Proaktives Management von Konsistenzbedingungen im Analytischen Performance Management

Stefan Brüggemann

OFFIS e.V.
Escherweg 2
D-26121 Oldenburg
brueggemann@offis.de

Abstract: Im Analytischen Performance Management (APM) definieren Fachanwender zur Überwachung von Zielsystemen benötigte Kennzahlen. Da hier die bereitzustellenden Daten definiert werden, können an dieser Stelle zusätzlich Datenqualitätsaspekte berücksichtigt werden. So lassen sich bereits proaktiv Konsistenzbedingungen festlegen, welche dann von Datenlieferanten berücksichtigt werden müssen. Dadurch wird die Entstehung von Inkonsistenzen vermieden. Die modellierten Konsistenzbedingungen werden ontologiebasiert verwaltet, damit diese selbst untereinander auf Widerspruchsfreiheit geprüft werden können. In diesem Beitrag wird das Vorgehensmodell des APM aufgegriffen und um explizite Modellierung von Datenqualitätsmerkmalen erweitert. Dazu wird mit ProCon ein Modell vorgestellt, welches es erlaubt, Konsistenzbedingungen in multidimensionalen Datenmodellen zu definieren. So wird ein wichtiger Beitrag zum unternehmensweiten Datenqualitätsmanagement geleistet, um präventiv Datenqualitätsmängel bereits bei der Entstehung zu vermeiden.

1 Einleitung

Organisationen setzen zunehmend auf organisationsweites Performance Management zur Kontrolle, Steuerung und Verbesserung der Organisationsleistung. Organisationen streben an, Strategie und Zielsystem zu operationalisieren, um die Organisationsleistung mittels Indikatoren zu überwachen. Mit dem Analytischen Performance Management (APM) wurde eine Methode zur zielgerichteten, kennzahlenbasierten Leistungsmessung in einer Organisation eingeführt.

Das APM stellt besondere Ansprüche an eine Informationslogistik. Informationslogistik befasst sich mit der Informationsbereitstellung mit Entscheidungsbezug. Ziel der Informationslogistik ist es, relevante Informationen in geeigneter Qualität zur Befriedigung der Informationsbedarfe bereitzustellen [WSDB08]. Getreu dem Prinzip des „Garbage in, Garbage out“ [BM99] können inkonsistente, fehlerhafte Daten zu Fehlinformationen und damit zu Fehlentscheidungen führen.

Unter dem Begriff des Datenqualitätsmanagements (DQM) werden die Modellierung, Verarbeitung, Speicherung und Darstellung von Daten mit dem Ziel der Sicherstellung einer hohen Datenqualität zusammengefasst [OWS+08]. DQM kann an verschiedenen Stellen ansetzen, z.B. als nachgelagerte Qualitätssicherung im ETL-Prozess beim Betrieb von Data Warehouse-Systemen. Präventives Datenqualitätsmanagement, beispielsweise durch die Kontrolle von Benutzereingaben, gewinnt immer stärker an Bedeutung. Gerade im aufkommenden Trend der Real Time Enterprises [29] ist eine frühe Datenqualitätskontrolle im Prozess der permanenten Datenintegration notwendig [MN07]. Heutzutage steht die Frage nach einem unternehmensweiten Datenqualitätsmanagement im Vordergrund, wo bereits bei der Entstehung der Daten Fehler vermieden werden können.

In diesem Beitrag wird mit ProCon (*Pro*aktives Management von *Kon*sistenzbedingungen) ein neues Konzept vorgestellt, welches es erlaubt, proaktiv Datenqualität im APM zu modellieren. Dieses Konzept ist ein Beitrag zum präventiven DQM, da es die Entstehung inkonsistenter Daten vermeidet. Bereits zum Zeitpunkt der Modellierung des Informationsbedarfs werden Konsistenzbedingungen modelliert. Hierzu wird eine Erweiterung des Vorgehensmodells des APM vorgeschlagen, so dass Fachexperten zur Modellierung der Datenqualität herangezogen werden können.

In Abschnitt 2 wird zunächst der Begriff „Analytisches Performance Management“ (APM) präzisiert, um das Problemumfeld zu konkretisieren. Zur Verdeutlichung wird ein Beispielszenario aus dem Gesundheitswesen gezeigt. Dann werden in Abschnitt 3 grundlegende Begriffe des Datenqualitätsmanagements erläutert und ausgewählte Datenqualitätsmerkmale erläutert. Die multidimensionale Analyse- und Managementsicht analytischer Informationssysteme wird in Abschnitt 4 eingeführt. Dort wird das multidimensionale Datenmodell MADEIRA erläutert, welches Grundlage für das im darauffolgenden Abschnitt eingeführte Metamodell ProCon ist. ProCon erlaubt es, Konsistenzbedingungen explizit in multidimensionalen Datenmodellen zu formulieren. Weiter wird eine Erweiterung des im APM definierten Vorgehensmodells um explizite Berücksichtigung von Konsistenzregeln vorgestellt. Zur prototypischen Umsetzung wird ein modellgetriebener Ansatz vorgestellt, der es erlaubt, aus diesem Metamodell eine domänenspezifische Sprache zu erzeugen. Für diese Sprache lässt sich dann ein Editor generieren, mit dem Konsistenzbedingungen erstellt werden können. Abschnitt 7 fasst den Beitrag zusammen und schließt mit einem Ausblick auf zukünftige Arbeiten ab.

2 Analytisches Performance Management

In [KM06] wurde der Begriff „Analytisches Performance Management“ eingeführt, vom weiter gefassten Begriff des „Performance Management“ abgegrenzt und wie folgt definiert: APM dient der Überwachung und Steuerung einer Organisation durch

- die kontinuierlich wiederkehrende Modellierung der Zielsysteme von Organisation und Organisationseinheiten,
- deren Verknüpfung untereinander und mit den zur Messung der Zielerreichung herangezogenen Indikatoren sowie

- der Überwachung von Indikatorausprägungen und Zielerreichungsgraden in einem analytischen Informationssystem.

