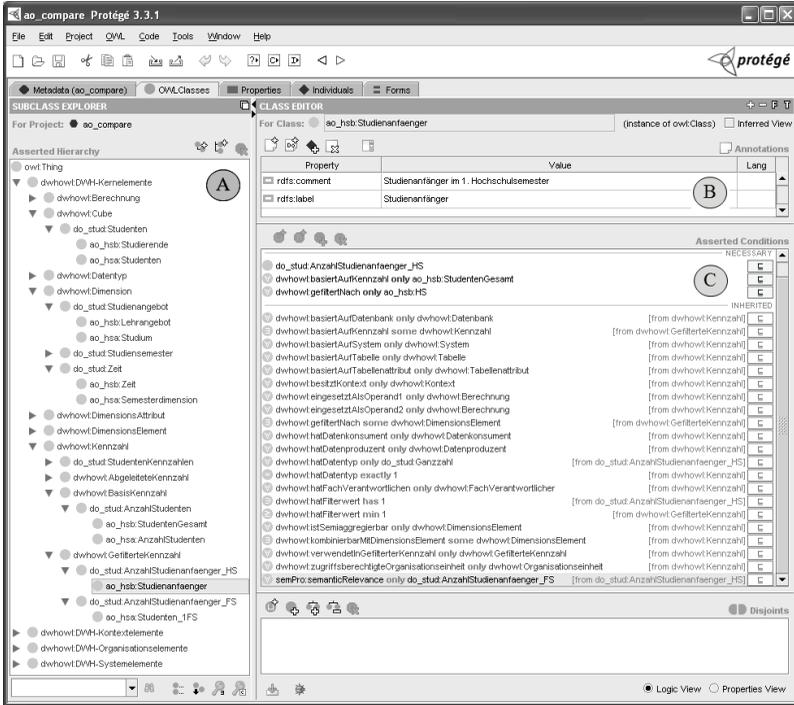


Abbildung 9-9: Gesamtansicht auf die zusammengeführten Applikationsontologien der Fallstudie in einer sHGS-DWH

Gestützt auf die hier dargestellte sHGS-DWH wird im Folgenden die Überwindung der semantischen Heterogenität der multiplen DWH-Systeme von Hochschule A und B exemplarisch aufgezeigt. Zu diesem Zweck wird auf die in Abschnitt 1.3 identifizierten Problemkategorien zurückgegriffen (*unterschiedliche Benennung bei gleicher Bedeutung, vermeintlich gleiche Kennzahlendefinitionen, verschiedenartige Ausgestaltung der Dimensionen*). Diese dienen als Leitlinie, um die Erreichung des Untersuchungsziels zu belegen.

UNTERSCHIEDLICHE BENENNUNG BEI GLEICHER BEDEUTUNG

Die beiden Applikationsontologien beinhalten unterschiedlich benannte Schemaelemente, die jedoch grundlegend dieselbe Ressource widerspiegeln.¹¹⁸ Die Abhängigkeit bestimmter Schemaelemente von einem gemeinsamen, übergeordneten Konzept erlaubt diesen Rückschluss (z. B. *ao_hsa:Studium* und *ao_hsb:Lehrangebot* stellen Unterklassen zu *do_stud:Studienangebot* dar). Eine semantische Beziehung zwischen diesen ungleich benannten Schemaelementen kann über die mit dem übergeordneten Konzept geteilten Eigenschaften (z. B. Beschreibungen, Restriktionen) hergestellt werden. Aufgrund ihrer taxonomischen Abhängigkeiten ist evident, dass es sich bei *ao_hsa:Studium* und *ao_hsb:Lehrangebot* jeweils um Dimensionen handelt, die über qualitative Merkmale zum Studienangebot (*do_stud:Studienangebot*) verfügen.

Des Weiteren sollen in dieser Problemkategorie die Dimensionselemente *ao_hsa:Fakultaet* und *ao_hsb:Fachbereich* genannt werden, die ebenfalls über ihr gemeinsames, übergeordnetes Konzept (*do_stud:Fakultaet*) als semantisch zusammengehörig identifiziert werden können. Überdies können die beiden Dimensionselemente anhand ihrer semantisch reichhaltigen Beschreibungen als *semantisch äquivalent* eingestuft werden. Über die Rolle *semPro:semanticEquivalence* kann dieser Grad der Ähnlichkeitsrelation in einer sHGS-DWH explizit dokumentiert werden.¹¹⁹

Neben Synonymen können auf diese Weise auch weitere sprachliche Defekte wie z. B. Äquipollenzen oder falsche Bezeichner erkannt und entsprechend behandelt werden. In bestimmten Fällen ist es sogar möglich, z. B. durch eine Konsolidierung der Benennungen der Schemaelemente, eine Auflösung der semantischen Heterogenität zu erzielen.

Die Abhängigkeit eines Schemaelements von einem übergeordneten Konzept bestimmt zugleich dessen Typ eindeutig. Exemplarisch sei hier das Dimensionselement *do_stud:Fachsemester* genannt, welches ebenso als Kennzahl bspw. zur Berechnung der durchschnittlichen Studiendauer eines Absolventen existieren könnte. In diesem Fall wäre *do_stud:Fachsemester* von *dwhowl:Kennzahl* abgeleitet. Dieser im Vergleich zur hier betrachteten Problemkategorie umgekehrte Sachverhalt (anscheinend gleiche Sche-

¹¹⁸ Diese Problemkategorie kann als ein Pendant zur Synonym-Problematik interpretiert werden.

¹¹⁹ Alternativ kann zur Kennzeichnung der semantischen Äquivalenz zweier Klassen auch das von OWL bereitgestellte Sprachkonstrukt *owl:equivalentClass* eingesetzt werden.

maelemente besitzen eine unterschiedliche Intension; Homonym-Problematik) kann demnach ebenfalls aufgrund der Abhängigkeit eines Schemaelements von seinem übergeordneten Konzept aufgedeckt und differenziert werden. Im Anschluss werden zur genauen Unterscheidung vermeintlich gleicher Schemaelemente deren konkrete Definitionen beleuchtet. Dies erfolgt am Beispiel von Kennzahlen.

VERMEINTLICH GLEICHE KENNZAHLEDEFINITIONEN

Anhand der Erfassung und Darstellung in einer sHGS-DWH können die Definitionen von Schemaelementen (hier im Speziellen von Kennzahlen) korrekt ermittelt, interpretiert und differenziert werden. Als Vorlage für die folgende Erläuterung sollen die gefilterten Kennzahlen *ao_hsa:Studenten_IFS* und *ao_hsb:Studienanfänger* dienen, die beide zur Berechnung der Studienanfängerzahlen eingesetzt werden. Letztere ist gemäß ihrer explizierten Abhängigkeiten (*dwhowl:gefiltertNach*) auf Basis des Hochschulsemesters definiert (*ao_hsb:HS*), erstgenannte wird dagegen über das Fachsemester bestimmt (*ao_hsa:Fachsemester*). Den zugehörigen Filterwert erben beide Kennzahlen von ihrem Pendant in der Domain-Ontologie (*dwhowl:hatFilterwert* = 1). Die betrachteten Kennzahlen beruhen zudem auf Basiskennzahlen, die ein und demselben Konzept der Domain-Ontologie untergeordnet sind (*ao_hsa:AnzahlStudenten* und *ao_hsb:StudentenGesamt* sind direkte Unterklassen von *do_stud:AnzahlStudenten*). In ihrer Summe bewirken die geschilderten Merkmale, die jeweilige Intension der beiden Studienanfänger-Kennzahlen exakt zu erfassen und deren korrekte Verwendung zu lancieren.

Ein gewisser Grad an semantischer Heterogenität zwischen den beiden diskutierten Varianten von Studienanfängerkennzahlen ist jedoch offensichtlich, da sie ihre Werte in Bezug auf unterschiedliche qualitative Merkmale ermitteln. Die Überwindung dieser semantischen Lücke wird im vorliegenden Fall – neben der obligatorischen Beachtung der Auszeichnungselemente mit den Erläuterungen zur Semantik – durch die auf Ebene der Domain-Ontologie für Studentendaten referenzierte, semantische Ähnlichkeitsbeziehung *semPro:semanticResemblance* unterstützt. Diese ist in der genannten Domain-Ontologie zwischen Studienanfängerkennzahlen auf Basis des Fachsemesters und denen, die anhand des Hochschulsemesters berechnet werden, eingesetzt (vgl. Abschnitt 9.2). Diese Ähnlichkeitsbeziehung wird somit auch auf die entsprechenden Studienanfängerkennzahlen in den beiden hier untersuchten Applikationsontologien vererbt. Dort dokumentiert sie den Grad der semantischen Relation zwischen den beiden Kennzahlen zur Bestimmung der Studienanfänger bei Hochschule A und B.

Evident ist, dass an dieser Stelle eine Auflösung der semantischen Heterogenität nicht zulässig ist, da jede Kennzahl für sich über eine korrekte Definition zur Ermittlung der Studienanfängerzahlen verfügt. Die ermittelten Charakteristika und Relationen legen jedoch offen, welche (semantischen) Annahmen diesen Kennzahlen zugrunde liegen und in welcher Beziehung diese stehen. Somit wird ein maßgeblicher Beitrag geleistet, deren semantische Heterogenität zu überwinden.

VERSCHIEDENARTIGE AUSGESTALTUNG DER DIMENSIONEN

Die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen erfordert auch die Berücksichtigung struktureller Beziehungen zwischen Schemaelementen sowie der Position eines Schemaelements innerhalb eines DWH-Schemas (vgl. Abschnitt 5.5). Die Dimensionen der DWH-Systeme der Fallstudie, die von einem gemeinsamen Konzept in der Domain-Ontologie abhängen, bestehen zum Teil aus unterschiedlichen Dimensionselementen. Dementsprechend differieren auch die jeweiligen Aggregationspfade (*dwhowl:verdichtetZu*).

Die zwischen den Dimensionselementen der Dimension *do_stud:Studienangebot* festgelegten Aggregationspfade erlauben die Schlussfolgerung, dass *ao_hsa:Fakultaet* und *ao_hsb:Fachbereich* dieselbe und zugleich höchste Granularitätsstufe in den zugehörigen Dimensionen der betrachteten DWH-Systeme besitzen. Die Transitivität der Rolle *dwhowl:verdichtetZu* bewirkt zudem, dass – trotz des Fehlens eines von *do_stud:Lehrereinheit* abhängigen Dimensionselements – die jeweils in den Applikationsontologien definierten Hierarchien zum Studienangebot hinsichtlich der Vorgabe der Domain-Ontologie konform sind. Darüber hinaus wird mithilfe der hier erstellten sHGS-DWH transparent, dass *ao_hsb:Studiengang* einen höheren Detaillierungsgrad aufweist als *ao_hsa:Studienfach*. Ein Bericht der Hochschule A, der Studentenzahlen anhand der Studienfächer ausweist, ist folglich nicht direkt mit einem Bericht von Hochschule B vergleichbar, der analog Studentenzahlen bezüglich der Studiengänge aufschlüsselt.