Das APM umfasst also nur die Tätigkeiten, die für die Vorbereitung und Durchführung der zielgerichteten, kennzahlen-basierten Leistungsmessung in einer Organisation relevant sind. Ausgeklammert werden zwei weitere wichtige Merkmale des Performance Managements: Die enge Kopplung zwischen Organisationsstrategie und den für die Zielerreichung implementierten Geschäftsprozessen (Prozessorientierung) sowie die fachliche Ausgestaltung des Performance Managements bezüglich präferierter Strategien, Auswahl der Ziele und Indikatoren, betrachteter Geschäftsprozesse sowie verwendeter Managementinstrumente (Managementmethode).

2.1 Vorgehensmodell des APM

Das APM identifiziert ausgehend von der Organisationsstruktur strategische und operationale Ziele, die durch Kennzahlen operationalisiert werden. Dabei wird für vorab definierte Ziele einer Organisation der Zielerreichungsgrad mittels Kennzahlen gemessen und überwacht. Dabei werden folgende Phasen durchlaufen, die in Abbildung 1 dargestellt sind:

- Bei der Modellierung des Zielsystems formulieren Entscheider, ggf. mit Unterstützung durch Fachexperten, ein Zielsystem. Dieses enthält alle relevanten Ziele der Organisation einschließlich ihrer Ursache-Wirkungsbeziehungen. Alle zur Zielerfüllung geplanten Aktivitäten und Geschäftsprozesse werden hier definiert.
- Werden organisationale Rahmenbedingungen geändert, oder wird unzureichende Effektivität der Ziele oder Strategien festgestellt, wird durch das Hinterfragen der Ziele eine Adaption des Zielsystems ausgelöst.
- Im Rahmen der Modellierung des Messsystems identifizieren und spezifizieren Fachexperten, ggf. unterstützt durch Statistiker, die für die Überwachung der Zielerreichung notwendigen Indikatoren und Kennzahlen.
- Durch das Hinterfragen der Indikatoren und Kennzahlen wird die Anpassung der Indikatoren und Kennzahlen ausgelöst, um die Messung der Zielerreichung zu verbessern.
- Im Subprozess Performance Measurement werden im Rahmen der operativen Planung quantifizierbare Zielwerte für die Indikatoren festgelegt. Diese werden für die fortlaufende Überwachung der Zielerreichung genutzt. In der Analysephase werden die Ursachen für die Planabweichungen durch Statistiker und Datenanalysten untersucht.

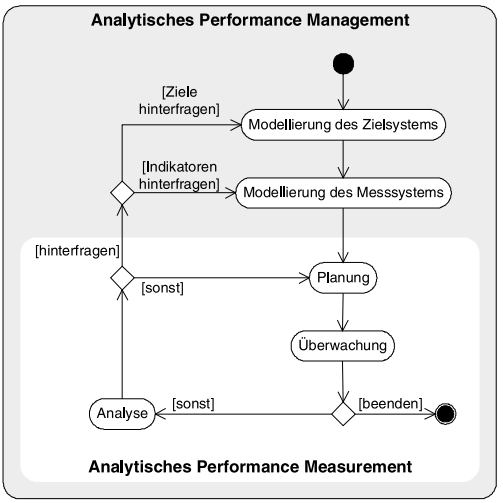

Abbildung 1: Vorgehensmodell des Analytischen Performance Management

Entsprechend der Theorie des organisationalen Lernens wird zwischen den zwei Lernprozessen Anpassungs- und Veränderungslernen unterschieden. Beim Anpassungslernen prüfen Fachexperten, ob sich Indikatoren für die Operationalisierung von Zielen als geeignet herausgestellt haben. Die verwendeten Indikatoren werden hinterfragt und ggf. angepasst, um die Messung der Zielerreichung zu verbessern. Eine mögliche Anpassung ist die Verbesserung der Risikoadjustierung. Bei Änderungen organisationaler Rahmenbedingungen oder unzureichender Effektivität der verfolgten Ziele oder Strategien hingegen wird durch das Hinterfragen der Ziele eine Adaption des Zielsystems ausgelöst. Entscheider und Fachexperten hinterfragen Strategie und Ziele und formulieren diese ggf. neu. Die Effektivität von Zielen im Hinblick auf die Erreichung übergeordneter Ziele und der Organisationsstrategie wird geprüft. Infolge der Neuformulierung werden auch Indikatoren und Kennzahlen durch Fachexperten angepasst.

Werden im Performance Measurement Abweichungen von Indikatoren von ihren Referenzwerten festgestellt, so können diese verschiedene Ursachen haben:

- Verfehlung der Zielerreichung: Entsprechend dem eigentlichen Zweck signalisiert ein Indikator, dass das Ziel, dem der Indikator zugeordnet ist, verfehlt wird.
- Unzureichende Referenzwerte: Der Referenzwert wurde für den Indikator zu restriktiv gewählt.
- Unzureichender Indikator: Der Indikator hat nicht alle relevanten Zusammenhänge repräsentiert und muss entsprechend adjustiert werden, z.B. durch geeignete Einbeziehung von Risikofaktoren in die Berechnungsvorschrift

– Unzureichende Datenqualität: Die Berechnung des Indikators beruht auf Daten unzureichender Qualität.

In diesem Beitrag soll die Ursache der mangelnden Datenqualität besonders betrachtet werden. Insbesondere wird ein Vorgehen eingeführt, wie im Prozessmodell des Analytischen Performance Management Datenqualität explizit berücksichtigt werden kann.