Gemäß der offiziellen Definition des Statistischen Bundesamts wird ein Studiengang als Kombination der angestrebten Abschlussprüfung mit einem oder dem ersten Studienfach erfasst.¹²⁰ In der Domain-Ontologie für Studentendaten sind derartige Vorgaben und Erläuterungen abgebildet. Diese steigern in Kombination mit den oben gefolgerten Erkenntnissen die Potentiale, um mithilfe einer sHGS-DWH konkrete Maßnahmen zur Überwindung der semantischen Heterogenität in Bezug auf die hier betrachtete Problemkategorie herzuleiten. Im vorliegenden Szenario können dadurch bspw. Hinweise zur Erreichung einer direkten Vergleichbarkeit der skizzierten Studienanfängerberichte erschlossen werden (z. B. Aufnahme des Dimensionselements Abschluss in das DWH-System von Hochschule A, Gruppierung der Zahlen des Berichts von Hochschule B unabhängig vom Abschluss).

AUSWIRKUNGEN AUF DAS PROJEKT CEUS^{HB}

Anhand der Ausschnitte zweier multipler DWH-Systeme aus dem Kontext des Projekts CEUS^{HB} wurde mithilfe einer sHGS-DWH aufgezeigt, wie deren semantische Heterogenität überwunden werden kann. Übertragen auf das Projekt CEUS^{HB} kann eine sHGS-DWH zu einem landesweit homogenen Berichtswesen auf Basis der lokalen DWH-Systeme der Hochschulen beitragen. Somit ist auch eine auf einer einheitlichen Semantik beruhende Berichterstattung an das StMWFK realisierbar. Die exakte und einheitliche Abbildung der Begriffsdefinitionen (z. B. nach Vorgabe vom Statistischen Bundesamt) in den lokalen DWH-Systemen der Hochschulen bildet hierfür eine entscheidende Voraussetzung. Eine entsprechend ausgestaltete Domain-Ontologie kann in einer sHGS-DWH als Hilfsmittel dienen, um diesbezüglich konforme DWH-Systeme zu modellieren bzw. die semantische Heterogenität bestehender DWH-Systeme in Bezug auf die erwähnten definitorischen Vorgaben zu überwinden.

Ein weiteres Nutzungsszenario stellt – neben der soeben diskutierten Vergleichbarkeit der Berichte aus den lokalen DWH-Systemen der Hochschulen – die Realisierung eines hochschulübergreifenden Berichtswesens dar. Gemäß der dem Projekt CEUS^{HB} zugrunde liegenden hierarchisch verteilten DWH-Architektur (vgl. Abschnitt 1.3) kann eine sHGS-DWH dazu eingesetzt werden, eine Domain-Ontologie für das landesweite DWH-System des StMWFK aufzubauen. Von dieser Domain-Ontologie sowie den da-

¹²⁰ Vgl. http://www.statistik.bayern.de/imperia/md/content/lfstad/definition_teil_1.pdf, Blatt 215. (Abruf am 10.06.2008).

rin formal erfassten amtlichen Definitionen der Fachbegriffe und Kennzahlen sind die Applikationsontologien für die lokalen DWH-Systeme der Hochschule abzuleiten. Durch die somit geschaffene Relation zwischen den hochschuleigenen und dem landesweiten DWH-System ist ein hochschulübergreifendes Berichtswesen realisierbar. Eine sHGS-DWH kann darüber hinaus auch zur Konzeption so genannter domänenübergreifender Auswertungen dienen. Dazu sind für die bei CEUS^{HB} standardmäßig unterstützten Domänen *Studenten/Prüfungen*, *Mittelbewirtschaftung* und *Personalverwaltung* zunächst jeweils eigene Domain-Ontologien zu beschreiben. Eine Variante zur Vorbereitung domänenübergreifender Auswertungen besteht darin, die Beziehungen zwischen den Domänen auf Ebene ihrer Domain-Ontologien zu identifizieren und in Form einer Ontologie als Referenzvorgabe für eine Implementierung in einem neu aufzubauenden DWH-System zu erfassen. Bei einem bereits bestehenden DWH-System greift die Variante, domänenübergreifend Abhängigkeiten in den Applikationsontologien zu den genannten Domänen aufzudecken und formal zu erfassen. Diese aus einer sHGS-DWH abgeleiteten konzeptionellen Erkenntnisse bilden einen grundlegenden Baustein für den Aufbau eines domänenübergreifenden Berichtswesens.

9.4 Zusammenfassung

Im zurückliegenden Kapitel wurde dargelegt, wie mithilfe einer sHGS-DWH die semantische Heterogenität bei multiplen DWH-Systemen aus dem Umfeld des Projekts CEUS^{HB} überwunden werden kann. Zu diesem Zweck wurde zunächst eine Domain-Ontologie für Studentendaten vorgestellt. Eingebunden in eine sHGS-DWH diente diese als Referenzmodell, um exemplarisch ein DWH-System zu modellieren. Dabei konnte aufgezeigt werden, wie eine hohe strukturelle und semantische Verwandtschaft der abgeleiteten DWH-Schemata erzielt werden kann.

Im Hinblick auf die Überwindung der semantischen Heterogenität der hier untersuchten DWH-Systeme erfolgte eine Anleitung zur Kopplung dieser DWH-Systeme an eine sHGS-DWH. Entlang der in Abschnitt 1.3 beispielhaft herausgestellten Problemkategorien wurden anschließend die Unterstützungspotentiale einer sHGS-DWH zur Überwindung semantischer Heterogenität belegt. Die Ausschnitte zweier multipler, dem Projekt CEUS^{HB} entlehnter DWH-Systeme traten hierbei als Stellvertreter für multiple DWH-Systeme in verteilten Organisationen auf.

10 Zusammenfassung und Ausblick

Die vorliegende Arbeit widmete sich der seit geraumer Zeit intensiv diskutierten Problematik der Heterogenitätsüberwindung bei (historisch gewachsenen) Anwendungssystemen. Untersuchungsgegenstand waren multiple DWH-Systeme. Das verfolgte Untersuchungsziel lag in der Überwindung der semantischen Heterogenität multipler DWH-Systeme. In diesem Zusammenhang wurde ebenfalls die Realisierung einer semantisch reichhaltigen Metadaten-Beschreibung für DWH-Systeme verfolgt. Letztere stellt eine Vorbedingung dar, um semantische Heterogenität zu überwinden.

ZUSAMMENFASSUNG

Im Folgenden werden die zentralen Erkenntnisse dieser Arbeit zusammengefasst. Die Ausführungen orientieren sich am Aufbau der Arbeit.

Nach der Abgrenzung der Untersuchungssituation und Explikation der Problemstellung (*Kapitel 1*) wurden in *Kapitel 2* betriebliche Organisationsstrukturen vorgestellt sowie die Führungsstruktur im Hochschulwesen dargelegt. Dabei konnten Gemeinsamkeiten zwischen der verteilten Organisationskonfiguration und der Führungsstruktur im Hochschulwesen nachgewiesen werden. Einen weiteren Schwerpunkt bildete die Erörterung von FIS im Allgemeinen und DWH-Systemen im Speziellen. In diesem Zusammenhang wurden vor allem Ursachen für das Auftreten multipler DWH-Lösungen in betrieblichen Organisationen identifiziert.

Kapitel 3 konzentrierte sich auf Metadaten in DWH-Systemen. Konsistente Metadaten stellen eine Voraussetzung für die Nutzung eines DWH-Systems dar. Das herausgearbeitete Praxisbild bescheinigte allerdings, dass zu einem DWH-System zumeist verschiedenartige Metadaten-Repositories existieren. Nennenswerte Metadaten-Standards beschränken sich im DWH-Kontext auf das CWM, dem jedoch einerseits eine geringe Akzeptanz und andererseits bezüglich der semantischen Auszeichnung von Metadaten ein schwaches Potential zuzuschreiben ist. Um insbesondere dem letztgenannten Defizit systematisch entgegenwirken zu können, wurde als Anhaltspunkt für die Entwicklung eines Vokabulars zur semantisch reichhaltigen Beschreibung von DWH-Metadaten die verwendungszweckgerichtete Metadatenkategorisierung eingeführt.

Gegenstand von *Kapitel 4* war die Einführung wichtiger Grundlagen zu Führungsinformationen in betrieblichen Organisationen. Neben den Beziehungen zwischen Daten, Information und Wissen wurde die Bedeutung der Kodifizierbarkeit von Wissen für betriebliche Organisationen herausgestellt. Die weiteren Erläuterungen umfassten Aspekte der Informationsversorgung von Entscheidungsträgern. Da Angaben über den Kontext, dem Daten entstammen, ein erforderliches Kriterium für deren semantisch eindeutige Beschreibung sind, wurde eine auf ein DWH bezogene Kontextdefinition mit zugehörigen Kontexttypen erarbeitet.

Kapitel 5 beschäftigte sich mit Grundlagen der Semiotik sowie den erforderlichen Merkmalen für eine vollständige Repräsentation eines Begriffs. Eine ungenaue oder nur fragmentarische Beschreibung eines Begriffs führt zur Entstehung sprachlicher Defekte. Zu deren Behandlung wurde die Methode zur Rekonstruktion von Fachbegriffen vorgeschlagen. Einen weiteren Untersuchungsaspekt dieses Kapitels bildete die (semantische) Heterogenität in Datenschemata. Dabei wurden Ursachen für das Auftreten von Heterogenität sowie Kriterien hinsichtlich der Überwindung spezifischer Heterogenitätsarten herausgearbeitet. Daneben konnte bei prominenten Forschungsansätzen, die sich der Überwindung von Heterogenität in Datenschemata widmen, ein überwiegender Fokus auf Fragen bezüglich struktureller Heterogenität festgestellt werden. Die Erkenntnisse aus diesem Kapitel resultierten in einer Anforderungsaufstellung zur semantisch reichhaltigen Beschreibung von DWH-Metadaten.

Kapitel 6 beinhaltete einen Exkurs zur Vorstellung und Diskussion ausgewählter Konzepte und Technologien des Semantic Web. Dabei wurde insbesondere die semantische Ausdrucksstärke von RDF und OWL untersucht und hinsichtlich der im vorausgegangenen Kapitel erarbeiteten Anforderungen an eine semantisch reichhaltige Beschreibung von DWH-Metadaten beleuchtet. Eine Erfüllung dieser Anforderungen konnte im Wesentlichen für die Ontologiesprache OWL festgestellt werden.

In *Kapitel 7* wurden entlang der verwendungszweckgerichteten Metadatenkategorisierung die Potentiale von RDF zur semantischen Auszeichnung von DWH-Metadaten ermittelt. Diese Ergebnisse stützten die Ausarbeitung eines RDFS für DWH-Metadaten (DWH-RDFS). Letzteres diente des Weiteren als Vorlage zur Ausgestaltung einer OWL-Ontologie zur semantisch reichhaltigen Beschreibung von DWH-Metadaten (DWH-OWL). Die Forschungsergebnisse DWH-RDFS und DWH-OWL sind über ihren URI frei verfügbar. Ein besonderer Mehrwert für die Modellierung eines DWH-Systems entsteht durch die in der DWH-OWL erfassten semantischen Auszeichnungen und Beziehungen der Schemaelemente. Jedes Modellierungsvorhaben geht somit von denselben Begrifflichkeiten und Restriktionen aus, wodurch eine hohe strukturelle und semantische Verwandtschaft in den abgeleiteten DWH-Schemata erreicht wird. Zur Wahrung der Nachhaltigkeit der modellierten DWH-Schemata wurden außerdem Konzepte zur Versionierung des DWH-RDFS und der DWH-OWL vorgeschlagen.

Zentraler Bestandteil von *Kapitel 8* war die Konzeption einer semantischen Homogenisierungsschicht für DWH-Systeme (sHGS-DWH). Für den Aufbau einer sHGS-DWH wurde eine Unterteilung in Top-Level-, Domain- und Applikationsontologien vorgeschlagen. Erstgenannte wird durch die im vorausgegangenen Kapitel definierte DWH-OWL repräsentiert. Domain-Ontologien stellen in diesem Anwendungsfall fachliche Referenzmodelle für konkrete Schemata von DWH-Systemen dar (Applikationsontologien). Im Hinblick auf die Anbindung multipler DWH-Systeme an eine sHGS-DWH erfolgte die Ausarbeitung einer entsprechenden Gesamt-Architektur. Anhand dieser konnte aufgezeigt werden, wie mithilfe einer sHGS-DWH die semantische Heterogenität multipler DWH-Systeme überwunden werden kann. Außerdem wurden der Nutzen und die erzielbare Unterstützung für Entwickler und Endanwender eines DWH-Systems in Bezug auf die identifizierten Defizite einer Infrastruktur mit multiplen DWH-Systemen herausgestellt. Die vorgegebenen Formalziele konnten als erfüllt bewertet werden.

Anhand repräsentativer Ausschnitte multipler DWH-Systeme aus dem Umfeld des Projekts CEUS^{HB} wurde in *Kapitel 9* nachgewiesen, dass deren semantische Heterogenität mithilfe einer sHGS-DWH überwunden werden kann. Die in Kapitel 1 herausgestellten Problemkategorien multipler DWH-Systeme dienten dabei als Orientierungspunkte der Argumentation. Des Weiteren konnte belegt werden, dass der Einsatz einer sHGS-DWH die Erstellung von DWH-Schemata mit hoher struktureller und semantischer Verwandtschaft fördert. Die dem Projekt CEUS^{HB} entlehnten DWH-Systeme traten in diesem Kapitel als Stellvertreter für multiple DWH-Systeme in verteilten Organisationen auf.

Mit der vorliegenden Arbeit wurde ein durchgängiger Ansatz zur Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen eingeführt. Neben der Kopplung von DWH-Systemen an eine sHGS-DWH stellt insbesondere die Bereitstellung eines Vokabulars zur semantisch reichhaltigen Beschreibung von DWH-Metadaten ein Herausstellungsmerkmal dieser Arbeit dar. Die mithilfe einer sHGS-DWH erzielbare, hohe strukturelle und semantische Verwandtschaft bei der Modellierung von DWH-Schemata war in der hier aufgezeigten Ausprägung bisher nicht möglich. Die geschaffene Möglichkeit, die semantische Heterogenität multipler DWH-Systeme zu überwinden, fördert die Ausgestaltung eines konsolidierten Berichtswesens und somit die Realisierung eines einheitlichen Lenkungssystems in betrieblichen Organisationen.

AUSBLICK

Der vorgestellte Lösungsansatz zur Realisierung einer semantisch reichhaltigen Beschreibung von DWH-Metadaten sowie der Vorschlag einer sHGS-DWH zur Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen eröffnen eine Reihe von Ansatzpunkten für weitere Forschungsaktivitäten:

- Zur Anbindung multipler DWH-Systeme an eine sHGS-DWH wurde auf Konzepte aus dem Bereich EAI zurückgegriffen (Kopplungskomponenten). Um eine vollständige Werkzeugunterstützung bei der Arbeit mit einer sHGS-DWH zu erzielen, bedarf es vor allem im Bereich der Überführung von DWH-Metadaten in eine Applikationsontologie der Entwicklung eines adäquaten Werkzeugs.

- Bei der Bereitstellung einer (Applikations-)Ontologie in einer sHGS-DWH ist außerdem die Beachtung von Zugriffsberechtigungen zu untersuchen. Eine Erweiterung des Konzepts einer sHGS-DWH sollte einem Nutzer – abhängig von seinem Zugriffsrecht – nur die für ihn freigegebenen Bereiche einer Ontologie zur Ansicht oder Bearbeitung anbieten.
- Ein weiterer Forschungsbereich, dessen Ergebnisse in einer sHGS-DWH noch einzubeziehen sind, ist das so genannte *Ontology-Matching*¹²¹. Ziel des *Ontology-Matchings* ist das Auffinden von ähnlichen oder gleichen Klassen zwischen zwei Ontologien. Zu den Anwendungsbereichen des *Ontology-Matchings* zählen der Vergleich von Ontologien oder die Informationsintegration. Einen zentralen Unterschied zum Forschungsgebiet des *Schema-Matchings* stellt die in den Ontologien erfasste und maschinenlesbare Semantik dar. Beim *Ontology-Matching* kann – analog zum *Schema-Matching* – zwischen elementbasierten und strukturbasierten Techniken unterschieden werden [Zaiß08, S. 13]. Der erzielbare Mehrwert durch den Einsatz derartiger *Matching-Systeme* in einer sHGS-DWH ist entsprechend zu ermitteln.
- Eine zusätzliche Unterstützung bei der Auflösung sprachlicher Defekte wie Synonyme oder Homonyme können lexikalische Datenbanken wie z. B. *WordNet*¹²² liefern. Zu analysieren sind deren Nutzung im Rahmen einer sHGS-DWH sowie deren Effekte im Hinblick auf die Überwindung der semantischen Heterogenität bei multiplen DWH-Systemen.
- Zuletzt soll die Einbindung der Ausprägungsebene bei der Überwindung semantischer Heterogenität multipler DWH-Systeme als ein Kriterium für die Weiterentwicklung einer sHGS-DWH genannt werden. Ein mögliches Untersuchungsziel könnte in der Zusammenführung der, aufgrund der semantischen Heterogenität auf Schemaebene, inhomogenen Daten liegen.

Die vorgestellten Ansätze zur Erweiterung des Vorschlags einer sHGS-DWH sollen Anregungen zur weiteren Forschung im Themengebiet der semantischen Heterogenität bei multiplen DWH-Systemen geben.

¹²¹ Einen fundierten Überblick über den Themenbereich des *Ontology-Matching* bietet z. B. [Zaiß08], detaillierte Erläuterungen sind u. a. in [EuSh07] gegeben.

¹²² Vgl. <http://wordnet.princeton.edu/> (Abruf am 10.06.2008).

Literaturverzeichnis

- [AaNy95] Aamodt, A.; Nygård, M.: Different roles and mutual dependencies of data, information, and knowledge – an AI perspective on their integration. In: *Data and Knowledge Engineering* 16 (1995) 3, S. 191-222.
- [Acko67] Ackoff, R. L.: Management Misinformation Systems. In: *Management Science* 14 (1967) 4, S. 147-156.
- [AdMo00] Adelman, S.; Moss, L. T.: *Data Warehouse Project Management*. Addison-Wesley, Boston 2000.
- [AnHa04a] Antoniou, G.; van Harmelen, F.: *Web Ontology Language: OWL*. In: Staab, S.; Studer, R.: *Handbook on Ontologies*. Springer, Berlin 2004, S. 67-92.
- [AnHa04b] Antoniou, G.; van Harmelen, F.: *A Semantic Web Primer*. MIT Press, Cambridge 2004.
- [AnLa04] Angele, J.; Lausen, G.: *Ontologies in F-Logic*. In: Staab, S.; Studer, R.: *Handbook on Ontologies*. Springer, Berlin 2004, S. 29-50.
- [Augu90] Augustin, S.: *Information als Wettbewerbsfaktor: Informationslogistik – Herausforderung an das Management*. Verlag TÜV Rheinland, Köln 1990.
- [AuMa02] Auth, G.; von Maur, E.: *A Software Architecture for XML-based Metadata Interchange in Data Warehouse Systems*. In: Chaudhri, A. K.; Unland, R.; Djeraba, C.; Lindner, W. (Hrsg.): *XML-Based Data Management and Multimedia Engineering – EDBT 2002 Workshops*. Springer, Berlin 2002, S. 1-14.
- [Auth04] Auth, G.: *Prozessorientierte Organisation des Metadatenmanagements für Data-Warehouse-Systeme*. Books on Demand, Norderstedt 2004.
- [Back91] Back-Hock, A.: *Executive Information Systeme (EIS)*. In: *krp Kostenrechnungspraxis – Zeitschrift für Controlling* (1991) 1, S. 48-50.
- [BaGü04] Bauer, A.; Günzel, H. (Hrsg.): *Data Warehouse Systeme – Architektur, Entwicklung, Anwendung*. 2. Auflage, dpunkt, Heidelberg 2004.
- [BaLN86] Batini, C.; Lenzerini, M.; Navathe, S. B.: *A comparative analysis of methodologies for database schema integration*. In: *ACM Computing Surveys (CSUR)* 18 (1986) 4, S. 323-364.
- [Bang02] Bange, C.: *Integration von Data und Document Warehouses – Verknüpfung strukturierter und unstrukturierter Daten in entscheidungsunterstützenden Informationssystemen durch den Knowledge Integrator*. Dissertation, Universität Würzburg 2002.
- [BDKK02] Becker, J.; Delfmann, P.; Knackstedt, R.; Kuroпка, D.: *Konfigurative Referenzmodellierung*. In: Becker, J.; Knackstedt, R. (Hrsg.): *Wissensmanagement mit Referenzmodellen – Konzepte für die Anwendungssystem- und Organisationsgestaltung*. Physica-Verlag, Heidelberg 2002, S. 25-144.