2.2 Beispielszenario: Qualitätsmessung im deutschen Gesundheitswesen

Im Gesundheitswesen wird im Rahmen der externen vergleichenden Qualitätssicherung die Leistung in den deutschen Krankenhäusern erfasst. Die medizinische und pflegerische Qualität wird mittels geeigneter Indikatoren sichtbar gemacht. Dazu dokumentieren alle Krankenhäuser qualitätsrelevante Daten für bestimmte Leistungsbereiche. Diese Daten werden gemäß der Vorgaben des Gesetzgebers (§ 137 Sozialgesetzbuch V) von der BQS Bundesgeschäftstelle Qualitätssicherung gGmbH [BQS08] analysiert, um beispielsweise die Qualität der Krankenhausversorgung sichtbar und vergleichbar zu machen. Die BQS, gegründet im Jahr 2000 von den Selbstverwaltungspartnern im Gesundheitswesen, leitet und koordiniert die inhaltliche Entwicklung und organisatorische Umsetzung der externen vergleichenden Qualitätssicherung.

Die BQS definiert in Fachgruppen Qualitätsziele und ermittelt für diese Qualitätsindikatoren und Auffälligkeitsgrenzen, mit denen die gesetzten Berechnungsergebnisse bewertet und die Zielerreichung beurteilt werden können. Gemäß dem Vorgehen des Analytischen Performance Managements werden die definierten Qualitätsindikatoren zyklisch hinterfragt und die Auswertungskonzepte fortlaufend überprüft und weiterentwickelt. Um die ca. 2200 deutschen Krankenhäuser miteinander vergleichen zu können, sind folgende Schritte nötig:

– Alle Krankenhäuser erheben qualitätsrelevante Daten und senden sie der BQS
– Die zuvor definierten Kennzahlen werden von der BQS gerechnet und analysiert
– Die Ergebnisse werden den Krankenhäusern berichtet und gegebenenfalls werden Empfehlungen ausgesprochen
– Sollten Auffälligkeitsgrenzen überschritten sein, werden diese Ergebnisse mit den Krankenhäusern in einem „Strukturierten Dialog“ gemeinsam analysiert und Maßnahmen zur Verbesserung umgesetzt.

Dem Aspekt der Datenqualität kommt im Kontext externer vergleichender Qualitätssicherung eine besondere Bedeutung zu. Basieren die durchgeführten Berechnungen und damit die Vergleiche der gemessenen Krankenhäuser auf fehlerhaften Daten, so sind auch die Vergleiche selbst fehlerhaft. Daher ist es einerseits wichtig, bereits bei der Modellierung der Indikatoren und Kennzahlen durch Fachexperten diese auch schon die Modellierung von (In-)Konsistenzen durchführen zu lassen. Andererseits ist es wichtig, nicht nur die Indikatoren und Kennzahlen, sondern auch die Konsistenzbedingungen kontinuierlich zu adjustieren. Dadurch, dass bereits bei der Definition des Informations-

bedarfs die Datenqualität explizit berücksichtigt und modelliert wird, lässt sich die Qualität der an die BQS gelieferten Daten proaktiv steigern.

Da die Krankenhäuser zyklisch neue Anforderungen der BQS erhalten und umsetzen, sind sie als Datenquelle bereits in der Lage, nur qualitativ hochwertige, den Konsistenzbedingungen genügende Daten zu liefern. Gleichzeitig ist die BQS in der Lage, unzureichende Daten von vornherein abzulehnen.

Beispiele für medizinische Dimensionen, die der BQS von Krankenhäusern zur Dokumentation in der Kardiologie geliefert werden, sind ICD-Kodierungen und OPS-Schlüssel. Die Weltgesundheitsorganisation WHO [ICD] definiert mit der International Classification of Diseases (ICD) eine Klassifikation von Erkrankungen. OPS-Schlüssel [Org78] dienen der formalen Bezeichnung von Operationen und Prozeduren im klinischen Kontext.

3 Datenqualität

Private und öffentliche Organisationen beginnen zunehmend, die Qualität der Daten, die sie verarbeiten, als ein wichtiges Wirtschaftsgut zu verstehen. Daher ist eine hohe Datenqualität anzustreben. Es existiert eine Vielzahl von Studien, die verschiedenste Prozentzahlen für fehlerhafte Daten in Organisationsdatenbanken nennen. Viele (Data Warehouse-) Projekte sind gescheitert, und oftmals sind es Datenqualitätsmängel, die zu fehlerhaften Entscheidungen führen. Data Warehouses liegen oftmals als Datenbasis im APM zu Grunde. Es ist von essentieller Wichtigkeit, dass diese Datenbasen genaue, vollständige und konsistente Daten enthalten.

Im Folgenden soll in Abschnitt 3.1 zunächst aufgezeigt werden, wie der Begriff der Datenqualität definiert werden kann und durch welche Merkmale Datenqualitätsmängel beschrieben werden können. In Abschnitt 3.2 werden Konsistenzprobleme explizit dargestellt. Abschnitt 3.3 zeigt, wo Datenfehler entstehen können und Abschnitt 3.4 schließlich zeigt, wie Konsistenzprobleme behoben werden können.

3.1 Datenqualitätsmängel und -merkmale

Die gegenwärtige Literatur verwendet keine einheitliche Definition des Begriffs Datenqualität. Vielmehr wird Datenqualität durch verschiedene Datenqualitätsmerkmale beschrieben. Im Folgenden wird ein Überblick über die gebräuchlichsten Merkmale gegeben. Im Beitrag von Scannapieco et al. [SMB05] werden die Merkmale Konsistenz, Genauigkeit und Vollständigkeit untersucht: Unter dem Begriff der Vollständigkeit werden Nullwertbehandlungen betrachtet, Genauigkeit betrachtet den syntaktischen und semantischen Abstand eines falschen Wertes zu einem richtigen. Konsistenz behandelt die Einhaltung semantischer Regeln, beispielsweise Integritätsbedingungen.