- [BeBe96] Becker, W.; Benz, K.: Effizienz des Controlling. Bamberger Betriebswirtschaftliche Beiträge Nr. 108, Bamberg 1996.
- [Beck04] Beckett, D.: RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004. <http://www.w3.org/TR/rdf-syntax-grammar> (Abruf am 03.06.2008).
- [BeFe01] Bertino, E.; Ferrari, E.: XML and Data Integration. In: IEEE Internet Computing 5 (2001) 6, S. 75-76.
- [BeHL01] Berners-Lee, T.; Hendler, J.; Lassila, O.: The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In: Scientific American (2001) 5. <http://www.sciam.com/article.cfm?chanID=sa006&coIID=1&articleID=00048144-10D2-1C70-84A9809EC588EF21> (Abruf am 03.06.2008).
- [BeKn04] Becker, J.; Knackstedt, R.: Referenzmodellierung im Data-Warehousing – State-of-the-Art und konfigurative Ansätze für die Fachkonzeption. In: Wirtschaftsinformatik 46 (2004) 1, S. 39-49.
- [BeMe07] Bernstein, P. A.; Melnik, S.: Model Management 2.0: Manipulating Richer Mappings. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. Beijing 2007, S. 1-12.
- [BeMu98] Behme, W.; Mucksch, H.: Die Notwendigkeit einer entscheidungsorientierten Informationsversorgung. In: Mucksch, H.; Behme, W.: Das Data Warehouse-Konzept: Architektur – Datenmodelle – Anwendungen, 3. Auflage, Gabler, Wiesbaden 1998, S. 3-31.
- [BeRa00] Bernstein, P. A.; Rahm, E.: Data Warehousing Scenarios for Model Management. In: Laender, A. H. F.; Liddle, S. W.; Storey, V. C. (Hrsg.): Proceedings of the 19th International Conference on Conceptual Modeling – ER 2000, Springer, Berlin 2000, S. 1-15.
- [Bern98a] Berners-Lee, T.: Semantic Web Road map. 1998. <http://www.w3.org/DesignIssues/Semantic.html> (Abruf am 03.06.2008).
- [Bern98b] Berners-Lee, T.: Why RDF model is different from the XML model. 1998. <http://www.w3.org/DesignIssues/RDF-XML> (Abruf am 03.06.2008).
- [BeSc06] Berger, S.; Schrefl, M.: Analysing Multi-dimensional Data Across Autonomous Data Warehouses. In: Tjoa, A. M.; Trujillo, J. (Hrsg.): Proceedings of the 8th International Conference on Data Warehousing Knowledge Discovery (DaWaK 2006). Springer, Berlin 2006, S. 120-133.
- [BeSc08] Berger, S.; Schrefl, M.: From Federated Databases to a Federated Data Warehouse System. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences 2008. <http://csdl2.computer.org/comp/proceedings/hicss/2008/3075/00/30750394.pdf> (Abruf am 03.06.2008).
- [BeSc93] Behme, W.; Schimmelpfennig, K.: Führungsinformationssysteme: Geschichtliche Entwicklung, Aufgaben und Leistungsmerkmale. In: Behme, W.; Schimmelpfennig, K. (Hrsg.): Führungsinformationssysteme, Gabler, Wiesbaden 1993, S. 3-16.

- [Bess04] Besser, M.: Data Warehouse Projekte: Sehr teuer und oft unnötig. 2004. http://www.sigs.de/presse/tdwi_Silicon_06_2004.pdf (Abruf am 03.06.2008).
- [BHH+04] Bechhofer, S.; van Harmelen, F.; Hendler, J.; Horrocks, I.; McGuinness, D. L.; Patel-Schneider, P. F.; Stein, L. A.: OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004. <http://www.w3.org/TR/owl-ref> (Abruf am 03.06.2008).
- [BHLT06] Bray, T.; Hollander, D.; Layman, A.; Tobin, R. (Hrsg.): Namespaces in XML 1.0 (Second Edition). W3C Recommendation 16 August 2006. <http://www.w3.org/TR/REC-xml-names> (Abruf am 03.06.2008).
- [Blas00] Blaschka, M.: FIESTA: A Framework for Schema Evolution in Multidimensional Databases. Dissertation, TU München 2000. http://deposit.d-nb.de/cgi-bin/dokserv?idn=962067342&dok_var=d1&dok_ext=pdf&filename=962067342.pdf (Abruf am 03.06.2008).
- [BISH99] Blaschka, M.; Sapia, C.; Höfling, G.: On Schema Evolution in Multidimensional Databases. In: Mohania, M.; Tjoa, A. M. (Hrsg.): Proceedings of the 1st International Conference on Data Warehousing and Knowledge Discovery (DaWak99). Springer, Berlin 1999, S. 153-164.
- [Böhn01] Böhnlein, M.: Konstruktion semantischer Data-Warehouse-Schemata. Deutscher Universitätsverlag, Wiesbaden 2001.
- [BöKU03] Böhnlein, M.; Knobloch, B.; Ulbrich-vom Ende, A.: Synergieeffekte zwischen Data Warehousing, OLAP und Data Mining – Eine Bestandsaufnahme. In: von Maur, E.; Winter, R. (Hrsg.): Data Warehouse Management – Das St. Galler Konzept zur ganzheitlichen Gestaltung der Informationslogistik. Springer, Heidelberg 2003, S. 167-193.
- [BöU100] Böhnlein, M.; Ulbrich-vom Ende, A.: Grundlagen des Data-Warehousing – Modellierung und Architektur. Bamberger Beiträge zur Wirtschaftsinformatik Nr. 55, Bamberg 2000.
- [BöU199] Böhnlein, M.; Ulbrich-vom Ende, A.: XML – Extensible Markup Language. In: Wirtschaftsinformatik 41 (1999) 3, S. 274-276.
- [Brat07] Bratt, S.: Semantic Web, and Other W3C Technologies to Watch. 2007. <http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/0130-sb-W3CTechSemWeb.pdf> (Abruf am 03.06.2008).
- [BrEG06] Braun, R.; Esswein, W.; Greiffenberg, S.: Einführung in die Programmierung – Grundlagen, Java, UML. Springer, Berlin 2006.
- [BrGu04] Brickley, D.; Guha, R. V.: RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. <http://www.w3.org/TR/rdf-schema> (Abruf am 03.06.2008).
- [Brid04] McBride, B.: The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS. In: Staab, S.; Studer, R.: Handbook on Ontologies. Springer, Berlin 2004, S. 51-65.

- [BuFK92] Bullinger, H.-J.; Friedrich, R.; Koll, P.: Management-Informationssysteme (MIS) – Vorgehensweisen, Trends und Entwicklungen. In: Office Management – Information / Organisation / Kommunikation 40 (1992) 11, S. 6-18.
- [Bühn04] Bühner, R.: Betriebswirtschaftliche Organisationslehre. 10. Auflage, Oldenbourg, München 2004.
- [BuKD06] Bucher, T.; Kurpjuweit, S.; Dinter, B.: Risikomanagement im Data Warehousing: Situative Komposition einer methodischen Vorgehensweise. In: Schelp, J.; Winter, R.; Frank, U.; Rieger, B.; Turowski, K. (Hrsg.): Integration, Informationslogistik und Architektur: DW2006. Köllen Druck+Verlag, Bonn 2006, S. 35-59.
- [BuMi06] Buxmann, P.; Miklitz, T.: IT-Standardisierung und -Integration bei M&A-Projekten – Ein Entscheidungsmodell zur Auswahl von Anwendungssystemen. In: Wirtz, B. W. (Hrsg.): Handbuch Mergers & Acquisitions Management. Gabler, Wiesbaden 2006, S. 1063-1083.
- [BVTS07] Banek, M.; Vrdoljak, B.; Tjoa, A. M.; Skočir, Z.: Automating the Schema Matching Process for Heterogeneous Data Warehouses. In: Song, I.; Eder, J.; Nguyen, T. M. (Hrsg.): Data Warehousing and Knowledge Discovery. 9th International Conference, DaWaK 2007 – Proceedings. Springer, Berlin 2007, S. 45-54.
- [Byrn97] Byrne, B.: Information Resource Dictionary System Standards and Support for Present and Future CASE Technology. In: Tolvanen, J.-P.; Winter, A.: CAiSE'97: 4th Doctoral Consortium on Advanced Information Systems Engineering – Proceedings. 1997.
- [CaNL06] Cao, L.; Ni, J.; Luo, D.: Ontological Engineering in Data Warehousing. In: Zhou, X. et al. (Hrsg.): Frontiers of WWW Research and Development – APWeb 2006. Springer, Berlin 2006, S. 923-929.
- [Carn47] Carnap, R.: Meaning and Necessity – A Study in Semantics and Modal Logic. University of Chicago, Chicago 1947.
- [CaZL06] Cao, L.; Zhang, C.; Lui, J.: Ontology-Based Integration of Business Intelligence. In: International Journal on Web Intelligence and Agent Systems 4 (2006) 3, S. 313-325.
- [ChDa97] Chaudhuri, S.; Dayal, U.: An Overview of Data Warehousing and OLAP Technology. In: ACM SIGMOD Record 26 (1997) 1, S. 65-74.
- [Chen76] Chen, P. P.-S.: The entity-relationship model – toward a unified view of data. In: ACM Transactions on Database Systems (TODS) 1 (1976) 1, S. 9-36.
- [ChGl04] Chamoni, P.; Gluchowski, P.: Integrationstrends bei Business-Intelligence-Systemen – Empirische Untersuchung auf Basis des Business Intelligence Maturity Model. In: Wirtschaftsinformatik 46 (2004) 2, S. 119-128.
- [ChGl06] Chamoni, P.; Gluchowski, P.: Analytische Informationssysteme – Einordnung und Überblick. In: Chamoni, P.; Gluchowski, P. (Hrsg.): Analytische Informationssysteme – Business Intelligence-Technologien und -Anwendungen. 3. Auflage, Springer, Berlin 2006, S. 3-22.

- [CHH+01] Connolly, D.; van Harmelen, F.; Horrocks, I.; McGuiness, D. L.; Patel-Schneider, P. F.; Stein, L. A.: DAML+OIL (March 2001) Reference Description. W3C Note 18 December 2001. <http://www.w3.org/TR/daml+oil-reference> (Abruf am 17.03.2008).
- [Conr97] Conrad, S.: Föderierte Datenbanksysteme – Konzepte der Datenintegration. Springer, Berlin 1997.
- [Czar07] Czarski, C.: Relationales Netz – RDF-Daten mit Oracle verarbeiten. In: iX – Magazin für professionelle Informationstechnik (2007) 3, S. 138-140.
- [DaOS03] Daconta, M. C.; Obrst, L. J.; Smith, K. T.: The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. Wiley, Indianapolis 2003.
- [DeAb99] Dey, A. K.; Abowd, G. D.: Towards a Better Understanding of Context and Context-Awareness. Technical Report GIT-GVU-99-22. Graphics, Visualization and Usability Center and College of Computing, Georgia Institute of Technology, Atlanta 1999.
- [DeMu88] Devlin, B. A.; Murphy, P. T.: An architecture for a business and information system. In: IBM Systems Journal 27 (1988) 1, S. 60-80.
- [Devl97] Devlin, B.: Data Warehouse: from Architecture to Implementation. Addison-Wesley, Reading, Massachusetts 1997.
- [Dint06] Dinter, B.: Metadatenmanagement – Ein Framework zur Positionierung. In: BI-Spektrum – Fachzeitschrift für Business Intelligence und Data Warehousing 1 (2006) 2, S. 29-31.
- [Ditt04] Dittmar, C.: Knowledge Warehouse – Ein integrativer Ansatz des Organisationsgedächtnisses und die computergestützte Umsetzung auf Basis des Data Warehouse-Konzepts. Deutscher Universitätsverlag, Wiesbaden 2004.
- [Ditt99] Dittmar, C.: Erfolgsfaktoren für Data Warehouse-Projekte – Eine empirische Studie aus Sicht der Anwendungsunternehmen. Arbeitsbericht Nr. 78, Institut für Unternehmensführung und Unternehmensforschung, Ruhr-Universität Bochum, Bochum 1999.
- [DoRa00] Do, H. H.; Rahm, E.: On Metadata Interoperability in Data Warehouses. Technischer Report 1-2000, Institut für Informatik, Universität Leipzig 2000. <http://lips.informatik.uni-leipzig.de/pub/2000-13> (Abruf am 03.06.2008).
- [Dude05] Dudenredaktion (Hrsg.): Duden – Das Fremdwörterbuch. 8. Auflage, Dudenverlag, Mannheim 2005.
- [EcEc04] Eckstein, R.; Eckstein, S.: XML und Datenmodellierung – XML-Schema und RDF zur Modellierung von Daten und Metadaten einsetzen. dpunkt, Heidelberg 2004.
- [Ecke04a] Eckerson, W.: Gauge Your Data Warehouse Maturity. In: DM Review Online, 11/2004. http://www.dmreview.com/article_sub.cfm?articleid=1012391 (Abruf am 03.06.2008).
- [Ecke04b] Eckerson, W.: In Search of a Single Version of Truth: Strategies for Consolidating Analytic Silos. TDWI Report Series, 2004. <http://www.tdwi.org/research/display.aspx?ID=7173> (Abruf am 03.06.2008).

- [ElNa07] Elmasri, R.; Navathe, S. B.: Fundamentals of Database Systems. 5. Auflage. Pearson Addison Wesley, Boston 2007.
- [EuSh07] Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg 2007.
- [Fers79] Ferstl, O. K.: Konstruktion und Analyse von Simulationsmodellen. Beiträge zur Datenverarbeitung und Unternehmensforschung, Bd. 22. Hain, Königstein/Ts. 1979.
- [FeSi06] Ferstl, O. K.; Sinz, E. J.: Grundlagen der Wirtschaftsinformatik, 5. Auflage, Oldenbourg, München 2006.
- [FHH+05] Filß, C.; Höhn, R.; Höppner, S.; Schumacher, M.; Wetzel, H.: Rahmen zur Auswahl von Vorgehensmodellen. In: Entscheidungsfall Vorgehensmodell, 12. Workshop der Fachgruppe WI-VM der Gesellschaft für Informatik e. V. (GI), Berlin 2005, S. 185-229.
- [FiHH03] Fikes, R.; Hayes, P.; Horrocks, I.: OWL-QL – A Language for Deductive Query Answering on the Semantic Web. In: Journal of Web Semantics: Science, Services and Agents on the World Wide Web 2 (2004) 1, S. 19-29.
- [Freg75] Frege, G.: Funktion, Begriff, Bedeutung – Fünf logische Studien. 4. Auflage, Vandenhoeck & Ruprecht, Göttingen 1975.
- [Fres05] Frese, E.: Grundlagen der Organisation – Entscheidungsorientiertes Konzept der Organisationsgestaltung. 9. Auflage, Gabler, Wiesbaden 2005.
- [FrWe00] Frie, T.; Wellmann, R.: Der Business Case im Kontext des Data Warehousing. In: Jung, R.; Winter, R. (Hrsg.): Data Warehousing Strategie – Erfahrungen, Methoden, Visionen. Springer, Berlin 2000.
- [Gabr99] Gabriel, R.: Strategische Bedeutung der analytischen Informationssysteme. In: Chamoni, P.; Gluchowski, P. (Hrsg.): Analytische Informationssysteme – Data Warehouse, On-Line Analytical Processing, Data Mining. 2. Auflage, Springer, Berlin 1999, S. 417-426.
- [Garr05] Garry, C.: What makes an Enterprise Data Warehouse? 2005. <http://www.eweek.com/c/a/Database/What-Makes-an-Enterprise-Data-Warehouse/> (Abruf am 03.06.2008).
- [Gass97] Gassmann, O.: Internationales F&E-Management – Potentiale und Gestaltungskonzepte transnationaler F&E-Projekte. Oldenbourg, München 1997.
- [GeFi92] Genesereth, M. R.; Fikes, R. E.: Knowledge Interchange Format, Version 3.0 Reference Manual. Report Logic-92-1. Computer Science Department, Stanford University 1992.
- [GFSH03] McGuinness, D.; Fikes, R.; Stein, L. A.; Hendler, J.: DAML-ONT: An Ontology Language for the Semantic Web. In: Fensel, D.; Hendler, J.; Lieberman, H.; Wahlster, W. (Hrsg.): Spinning the Semantic Web – Bringing the World Wide Web to Its Full Potential. MIT Press, Cambridge 2003, S. 65-93.

- [Glas97] Glasen, F.: Die Potentiale der Internet-Technologie zur Lösung von Integrationsproblemen. In: Reiterer, H.; Mann, T. (Hrsg.): Informationssysteme als Schlüssel zur Unternehmensführung – Anspruch und Wirklichkeit; Proceedings des 3. Konstanzer Informationswissenschaftlichen Kolloquiums (KIK '97). Universitätsverlag Konstanz (UVK), Konstanz 1997, S. 90-101.
- [GlCh06] Gluchowski, P.; Chamoni, P.: Entwicklungsleitlinien und Architekturkonzepte des On-Line Analytical Processing. In: Chamoni, P.; Gluchowski, P. (Hrsg.): Analytische Informationssysteme – Business Intelligence-Technologien und -Anwendungen. 3. Auflage, Springer, Berlin 2006, S. 143-176.
- [GlGC97] Gluchowski, P.; Gabriel, R.; Chamoni, P.: Management Support Systeme. Springer, Berlin 1997.
- [GlKe06] Gluchowski, P.; Kemper, H.-G.: Quo Vadis Business Intelligence? Aktuelle Konzepte und Entwicklungstrends. In: BI Spektrum – Fachzeitschrift für Business Intelligence und Data Warehousing 1 (2006) 1, S. 12-19.
- [GoFC04] Gómez-Pérez, A.; Fernández-López, M.; Corcho, O.: Ontological Engineering – with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer, London 2004.
- [GrLe02] Gruninger, M.; Lee, J.: Ontology – Applications and Design. In: Communications of the ACM 45 (2002) 2, S. 39-41.
- [Grub93] Gruber, T. R.: A Translation Approach to Portable Ontology Specifications. In: Knowledge Acquisition 5 (1993) 2, S. 199-220.
- [Grub95] Gruber, T. R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In: International Journal of Human-Computer Studies 43 (1995) 5-6, S. 907-928.
- [GrZa92] Greschner, J.; Zahn, E.: Strategischer Erfolgsfaktor Information. In: Krallmann, H.; Papke, J.; Rieger, B. (Hrsg.): Rechnergestützte Werkzeuge für das Management – Grundlagen, Methoden, Anwendungen. Erich Schmidt, Berlin 1992, S. 9-28.
- [Guar98] Guarino, N.: Formal Ontology and Information Systems. In: Guarino, N. (Hrsg.): Formal Ontology in Information Systems – Proceedings of the 1st International Conference on Formal Ontologies in Information Systems (FOIS'98). IOS Press, Amsterdam 1998, S. 3-15.
- [Hack98] Hackathorn, R. D.: Web Farming for the Data Warehouse – Exploiting Business Intelligence and Knowledge Management. Morgan Kaufmann, San Francisco 1998.
- [Hafn02] Hafner, M.: Datenschutz im Data Warehousing. Arbeitsberichte des CC DW2, Bericht-Nr.: BE HSG/CC DW2/04, Universität St. Gallen 2002. [http://web.iwi.unisg.ch/org/iwi/iwi_pub.nsf/wwwPublAuthorGer/FF93A9CC2CCEF719C12570A30051E533/\\$file/Datenschutz%20im%20Data%20Warehousing.pdf](http://web.iwi.unisg.ch/org/iwi/iwi_pub.nsf/wwwPublAuthorGer/FF93A9CC2CCEF719C12570A30051E533/$file/Datenschutz%20im%20Data%20Warehousing.pdf) (Abruf am 03.06.2008).

- [Hahn07] Hahne, M.: Meta data in Data Warehousing – The Common Warehouse Metamodel. In: BI-Spektrum – Fachzeitschrift für Business Intelligence und Data Warehousing 2 (2007) 1, S. 14-21.
- [HaKa06] Han, J.; Kamber, M.: Data Mining – Concepts and Techniques. 2. Auflage, Elsevier, Amsterdam 2006.
- [HaLe93] Habermann, H.-J.; Leymann, F.: Repository – Eine Einführung. In: Endres, A. et al. (Hrsg.): Handbuch der Informatik, Band 8.1. Oldenbourg, München 1993.
- [Hark04] Harka, Ö.: Mergers and Acquisitions (Theoretical and Practical Aspects). Officina Press, Szeged 2004.
- [HaTB01] Habermann, F.; Thomas, O.; Botta, C.: Organisational-Memory-System zur Unterstützung informationstechnisch basierter Verbesserungen von Geschäftsprozessen. Kurzbeitrag auf der Referenzmodellierung 2001. <http://www.wi.uni-muenster.de/is/Tagung/Ref2001/Kurzbeitrag10.pdf> (Abruf am 03.06.2008).
- [HaUl05] Hartmann, S.; Ulbrich-vom Ende, A.: Mehr Intelligenz an Bayerns Hochschulen. In: Köster, T. (Hrsg.): Staat&IT 2/2005, S. 18-20.
- [HaWe08] Hartmann, S.; Weber, M.: Semantisch homogene Beschreibung von Data-Warehouse-Metadaten mit RDF. In: Bichler, M.; Hess, T.; Kremer, H.; Lechner, U.; Matthes, F.; Picot, A.; Speitkamp, B.; Wolf, P. (Hrsg.): Multikonferenz Wirtschaftsinformatik 2008, GITO-Verlag, Berlin 2008, S. 117-128.
- [Helf02] Helfert, M.: Planung und Messung der Datenqualität in Data-Warehouse-Systemen. Dissertation Nr. 2648. Difo-Druck, Bamberg 2002.
- [Hell97] Hellmuth, T. W.: Terminologiemanagement – Aspekte einer effizienten Kommunikation in der computerunterstützten Informationsverarbeitung. Dissertation. 1997. http://deposit.ddb.de/cgi-bin/dokserv?idn=957056079&dok_var=d1&dok_ext=pdf&filename=957056079.pdf (Abruf am 10.06.2008).
- [HeMH01] Hernández, M. A.; Miller, R. J.; Haas, L. M.: Clio: A Semi-Automatic Tool For Schema Mapping. In: Proceedings of the 2001 ACM SIGMOD international conference on Management of Data. ACM, New York 2001, S. 607.
- [HeMo07] Henrich, A.; Morgenroth, K.: Kontextbasiertes Information Retrieval zur Unterstützung von Entwicklungsprozessen. In: Oberweis, A.; Weinhardt, C. et al. (Hrsg.): eOrganisation – Service-, Prozess-, Market-Engineering: 8. Internationale Tagung Wirtschaftsinformatik – Band 2. Universitätsverlag Karlsruhe, Karlsruhe 2007, S. 575-592.
- [Hess02] Hesse, W.: Ontologie(n). In: Informatik Spektrum 25 (2002) 6, S. 477-480.
- [HiRW06] Hippner, H.; Rentzmann, R.; Wilde, K. D.: Aufbau und Funktionalitäten von CRM-Systemen. In: Hippner, H.; Wilde, K. D.: Grundlagen des CRM – Konzepte und Gestaltung. 2. Auflage, Gabler, Wiesbaden 2006, S. 45-74.