Naumann [Nau02] definiert Genauigkeit einer Datensammlung als den Quotienten der Anzahl korrekter Werte und der Anzahl aller Werte. Müller und Freytag [MF03] definie-

ren Konsistenz Schemakonformität. Sie wird gemessen als Anzahl fehlerfreier Tupel einer Relation in Bezug zur Anzahl aller Tupel einer Relation. Fehlerfreiheit bezieht sich hier auf die Verletzung der syntaktischen Struktur einer Relation. Müller und Freytag definieren die Korrektur domänenspezifischer Datenfehler als größte offene Herausforderung bei der Datenbereinigung. Dies ist darin begründet, dass Wissen bislang nicht ausreichend modelliert wird, um Korrekturwerte für invalide Tupel zu identifizieren. Da sich Ontologien zur Wissensmodellierung eignen wird in dieser Arbeit ein ontologiebasierter Ansatz eingeführt.

Ontologien sind formale explizite Spezifikationen einer gemeinsamen Konzeptualisierung [Gru93]. Sie repräsentieren Konzepte und ihre Beziehungen untereinander. Ontologien enthalten weiter Inferenz- und Integritätsregeln, mit denen Plausibilitätszusammenhänge geprüft werden können.

3.2 Betrachtung von Konsistenzproblemen

Konsistenz bezeichnet die korrekte Kombination der Attributwerte, durch die ein Tupel beschrieben werden kann. Attributwerte können sowohl in linearen Wertebereichen (bspw. Alter) definiert sein, als auch hierarchischen Metadatenkatalogen entstammen. Es wird bei der Prüfung auf Konsistenz hier die Einhaltung semantischer Regeln betrachtet, durch die eine multidimensionale Entität konsistent beschrieben werden kann. In multidimensionalen Datenmodellen können semantische Regeln an Dimensionsmetadaten geknüpft werden. Der Begriff der Konsistenz bezeichnet allgemein die Verletzung semantischer Regeln, die für Daten und deren Attribute definiert sein können. Im relationalen Bezug können dies beispielsweise Integritätsbedingungen sein. Diese Integritätsbedingungen müssen von allen Tupeln einer Datenbank erfüllt sein. Semantische Regeln werden üblicherweise in zwei Kategorien eingeteilt:

- Intra-relationale Konsistenzbedingungen: Diese Bedingungen betreffen einzelne oder mehrere Attribute einer Relation. Diese können sowohl auf Schemaebene als auch auf Instanzebene auftreten. Auf Schemaebene werden Bedingungen definiert, die für alle Tupel dieser Relation gelten, beispielsweise „Sterbedatum jünger als Geburtsdatum“. Auf Instanzebene werden Bedingungen zwischen konkreten Werte-Ausprägungen definiert.
- Inter-relationale Konsistenzbedingungen: Diese Bedingungen definieren Regeln, die Attribute betreffen, die aus verschiedenen Relationen stammen können.

3.3 Auftreten von Datenqualitätsmängeln

In der Vergangenheit wurden vielfältige Maßnahmen entwickelt, die aufgetretene Datenfehler aufspüren und bereinigen können. Datenbereinigung wurde als eines der größten Probleme im Data Warehousing betrachtet [RD00]. Datenprobleme werden klassifiziert nach dem Ort ihres Auftretens:

- Single Source-Probleme auf Schemaebene: Hier wird von einer Datenbasis ausgegangen, in der es unzureichende Integritätsbedingungen und mangelhaftes Schemadesign geben kann. Besondere Probleme sind die Nichteindeutigkeit von Schlüsseln, mangelhafte referentielle Integrität oder Werte außerhalb des zulässigen Wertebereichs.
- Single Source-Probleme auf Instanzebene: Auf Tupelebene innerhalb einer Datenbasis kann es Probleme bezüglich konkreter Ausprägungen der Attributwerte geben. Diese können ungültige, fehlende, widersprüchliche oder deplatzierte Werte sein, aber auch ungültige Attributwertkombinationen. Solche Probleme lassen sich in der Regel nur mit Hilfe domänenspezifischen Wissens identifizieren und beheben.
- Multi Source-Probleme auf Schemaebene: Bei mehreren, sich inhaltlich überlappenden Problemen kann es auf Schemaebene zu Namenskonflikten und strukturellen Konflikten kommen. Ursächlich hierfür sind heterogene Datenmodelle und Schemadesigns.
- Multi Source-Probleme auf Instanzebene: Auch auf Instanzebene kann es bei der Betrachtung mehrerer Datenquellen zu Problemen kommen. Besonders auffällig sind hier Redundanzen und Inkonsistenzen. Benutzen verschiedene Krankenhäuser beispielsweise unterschiedliche Klassifikationssysteme zur Befundung, lassen sich diese Daten nicht oder nur eingeschränkt vergleichen.

In diesem Beitrag werden Single Source-Konsistenzprobleme auf Instanzebene betrachtet.

3.4 Edit/Imputation – Systeme

Die Erkennung und Behebung von Konsistenz-Verletzungen ist ein wohlbekanntes Problem. Im statistischen Bereich, beispielsweise bei der Verarbeitung von Fragebögen, werden Regeln (edits) definiert. In einem Fragebogen kann zum Beispiel die ungültige Aussage ICD=I21.1 und OPS=5.728.0 erfasst worden sein. Eine Regel, um einen solchen Fehler zu erkennen, könnte sein: Wenn ICD=I21.1, dann muss OPS <=5.72 sein. Nachdem solch ein Fehler erkannt worden ist, muss dieser fehlerhafte Datensatz korrigiert werden. Dazu werden Daten in fehlerhaften Feldern mit korrekten Werten ersetzt (imputation). Im amerikanischen Census Bureau[1] werden demographische Daten mit Hilfe solcher Regeln qualitätsgesichert. Dieses Vorgehen der Fehlererkennung und Ersetzung fehlerhafter Werte wird in der Literatur als Edit-Imputation-Problem bezeichnet.

Die Fellegi-Holt-Methode [FH76] ist ein bekanntes theoretisches Modell [LKM06, SMB05] dieses Ansatzes. Es hat drei wesentliche Ziele:

[1] Das Census Bureau befasst sich mit der Erhebung von Bevölkerungsdaten (siehe www.census.gov, zuletzt besucht am 12. März 2008).