- [HKRS08] Hitzler, P.; Krötzsch, M.; Rudolph, S.; Sure, Y.: Semantic Web – Grundlagen. Springer, Berlin 2008.
- [HoKB01] Holten, R.; Knackstedt, R.; Becker, J.: Betriebswirtschaftliche Herausforderungen durch Data-Warehouse-Technologien. In: Schütte, R; Rothhowe, T.; Holten, R.: Data Warehouse Management-Handbuch – Konzepte, Software, Erfahrungen. Springer, Berlin 2001, S. 41-64.
- [Holt03] Holten, R.: Integration von Informationssystemen. In: Wirtschaftsinformatik 45 (2003) 1, S. 41-52.
- [Holt97] Holten, R.: Die drei Dimensionen des Inhaltsaspektes von Führungsinformationssystemen. In: Becker, J.; Grob, H. L.; Müller-Funk, U.; Vossen, G. (Hrsg.): Arbeitsberichte des Instituts für Wirtschaftsinformatik, Arbeitsbericht Nr. 56. Münster 1997.
- [Holt99a] Holten, R.: Entwicklung von Führungsinformationssystemen – Ein methodenorientierter Ansatz. Deutscher Universitätsverlag, Wiesbaden 1999.
- [Holt99b] Holthuis, J.: Der Aufbau von Data Warehouse-Systemen – Konzeption – Datenmodellierung – Vorgehen. 2. Auflage, Deutscher Universitätsverlag, Wiesbaden 1999.
- [HoRS01] Holten, R.; Rothhowe, T.; Schütte, R.: Grundlagen, Einsatzbereiche, Modelle. In: Schütte, R; Rothhowe, T.; Holten, R.: Data Warehouse Management-Handbuch – Konzepte, Software, Erfahrungen. Springer, Berlin 2001, S. 3-24.
- [Horv06] Horváth, P.: Controlling. 10. Auflage, Vahlen, München 2006.
- [HoST99] Horrocks, I.; Sattler, U.; Tobies, S.: Practical Reasoning for Description Logics with Functional Restrictions, Inverse and Transitive Roles, and Role Hierarchies. In: Proceedings of the 1st workshop on Methods for Modalities, Amsterdam 1999.
- [Inmo05] Inmon, W. H.: Building the Data Warehouse, 4. Auflage, Wiley, Indianapolis 2005.
- [Inmo06] Inmon, W. H.: DW 2.0 – Architecture for the Next Generation of Data Warehousing. In: DM Review Magazine, 04/2006. <http://www.dmreview.com/issues/20060401/1051111-1.html> (Abruf am 03.06.2008).
- [Inmo93] Inmon, W. H.: Building the Data Warehouse, Wiley, New York 1993.
- [Isse07] Isselhorst, T.: Modellierung von Kontext für Führungsinformationssysteme. Dissertation. WiKu, Duisburg 2007.
- [Jans01] Jansen, S. A.: Mergers & Acquisitions – Unternehmensakquisitionen und -kooperationen. 4. Auflage, Gabler, Wiesbaden 2001.
- [Jeck03] Jeckle, M.: Erste praktische Anwendungen des Semantic Web. Vortrag auf der Web-Services/XML-One-Konferenz. München 2003. <http://www.jeckle.de/files/sw.pdf> (Abruf am 03.06.2008).

- [Jeck04] Jeckle, M.: OMG's XML Metadata Interchange Format XMI. In: Nüttgens, M.; Mendling, J. (Hrsg.): XML4BPM 2004, Proceedings of the 1st GI Workshop XML4BPM - XML Interchange Formats for Business Process Management at 7th GI Conference Modellierung 2004, Marburg 2004, S. 25-42.
- [JLVV03] Jarke, M.; Lenzerini, M.; Vassiliou, Y.; Vassiliadis, P.: Fundamentals of Data Warehouses. 2. Auflage, Springer, Berlin 2003.
- [Jone98] Jones, K.: An Introduction to Data Warehousing: What Are the Implications for the Network? In: International Journal of Network Management 8 (1998) 1, S. 42-56.
- [Jung03] Jung, R.: Datenintegration im Kontext unternehmerischer Strategien und Konzepte – Ein Ansatz zur Spezifikation von Anforderungen. In: Österle, H.; Winter, R. (Hrsg.): Business Engineering – Auf dem Weg zum Unternehmen des Informationszeitalters. 2. Auflage, Springer, Berlin 2003, S. 305-328.
- [Kell97] Keller, F.: Aufbau eines Center-Informationssystems auf Basis eines Data Warehouse. In: Reiterer, H.; Mann, T. (Hrsg.): Informationssysteme als Schlüssel zur Unternehmensführung – Anspruch und Wirklichkeit; Proceedings des 3. Konstanzer Informationswissenschaftlichen Kolloquiums (KIK '97). UVK Universitätsverlag Konstanz, Konstanz 1997, S. 189-203.
- [KeMU06] Kemper, H.-G.; Mehanna, W.; Unger, C.: Business Intelligence – Grundlagen und praktische Anwendungen – Eine Einführung in die IT-basierte Managementunterstützung. 2. Auflage, Vieweg, Wiesbaden 2006.
- [KHL+07] Katifori, A.; Halatsis, C.; Lepouras, G.; Vassilakis, C.; Giannopoulou, E.: Ontology Visualization Methods – A Survey. ACM Computing Surveys (CSUR) 39 (2007) 4, Article 10.
- [KilW95] Kifer, M.; Lausen, G.; Wu, J.: Logical Foundations of Object-Oriented and Frame-Based Languages. In Journal of the ACM 42 (1995) 4, S. 741-843.
- [KiWa07] Kieser, A.; Walgenbach, P.: Organisation. 5. Auflage, Schäffer-Poeschel, Stuttgart 2007.
- [KICa04] Klyne, G.; Carrol, J. J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation 10 February 2004. <http://www.w3.org/TR/rdf-concepts> (Abruf am 03.06.2008).
- [Knac06] Knackstedt, R.: Fachkonzeptionelle Referenzmodellierung einer Managementunterstützung mit quantitativen und qualitativen Daten – Methodische Konzepte zur Konstruktion und Anwendung. Dissertation. Logos-Verlag, Berlin 2006.
- [Krcm05] Krcmar, H.: Informationsmanagement. 4. Auflage, Springer, Berlin 2005.
- [Krst02] Kromer, G.; Stucky, W.: Die Integration von Informationsverarbeitungsressourcen im Rahmen von Mergers & Acquisitions. In: Wirtschaftsinformatik 44 (2002) 6, S. 523-533.

- [Kure06] Kruczynski, K.: DW 2.0 – das neue Konzept für Data Warehousing. In: isreport 10 (2006) 12, S. 22-25.
- [Kurz99] Kurz, A.: Data Warehousing – Enabling Technology. MITP-Verlag, Bonn 1999.
- [Lacy05] Lacy, L. W.: OWL: Representing Information Using the Web Ontology Language. Trafford, Victoria (BC) 2005.
- [LeE197] Lehmann, P.; Ellerau, P.: Definierte Fachbegriffe als kritischer Erfolgsfaktor eines Data Warehouses. In: Reiterer, H.; Mann, T. (Hrsg.): Informationssysteme als Schlüssel zur Unternehmensführung – Anspruch und Wirklichkeit; Proceedings des 3. Konstanzer Informationswissenschaftlichen Kolloquiums (KIK '97). Universitätsverlag Konstanz (UVK), Konstanz 1997, S. 169-188.
- [LeGu90] Lenat, D. B.; Guha, R. V.: Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, Reading (MA) 1990.
- [Lehm01] Lehmann, P.: Meta-Datenmanagement in Data-Warehouse-Systemen – Rekonstruierte Fachbegriffe als Grundlage einer konstruktiven, konzeptionellen Modellierung. Dissertation. Shaker, Aachen 2001.
- [Lehn03] Lehner, W.: Datenbanktechnologie für Data-Warehouse-Systeme – Konzepte und Methoden. dpunkt, Heidelberg 2003.
- [Lehn06] Lehner, F.: Wissensmanagement – Grundlagen, Methoden und technische Unterstützung. Hanser, München 2006.
- [LeJa99] Lehmann, P.; Jaszewski, J.: Business Terms as a Critical Success Factor for Data Warehousing. In: Gatzju, S.; Jeusfeld, M.; Staudt, M.; Vassiliou, Y. (Hrsg.): Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW '99), Heidelberg 1999, S. 7.1-7.5. <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-19/paper7.pdf> (Abruf am 03.06.2008).
- [LeNa07a] Leser, U.; Naumann, F.: Informationsintegration – Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. dpunkt, Heidelberg 2007.
- [LeNa07b] Legler, F.; Naumann, F.: A Classification of Schema Mappings and Analysis of Mapping Tools. In: Kemper, A.; Schöning, H.; Rose, T.; Jarke, M.; Seidl, T.; Quix, C.; Brochhaus, C. (Hrsg.): Datenbanksysteme in Business, Technologie und Web (BTW 2007). Proceedings der 12. Fachtagung des GI-Fachbereichs Datenbanken und Informationssysteme (DBIS). Köllen Druck + Verlag, Bonn 2007, S. 449-463.
- [LeWh58] Leavitt, H. J.; Whisler, T. L.: Management in the 1980's – New Information Flows Cut New Organization Channels. In: Harvard Business Review 36 (1958) 6, S. 41-48.
- [LiGa03] Li, M.; Gao, F.: Why Nonaka highlights tacit knowledge: a critical review. In: Journal of Knowledge Management 7 (2003) 4, S. 6-14.
- [MaBR01] Madhavan, J.; Bernstein, P. A.; Rahm, E.: Generic Schema Matching with Cupid. In: Proceedings of the 27th International Conference on Very Large Data Bases (VLDB 2001). Morgan Kaufmann, Orlando 2001, S. 49-58.

- [MaSc02] Mantel, S.; Schissler, M.: Application Integration. In: Wirtschaftsinformatik 44 (2002) 2, S. 171-174.
- [MaWe05] Matheis, T.; Werth, D.: Konzeption und Potenzial eines kollaborativen Data-Warehouse-Systems. In: Scheer, A.-W. (Hrsg.): Veröffentlichungen des Instituts für Wirtschaftsinformatik. Heft 185, Saarbrücken 2005. http://www.iwi.uni-sb.de/Download/iwihefte/Heft_185.pdf (Abruf am 03.06.2008).
- [MaWL07] Matheis, T.; Werth, D.; Loos, P.: Kollaboratives Data Warehousing – Konzeption und prototypische Realisierung flexibler Schema- und Datenintegration. In: Oberweis, A.; Weinhardt, C. et al. (Hrsg.): eOrganisation – Service-, Prozess-, Market-Engineering: 8. Internationale Tagung Wirtschaftsinformatik – Band 1. Universitätsverlag Karlsruhe, Karlsruhe 2007, S. 569-586.
- [Melc03] Melchert, F.: Das Common Warehouse Metamodel als Standard für Metadaten im Data Warehousing. In: von Maur, E.; Winter, R. (Hrsg.): Data Warehouse Management – Das St. Galler Konzept zur ganzheitlichen Gestaltung der Informationslogistik. Springer, Berlin 2003, S. 89-111.
- [MeRB03] Melnik, S.; Rahm, E.; Bernstein, P. A.: Rondo: A Programming Platform for Generic Model Management. In: Proceedings of the 2003 ACM SIGMOD international conference on Management of Data. ACM, New York 2003, S. 193-204.
- [Mert01] Mertens, P.: Integrierte Informationsverarbeitung. In: Mertens, P. (Haupt-Hrsg.): Lexikon der Wirtschaftsinformatik. 4. Auflage, Springer, Berlin 2001, S. 244-245.
- [MES+02] Mantel, S.; Eckert, S.; Schissler, M.; Ferstl, O. K.; Sinz, E. J.: Entwicklungsmethodik für überbetriebliche Kopplungsarchitekturen von Anwendungssystemen. Bayerischer Forschungsverbund Wirtschaftsinformatik, FORWIN-Bericht Nr. FWN-2002-009, Bamberg u. a. 2002.
- [Metz02] Metz, M.: Controlling des Integrationsprozesses bei Mergers & Acquisitions. Dissertation. Deutscher Universitätsverlag, Wiesbaden 2002.
- [Mint79] Mintzberg, H.: The Structuring of Organizations – A Synthesis of the Research. Prentice-Hall, Englewood Cliffs 1979.
- [Mint91] Mintzberg, H.: The Effective Organization: Forces and Forms. In: Sloan Management Review 32 (1991) 2, S. 54-67.
- [Mint92] Mintzberg, H. (Übersetzung: Höhle, H.): Die Mintzberg-Struktur – Organisationen effektiver gestalten. Verlag Moderne Industrie, Landsberg/Lech 1992.
- [Morg06] Morgenroth, K.: Kontextbasiertes Information Retrieval – Modell, Konzeption und Realisierung kontextbasierter Information Retrieval Systeme. Dissertation. Logos, Berlin 2006.
- [Morr79] Morris, C. W.: Grundlagen der Zeichentheorie – Ästhetik und Zeichentheorie. Ullstein, Frankfurt/Main 1979.

- [MuBe00] Mucksch, H.; Behme, W.: Das Data Warehouse-Konzept als Basis einer unternehmensweiten Informationslogistik. In: Mucksch, H.; Behme, W. (Hrsg.): Das Data Warehouse-Konzept – Architektur – Datenmodelle – Anwendungen. 4. Auflage, Gabler, Wiesbaden 2000, S. 3-79.
- [Neue81] Newell, A.: The Knowledge Level. In: AI Magazine 2 (1981) 2, S. 1-20, S. 33.
- [Nort05] North, K.: Wissensorientierte Unternehmensführung – Wertschöpfung durch Wissen. 4. Auflage, Gabler, Wiesbaden 2005.
- [NoTa95] Nonaka, I.; Takeuchi, H.: The Knowledge-Creating Company – How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York 1995.
- [Ortn00] Ortner, E.: Wissensmanagement – Teil 1: Rekonstruktion des Anwendungswissens. In: Informatik Spektrum 23 (2000) 2, S. 100-108.
- [Ortn97] Ortner, E.: Methodenneutraler Fachentwurf – Zu den Grundlagen einer anwendungsorientierten Informatik. B. G. Teubner, Stuttgart 1997.
- [Öste95] Österle, H.: Business Engineering – Prozeß- und Systementwicklung – Band 1: Entwurfstechniken. 2. Auflage, Springer, Berlin 1995.
- [PiFr88] Picot, A.; Franck, E.: Die Planung der Unternehmensressource Information (I). In: Das Wirtschaftsstudium 17 (1988) 10, S. 544-549.
- [PiRW03] Picot, A.; Reichwald, R.; Wigand, R. T.: Die grenzenlose Unternehmung – Information, Organisation und Management. 5. Auflage, Gabler, Wiesbaden 2003.
- [Pola58] Polanyi, M.: Personal Knowledge – Towards a Post-Critical Philosophy. Routledge and Kegan Paul Ltd., London 1958.
- [Pott98] Potthof, I.: Kosten und Nutzen der Informationsverarbeitung – Analyse und Beurteilung von Investitionsentscheidungen. Dissertation. Deutscher Universitätsverlag, Wiesbaden 1998.
- [Powe03] Powers, S.: Practical RDF – Solving Problems with the Resource Description Framework. O'Reilly, Beijing 2003.
- [QuKC05] Quix, C.; Kensch, D.; Chatti, M. A.: Rollenbasierte Metamodellierung zur Datenintegration. In: Datenbank-Spektrum 5 (2005) 15, S. 5-11.
- [RaBe01] Rahm, E.; Bernstein, P. A.: A survey of approaches to automatic schema matching. The VLDB Journal 10 (2001) 4, S. 334-350.
- [Rahm07] Rahm, E.: Model Management. In: Datenbank-Spektrum 7 (2007) 23, S. 60-61.
- [Rech06] Rechenberg, P.: Formale Sprachen und Automaten. In: Rechenberg, P.; Pomberger, G.: Informatikhandbuch. 4. Auflage, Hanser, München 2006, S. 89-110.
- [Reil83] O'Reilly, C. A.: The Use of Information in Organizational Decision Making: A Model and Some Propositions. In: Cummings, L. L.; Staw, B. M. (Hrsg.): Research in Organizational Behavior – An Annual Series of Analytical Essays and Critical Reviews. Vol. 5, Jai Press, Greenwich 1983, S. 103-140.

- [Riem05] Riemp, G.: Integriertes Wissensmanagement – Strategie, Prozesse und Systeme wirkungsvoll verbinden. In: Fröschle, H.-P.: Wissensmanagement. HMD Heft 246, dpunkt, Heidelberg 2005, S. 6-19.
- [RoTr82] Rockart, J. F.; Treacy, M. E.: The CEO goes on-line. In: Harvard Business Review 60 (1982) 1, S. 82-88.
- [RuNo04] Russel, S.; Norvig, P.: Künstliche Intelligenz – Ein moderner Ansatz. 2. Auflage, Pearson Studium, München 2004.
- [Russ06] Russom, P.: Master Data Management: Consensus-Driven Data Definitions for Cross-Application Consistency. TDWI Best Practices Report, October 2006. http://download.101com.com/pub/tdwi/Files/TDWI_MDM_Report_Q406REVISED.pdf (Abruf am 03.06.2008).
- [SaSH95] Sander, P.; Stucky, W. Herschel, R.: Automaten, Sprachen, Berechenbarkeit. In: Stucky, W. (Hrsg.): Grundkurs Angewandte Informatik IV. 2. Auflage, Teubner, Stuttgart 1995.
- [SBHD98] Sapia, C.; Blaschka, M.; Höfling, G.; Dinter, B.: Extending the E/R Model for the Multidimensional Paradigm. In: Kambayashi, Y.; Lee, D. L. et al. (Hrsg.): Proceedings of the Workshops on Data Warehousing and Data Mining: Advances in Database Technologies. LNCS Vol. 1552. Springer, Berlin 1998, S. 105-116.
- [SBPU01] Sinz, E. J.; Böhnlein, M.; Ulbrich-vom Ende, A.; Plaha, M.: Architekturkonzept eines verteilten Data Warehouse-Systems für das Hochschulwesen. Bamberger Beiträge zur Wirtschaftsinformatik Nr. 59, Bamberg 2001.
- [ScDi06a] Schmaltz, M.; Dinter, B.: Wartung von Dimensionsdaten in verteilten Data Warehouse-Systemen. In: Schelp, J.; Winter, R.; Frank, U.; Rieger, B.; Turowski, K. (Hrsg.): Integration, Informationslogistik und Architektur (DW2006). Köllen Druck+Verlag, Bonn 2006, S. 83-106.
- [ScDi06b] Schulze, K.-D.; Dittmar, C.: Business Intelligence Reifegradmodelle – Reifegradmodelle als methodische Grundlage für moderne Business Intelligence Architekturen. In: Chameni, P.; Gluchowski, P. (Hrsg.): Analytische Informationssysteme – Business Intelligence-Technologien und -Anwendungen. 3. Auflage, Springer, Berlin 2006, S. 71-87.
- [Schm98] Schmitt, I.: Schemaintegration für den Entwurf Föderierter Datenbanken. Infix, Sankt Augustin 1998.
- [Scho04] Scholl, W.: Innovation und Information – Wie in Unternehmen neues Wissen produziert wird. Hogrefe, Göttingen, 2004.
- [ScWe07] Schneider, U.; Werner, D. (Hrsg.): Taschenbuch der Informatik. 6. Auflage, Hanser, München 2007.
- [ScZe99] Schütte, R.; Zelewski, S.: Wissenschafts- und erkenntnistheoretische Probleme beim Umgang mit Ontologien. In: König, W.; Wendt, O. (Hrsg.): Wirtschaftsinformatik und Wissenschaftstheorie – Verteilte Theoriebildung. Frankfurt/Main 1999, S. 1-19.