- Die Daten jedes Datensatzes sollen alle Edits erfüllen, indem nur die wenigsten Felder geändert werden. So bleibt ein Großteil der Originaldaten unverändert.
- Ersetzungsregeln sollen automatisch aus den Edits abgeleitet werden. So wird gewährleistet, dass ersetzte Tupel die Edits erfüllen. Weiter wird so die Wartbarkeit verbessert.
- Ersetzte Werte sollen in Hinblick auf die Häufigkeitsverteilung des betrachteten Attributs gewählt werden. So soll nicht nur der Durchschnittswert des Attributes zur Ersetzung verwendet werden, um die Verteilung nicht zu verfälschen.

Winkler et al. [HSW07, Win04] geben eine Einführung in das Themengebiet. Es existieren diverse Implementierungen dieses Modells [WQ03, WP97, MR01], dennoch besteht massiver Forschungsbedarf insbesondere beim Problem der Fehlerlokalisierung.

Es gibt leider einige Einschränkungen im Bezug auf Edit/Imputation Systeme, so dass diese leider nicht bei der Definition des Informationsbedarfes in einer Data Warehouse Umgebung zum Einsatz kommen können:

- Bisherige Implementierungen sind nicht auf Data Warehouses angewandt, so dass es hier insbesondere keine Anpassung des Modells an multidimensionale Datenstrukturen gibt. Durch die Ausnutzung hierarchischer Strukturen lassen sich invalide Tupel besser korrigieren, da nicht nur statistisch ermittelte Werte, sondern semantisch passende gewählt werden können.
- Edit/Imputation-Systeme sind nicht dazu konzipiert, proaktiv in den Prozess der Informationsbeschaffung und die Definition des Informationsbedarfes einzugreifen. Ihre Anwendung findet sich beispielsweise in ETL-Prozessen, wo sie reaktiv Inkonsistenzen in Daten finden und beheben können.
- Regeln (edits) in Edit/Imputation Systemen sind weniger mächtig als Integritätsbedingungen in Informationssystemen, da sie nicht auf einem Datenmodell wie dem Relationalen basieren [SMB05].
- Bei der Definition von Konsistenzbedingungen können Fehler und Widersprüche auftreten, so dass zwei Regeln sich gegenseitig ausschließen. Edit/Imputation Systeme beachten diese Widersprüche bisher nicht.

Um Widersprüche in Konsistenzbedingungen aufzufinden, wird in dieser Arbeit in Abschnitt 5 ein ontologiebasierter Ansatz eingeführt. Ontologien wurden bereits in [LN07] zum DQM bei der Informationsintegration und in [BA07] zur ontologiebasierten Datenvalidierung und -bereinigung eingeführt.

4 Multidimensionale Analyse- und Managementsicht

Analytische Informationssysteme bieten für die explorative Analyse von Daten typischerweise eine multidimensionale, auswertungsorientierte Sicht auf die in einem Data Warehouse integrierten Daten [CG06]. Dabei wird von einzelnen Individuen, Fällen oder Transaktionen (Mikrodaten) abstrahiert. Stattdessen werden aggregierte, beispielsweise

nach Zeit oder Organisationseinheit klassifizierte Daten (Makrodaten) in Form multidimensionaler Datenräume (Datenwürfel, Cubes) analysiert. In Forschung und Praxis gibt es diverse multidimensionale Datenmodelle, die für die explorative Analyse von Daten konzipiert sind. Multidimensionale Datenmodelle sind aus folgenden Gründen als Analyse- und Managementsicht für die im APM zur Modellierung und Überwachung von Messsystemen benötigten Informationen geeignet:

- Die Konzepte und Begriffe multidimensionaler Datenmodelle spiegeln die Sichtweise von Analysten und Entscheidern in Dimensionen und Klassifikationshierarchien wider.
- Die zur Messung der Zielerreichung herangezogenen Kennzahlen bzw. Indikatoren haben einen verdichtenden Charakter.

Das multidimensionale Datenmodell MADEIRA (Modelling Analyses of Data in Epidemiological InteRActive studies) [Wie00] integriert die Konzepte und Begriffe verschiedener Modelle aus den Bereichen On-Line Analytical Processing (OLAP) sowie Scientific & Statistical Databases. Dabei wird ein konzeptioneller Ansatz verfolgt, der unabhängig von der Implementierung in einer konkreten Datenbank (beispielsweise Star Schema oder Snowflake Schema [Leh03]) ist. Der multidimensionale Datenraum bildet die zentrale Struktur für Daten und Berechnungsergebnisse. Aufgrund des verdichtenden Charakters der zur Messung der Zielerreichung herangezogenen Kennzahlen eignet sich die multidimensionale Analyse- und Managementsicht für die Modellierung von Messsystemen und deren Überwachung im Rahmen des APM.

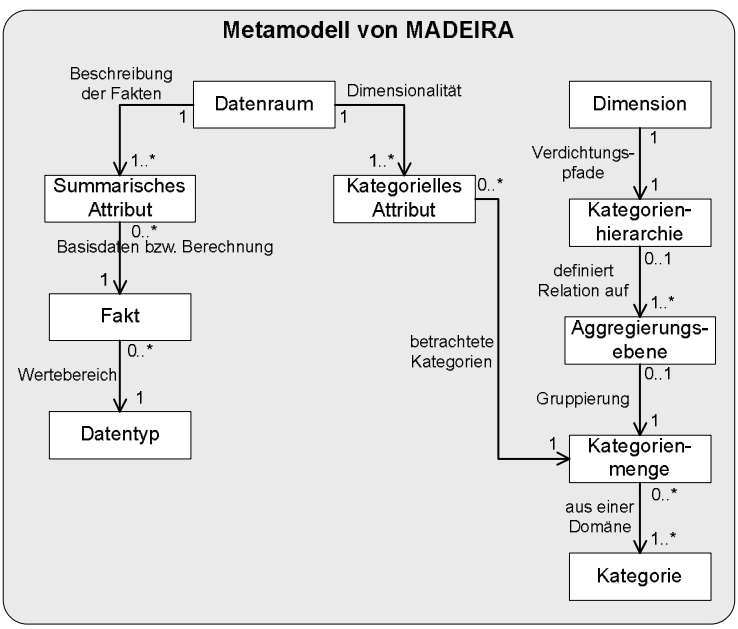

Abbildung 2: Metamodell von MADEIRA

Das Metamodell von MADEIRA ist in Abbildung 2 dargestellt. In MADEIRA wird ein multidimensionaler Datenraum durch summarische und kategorielle Attribute beschrieben. Einzelne Fakten und deren Berechnungen werden durch summarische Attribute definiert. Kategorielle Attribute werden durch mehrere Kategorien gebildet. Kategorien sind elementare Ausprägungen. Nach diesen kann klassifiziert werden (z.B. ICD-Code = 'I21') und sie spannen so den multidimensionalen Datenraum auf. Zugeordnet sind Kategorien zur Aggregierungsebene der Kategorienhierarchie einer Dimension. Dimensionen können beispielsweise 'ICD' oder 'Organisationseinheit' sein. Dimensionen sind unabhängig von betrachteten multidimensionalen Datenräumen und werden organisationsspezifisch modelliert.

5 ProCon: Proaktives Management von Konsistenzbedingungen im APM

Konsistenzbedingungen lassen sich im Analytischen Performance Management direkt an die betrachteten Kennzahlen binden. Da Fachanwender Kennzahlen auf Basis der multidimensionalen Sicht auf das Zielsystem definieren, ist es empfehlenswert, Konsistenzbedingungen direkt in multidimensionalen Datenmodellen zu modellieren.

Rückblickend auf die im Abschnitt 2 beschriebene Phase der Modellierung des Messsystems wird diese hier noch einmal vertieft dargestellt. In dieser Phase werden insbesondere die folgenden Schritte durchgeführt:

- *Datengrundlage prüfen:* Für die ausgewählten Indikatoren wird überprüft, ob die benötigten Daten und Metadaten aus den operativen Applikationen oder aus externen Quellen mit vertretbarem Aufwand beschafft werden können. Es werden also Informationsbedarf und Informationsangebot abgeglichen. Ggf. wird auf Indikatoren verzichtet.
- *Operationalisierung festlegen:* Die Operationalisierung von Zielen durch Indikatoren wird abschließend festgelegt.
- *Berechnungsvorschriften konkretisieren:* Die Berechnung der Kennzahlen, auf denen die Indikatoren basieren, wird exakt in Form von Berechnungsvorschriften beschrieben. Für alle berechneten Kennzahlen wird die komplette Herleitung aus Basiskennzahlen beschrieben.

Werden diese Schritte durchgeführt, so sollte hier auch das Datenqualitätsmanagement berücksichtigt werden. Der Fachexperte kann an dieser Stelle sinnvolle Einschränkungen für die betrachteten Daten und Dimensionen modellieren. Zwischen Kennzahlen, Aggregierungsebenen und Dimensionen lassen sich semantische Regeln definieren, die im Vorgehensmodell des APM berücksichtigt werden müssen. Daher wird folgende Teilaktivität als Erweiterung der Aktivität der Modellierung des Messsystems definiert:

- *Konsistenzbedingungen modellieren:* Zur Optimierung der Datenqualität lassen sich für Kennzahlen, auf denen Indikatoren basieren, Konsistenzbedingungen definieren.

Es wird nun ein Metamodell eingeführt, mit welchem sich Attributwertkombinationen in multidimensionalen Datenmodellen definieren lassen. Das multidimensionale Datenmodell gibt für das betrachtete Zielsystem die geforderten Daten vor, so dass der Informationsbedarf für das Zielsystem explizit modelliert werden kann. Dimensionen können sowohl aus einzelnen Datenquellen als auch aus Metadatenquellen wie ICD- oder OPS-Kataloge einbezogen werden. Durch das Metamodell können für jedes Zielsystem valide und invalide Verbindungen zwischen Ebenen und Kategorien angegeben werden, so dass bereits bei der Modellierung der ETL-Prozesse Konsistenzbedingungen definiert werden können.