- [ShKa93] Sheth, A.; Kashyap, V.: So Far (Schematically) yet So Near (Semantically). In: Hsiao, D. K.; Neuhold, E. J.; Sacks-Davis, R. (Hrsg.): Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5). North-Holland Publishing Co., Amsterdam 1993, S. 283-312.
- [Shos82] Shoshani, A.: Statistical Databases: Characteristics, Problems, and some Solutions. In: Proceedings of the 8th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco 1982, S. 208-222.
- [SiBU99] Sinz, E. J.; Böhnlein, M.; Ulbrich-vom Ende, A.: Konzeption eines Data Warehouse-Systems für Hochschulen. Bamberger Beiträge zur Wirtschaftsinformatik Nr. 52, Bamberg 1999.
- [Sinz02] Sinz, E. J.: Data Warehouse. In: Küpper, H.-U.; Wagenhofer, A. (Hrsg.): Handwörterbuch Unternehmensrechnung und Controlling. 4. Auflage, Schäffer-Poeschel, Stuttgart 2002, S. 309-318.
- [Sinz04] Sinz, E. J.: Unternehmensarchitekturen in der Praxis – Architekturdesign am Reißbrett vs. situationsbedingte Realisierung von Informationssystemen. In: Wirtschaftsinformatik 46 (2004) 4, S. 315-316.
- [Sinz88] Sinz, E. J.: Das Strukturierte Entity-Relationship-Modell (SER-Modell). In: Angewandte Informatik 30 (1988) 5, S. 191-202.
- [Sinz95] Sinz, E. J.: Das Informationssystem der Universität als Instrument zur zielgerichteten Lenkung von Universitätsprozessen. In: Wolff, K. D. (Hrsg.): Qualitätskonzepte einer Universität – Differenzierung, Effektivierung und Vernetzung. Erfurter Beiträge zur Hochschulforschung und Wissenschaftspolitik – Band 1. iudicium Verlag, München 1995, S. 65-83.
- [SiPa07] Sirin, E.; Parsia, B.: SPARQL-DL: SPARQL Query for OWL-DL. In: Golbreich, C.; Kalyanpur, A.; Parsia, B. (Hrsg.): Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions. 2007. <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-258/paper14.pdf> (Abruf am 10.06.2008).
- [SiPU02] Sinz, E. J.; Plaha, M.; Ulbrich-vom Ende, A.: Datenschutz und Datensicherheit in einem landesweiten Data-Warehouse-System für das Hochschulwesen. Bamberger Beiträge zur Wirtschaftsinformatik und Angewandten Informatik Nr. 62, Bamberg 2002.
- [Skod99] Skodvin, O.-J.: Mergers in Higher Education – Success or Failure? In: Tertiary Education and Management 5 (1999) 1, S. 63-78.
- [SMFS02] Schissler, M.; Mantel, S.; Ferstl, O. K.; Sinz, E. J.: Kopplungsarchitekturen zur überbetrieblichen Integration von Anwendungssystemen und ihre Realisierung mit SAP R/3. In: Wirtschaftsinformatik 44 (2002) 5, S. 459-468.
- [SmWe01] Smith, B.; Welty, C.: Ontology: Towards a New Synthesis. In: Welty, C.; Smith, B. (Hrsg.): Proceedings of the international conference on Formal Ontology in Information Systems (FOIS-2001). ACM Press, New York 2001, S. iii-x.

- [SmWG04] Smith, W. K.; Welty, C.; McGuinness, D. L.: OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004. <http://www.w3.org/TR/owl-guide> (Abruf am 10.06.2008).
- [SpPD92] Spaccapietra, S.; Parent, C.; Dupont, Y.: Model Independent Assertions for Integration of Heterogeneous Schemas. In: The VLDB Journal – The International Journal on Very Large Data Bases 1 (1992) 1, S. 81-126.
- [Staa02] Staab, S.: Wissensmanagement mit Ontologien und Metadaten. In: Informatik Spektrum 25 (2002) 3, S. 194-209.
- [Stae99] Staehle, W. H.: Management – Eine verhaltenswissenschaftliche Perspektive. 8. Auflage (überarbeitet von Conrad, P.; Sydow, J.), Vahlen, München 1999.
- [StHa05a] Stahlknecht, P.; Hasenkamp, U.: Einführung in die Wirtschaftsinformatik. 11. Auflage, Springer, Berlin 2005.
- [StHa05b] Stuckenschmidt, H.; van Harmelen, F.: Information Sharing on the Semantic Web. Springer, Heidelberg 2005.
- [Stra02] Strauch, B.: Entwicklung einer Methode für die Informationsbedarfsanalyse im Data Warehousing. Dissertation Nr. 2603. Difo-Druck, Bamberg 2002.
- [Stül91] von Stülpnagel, A.: Repositories – Konzepte, Architekturen, Standards. In: HMD Theorie und Praxis der Wirtschaftsinformatik 28 (1991) 161, S. 10-25.
- [StVV99a] Staudt, M.; Vaduva, A.; Vetterli, T.: Metadata Management and Data Warehousing. Technical Report 99.04, Department of Computer Science, University of Zurich. <ftp://ftp.ifi.unizh.ch/pub/techreports/TR-99/ifi-99.04.pdf.gz> (Abruf am 10.06.2008).
- [StVV99b] Staudt, M.; Vaduva, A.; Vetterli, T.: The Role of Metadata for Data Warehousing. Technical Report 99.06, Department of Computer Science, University of Zurich. <ftp://ftp.ifi.unizh.ch/pub/techreports/TR-99/ifi-99.06.ps.gz> (Abruf am 10.06.2008).
- [SUPH07a] Sinz, E. J.; Ulbrich-vom Ende, A.; Plaha, M.; Hartmann, S.: BI für das Hochschulwesen – E-Government, Teil 1. In: BI-Spektrum – Fachzeitschrift für Business Intelligence und Data Warehousing, TDWI-Sonderausgabe 2007, S. 9-12.
- [SUPH07b] Sinz, E. J.; Ulbrich-vom Ende, A.; Plaha, M.; Hartmann, S.: Nutzungs- und Betriebskonzept – E-Government, Teil 2. In: BI-Spektrum – Fachzeitschrift für Business Intelligence und Data Warehousing 2 (2007) 3, S. 22-26.
- [ThSM01] Thalhammer, T.; Schrefl, M.; Mohania, M.: Active Data Warehouses: Complementing OLAP with Active Rules. In: Data Knowledge Engineering 39 (2001) 3, S. 241-269.
- [Toto06] Totok, A.: Entwicklung einer Business-Intelligence-Strategie. In: Chamoni, P.; Gluchowski, P.: Analytische Informationssysteme – Business Intelligence-Technologien und -Anwendungen. 3. Auflage, Springer, Berlin 2006, S. 51-70.
- [VaVe01] Vaduva, A.; Vetterli, T.: Metadata Management for Data Warehousing: An Overview. In: International Journal of Cooperative Information Systems 10 (2001) 3, S. 273-298.

- [Voss08] Vossen, G.: Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme. 5. Auflage, Oldenbourg, München 2008.
- [Wals98] Walsh, N.: A Technical Introduction to XML. 1998. <http://xml.com/lpt/a/316> (Abruf am 10.08.2008).
- [WeOr80] Wedekind, H.; Ortner, E.: Systematisches Konstruieren von Datenbankanwendungen – Zur Methodologie der Angewandten Informatik. Carl Hanser, München 1980.
- [WiDV04] Wilmes, C.; Dietl, H. M.; van der Velden, R.: Die strategische Ressource „Data Warehouse“ – Eine ressourcentheoretisch empirische Analyse. Deutscher Universitätsverlag, Wiesbaden 2004.
- [Wile67] Wilensky, H. L.: Organizational Intelligence – Knowledge and Policy in Government and Industry. Basic Books, New York 1967.
- [Wint04] Winter, R.: Architektur braucht Management. In: Wirtschaftsinformatik 46 (2004) 4, S. 317-319.
- [Wirt00] Wirth, W.: Nadelöhr Wissen: Wie aus Daten Information und aus Information Wissen wird. Schriftliche Fassung eines Vortrags für die Münchner Medientage, München 2000. <http://www.medientage-muenchen.de/archiv/1999-2000/wirth.pdf> (Abruf am 10.06.2008).
- [Wüst91] Wüster, E.: Einführung in die allgemeine Terminologielehre und terminologische Lexikographie. 3.Auflage, Romanistischer Verlag, Bonn 1991.
- [Zach87] Zachman, J. A.: A framework for information systems architecture. In: IBM Systems Journal 26 (1987) 3, S. 276-292.
- [Zaiß08] Zaifß, K.: Ontologie-Matching: Überblick und Evaluation. In: Datenbank-Spektrum 8 (2008) 24, S. 12-17.
- [Zieg07] Ziegler, C.: Sinn oder nicht Sinn – Vom Suchen und Finden der Semantik im Web. In: c't – Magazin für Computertechnik (2007) 21, S. 172-179.



Der bedarfsorientierte Zugriff auf ein konsolidiertes Berichtswesen stellt für betriebliche Organisationen einen entscheidenden Erfolgsfaktor dar, um einen Informationsvorsprung gegenüber ihren Wettbewerbern erzielen zu können. Data-Warehouse-Systeme (DWH-Systeme) dienen dazu, Entscheidungsträgern die für ihre aktuelle Aufgabe relevanten Informationen rechtzeitig und in geeigneter Form zur Verfügung zu stellen. In der Praxis existieren jedoch multiple DWH-Systeme, die vor allem auf Ebene der Datenschemata sowie der zugrunde liegenden Terminologien eine starke Heterogenität aufweisen. Inkonsistenzen zwischen den generierten Berichten sowie eine divergente Interpretation der Berichtsergebnisse sind die Folge. Der Aufbau eines konsolidierten Berichtswesens ist daher nur mit nachträglichem, manuellem Abstimmungsaufwand möglich.

Die vorliegende Arbeit widmet sich insbesondere dem durch multiple DWH-Systeme offerierten Informationsangebot. Mithilfe der im Semantic Web etablierten Technologie Resource Description Framework (RDF) und der Web Ontology Language (OWL) wird ein Rahmen zur semantisch reichhaltigen Beschreibung von DWH-Metadaten geschaffen. Derartig ausgezeichnete Metadaten unterschiedlicher DWH-Systeme werden anschließend in einer semantischen Homogenisierungsschicht (sHGS) zusammengeführt. Mit diesem Hilfsmittel können Beziehungen zwischen multiplen DWH-Systemen identifiziert und deren semantische Heterogenität überwunden werden.

ISBN 978-3-923507-41-2

ISSN 1867-7401

16,50 €