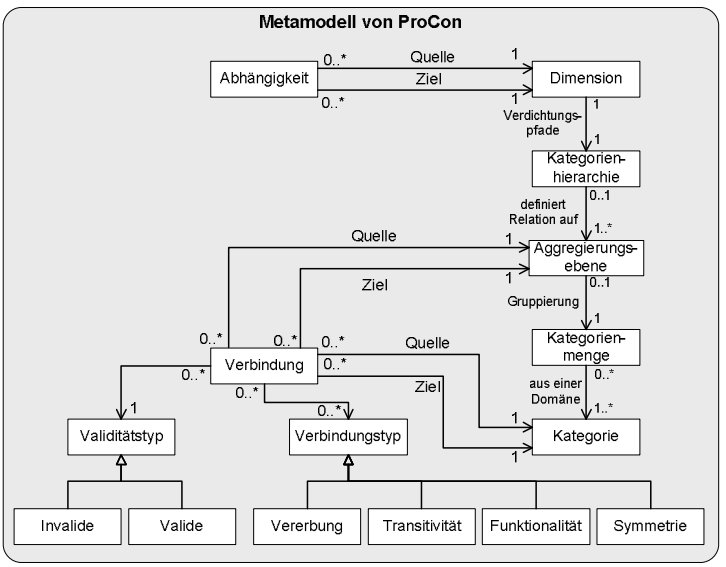

Abbildung 3: Metamodell zur Definition von Konsistenzbedingungen

In Abbildung 3 ist ein Ausschnitt aus dem MADEIRA-Metamodell um Verbindungen erweitert worden. Verbindungen haben immer eine Quelle und ein Ziel und können somit zwei Entitäten verbinden. Kategorien lassen sich mit Kategorien und Ebenen aus anderen Dimensionen verbinden. Gleiches gilt für Ebenen. Für jede Verbindung lässt sich definieren, ob sie einen validen oder invaliden Sachverhalt beschreibt. Valide Sachverhalte beschreiben gültige Attributwertkombinationen und geben somit erlaubte Kombinationen vor. Durch die Modellierung invalider Sachverhalte wird festgelegt, welche Attributwertkombinationen ungültig sind. Solche Daten lassen sich somit direkt bei der Informationsbeschaffung abweisen. Weiter können Verbindungen Eigenschaften zugeordnet werden. Die Eigenschaft „Vererbung“ einer Verbindung zu einer Ebene beschreibt, dass diese Verbindung auch zu allen Kategorien in dieser Ebene gilt. Transitivität gibt an, dass eine Verbindung traversiert werden kann. Eine funktionale Verbindung beschreibt, dass eine Kategorie oder eine Ebene ausschließlich mit der verbundenen gültig ist, und mit keiner anderen. Ist eine solche Verbindung zusätzlich symmetrisch, so ist ausgesagt, dass beide Verbindungsenden ausschließlich mit dem anderen Ende verbunden werden dürfen. Damit Regeln nicht zwischen beliebigen Dimensionen erstellt

werden, sondern nur zwischen solchen, zwischen denen es semantische Zusammenhänge gibt, kann man mit Hilfe des Metamodells solche Abhängigkeiten modellieren. Da Abhängigkeiten jeweils eine Quelle und ein Ziel haben, lassen sich so zwischen Dimensionen semantische Beziehungen modellieren.

Nach der Modellierung der Konsistenzbedingungen werden diese in Ontologiestrukturen überführt. Ontologien lassen sich ideal auf Widersprüche analysieren. Hier können sich gegenseitig widersprechende Konsistenzbedingungen aufgespürt werden. Diese können dann innerhalb der Aktivität *Konsistenzbedingungen modellieren* dem Fachanwender wieder präsentiert werden, so dass diese analysiert und neu adjustiert werden können.

6 Evaluation

Zum Zwecke eines „Proof of Concept“ wurde ein Prototyp entwickelt (siehe Abschnitt 6.1). In Abschnitt 6.2 ist ein Konsistenzszenario exemplarisch dargestellt, welches mit diesem Prototyp modelliert worden ist.

Abbildung 4: Definition einer domänenspezifischen Sprache zur Erstellung von Konsistenzbedingungen

6.1 Prototypische Realisierung

Zur Erstellung der Konsistenzregeln wurde eine domänenspezifische Sprache (DSL) [LKT04] definiert, die in Abbildung 4 dargestellt ist. Auch hier sind wieder die Elemente Dimension, Ebene und Kategorie enthalten. Um multidimensionale Strukturen zu modellieren, lassen sich Kategorien Ebenen und auch Ebenen Dimensionen zuordnen. Ebenso gibt es die Elemente zur Definition von Verbindungen zwischen Kategorien und

Kategorien, Kategorien und Ebenen sowie Ebenen und Ebenen. Diese Sprache wurde mit Hilfe der Microsoft DSL Tools [Cor08] erstellt. Für diese Sprache lässt sich automatisch ein grafischer Editor erzeugen, mit welchem sich dann Konsistenzbedingungen in multidimensionalen Datenmodellen erstellen lassen, die gültige Instanzen der DSL und des Metamodells sind.

Ein Überblick über den modellgetriebenen Ansatz der Modellierung und Erstellung der Konsistenzregeln ist in Abbildung 5 dargestellt. In Teil A ist das Metamodell ProCon dargestellt. Es enthält die relevanten Elemente aus MADEIRA und die Konzepte zur Definition von Konsistenzbedingungen. Dieses Metamodell wird im nächsten Schritt auf eine DSL abgebildet (Teil B). Diese Sprache besteht aus den Entitäten und Beziehungen des Metamodells. Aus dieser DSL lässt sich dann automatisch ein Editor erzeugen (Teil C). Die mit diesem Editor modellierten Strukturen werden dann in eine Ontologie überführt und dort auf Widerspruchsfreiheit geprüft. Auf Grund dieses Vorgehens ist die erzeugte Ontologie sowohl Instanz des Metamodells als auch der domänenspezifischen Sprache.

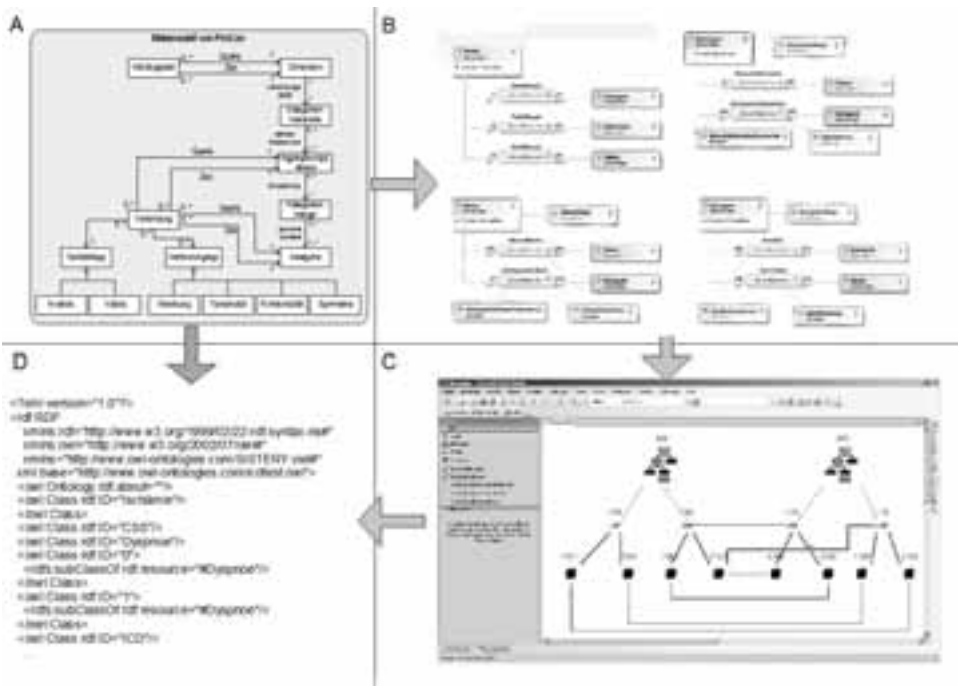

Abbildung 5: Abbildungsschritte: Das Metamodell ProCon wird in eine DSL überführt. Aus dieser wird dann ein Editor generiert. Die hierin erzeugte Struktur wird in eine Ontologie transformiert. Diese Ontologie ist Instanz des Metamodells ProCon.

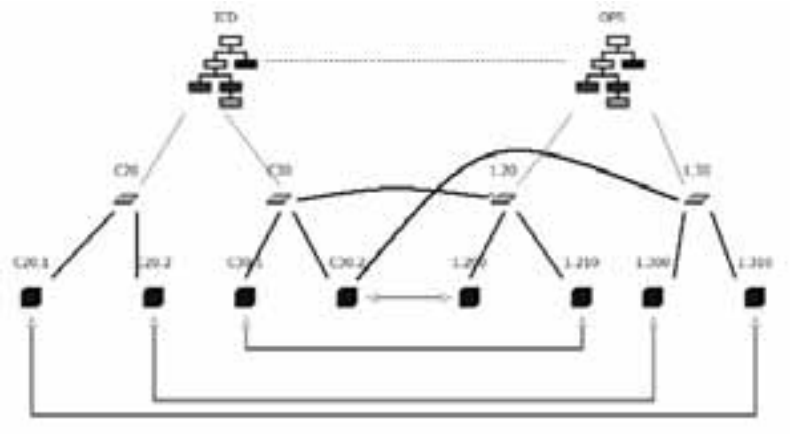

Abbildung 6: Einige mittels des generierten Editors modellierte Konsistenzbedingungen im Kontext der Dokumentation für die Qualitätskontrolle der BQS als exemplarische Anwendung des Ansatzes auf eine konkrete Domäne

6.2 Beispiel

Mit ProCon wurde in dieser Arbeit ein allgemeingültiger Ansatz zur Definition von Konsistenzbedingungen im APM eingeführt. Durch die Verwendung des Metamodells ist der Ansatz auf beliebige Domänen anwendbar. Der realisierte Prototyp wurde genutzt, um die in diesem Beitrag eingeführten Konzepte anhand des Beispielszenarios „Qualitätsmessung im deutschen Gesundheitswesen“ zu evaluieren. Zu diesem Zweck wurden zunächst mit dem auf dem ProCon-Metamodell basierenden Werkzeug die Dimensionen ICD und OPS beispielhaft modelliert.

Abbildung 6 verdeutlicht dieses Szenario. Zwischen ICD und OPS wurde eine semantische Abhängigkeit definiert (gestrichelte Linie). In der Dimension ICD wurden die Ebenen und Kategorien „C20“, „C30“, „C20.1“, „C20.2“, „C30.1“ sowie „C30.2“ modelliert. In der Dimension OPS wurden in der Ebene „1.20“ die Kategorien „1.200“ und „1.210“, in der Ebene „1.30“ die Kategorien „1.300 und „1.310“ erzeugt. Daraufhin wurden folgende Konsistenzbedingungen definiert: Es ist angegeben, dass der ICD-Code „C30“ nicht in Kombination mit den jeweiligen „1.20“-Werten der anderen Dimensionen auftreten darf (geschwungene Verbindung). Da diese Verbindung auch Grund der hierarchischen Struktur der OPS-Klassifikation und der Nutzung des Verbindungstyps „Vererbung“ zu den Kategorien „1.200“ und „1.210“ vererbt wird, ist der Wert „C30“ auch in Kombination mit „1.200“ und „1.210“ ungültig. Weiter ist modelliert worden, dass die Kombination von „C30.2“ und den Kategorien der OPS-Ebene „1.30“ „1.300“ und „1.310“ ungültig ist (stark geschwungene Verbindung). Im unteren Teil der Abbildung 6 sind einige valide Verbindungen (durchgezogene Linien) angegeben, beispielsweise darf die Kategorie „C20.2“ mit „1.300“ kombiniert werden. Ebenfalls darf „C30.1“ mit „1.210“ in Kombination verwendet werden. Dies ist jedoch eine widersprüchliche Modellierung, da dies im Gegensatz zur vormals definierten Ausage steht, dass „C30“ und alle spezielleren Werte nicht mit den OPS-Werten der Ebene „1.20“

kombiniert werden darf. Dieser Widerspruch lässt sich mit den in Kapitel 3.4 beschriebenen Methoden nicht aufdecken. Dies wird erst durch eine Überführung der ProCon-Modelle in Ontologien ermöglicht. Durch eine Validierung der Ontologie lässt sich dieser Widerspruch erkennen. Damit lässt sich die Definition solcher widersprüchlicher Aussagen vermeiden.

7 Zusammenfassung und Ausblick

In dieser Arbeit wurde mit ProCon ein Metamodell eingeführt, mit welchem Konsistenzbedingungen in multidimensionalen Datenmodellen definiert werden können. Das Vorgehensmodell des Analytischen Performance Management wurde aufgegriffen und um das proaktive Management des Datenqualitätsmerkmals Konsistenz erweitert.

Die in diesem Beitrag eingeführten Konzepte wurden beispielhaft auf das Szenario „Qualitätsmessung im deutschen Gesundheitswesen“ angewandt. Hier wurde gezeigt, daß bekannte Konzepte wie multidimenionale Modellierung in diesem Ansatz genutzt werden. Durch die explizite Verwendung eines Metamodells ist die Anwendung des Ansatzes nicht auf die beschriebene Domäne beschränkt, sondern auf andere Domänen übertragbar.

Durch die konsequente Definition von Konsistenzbedingungen direkt bei der Modellierung multidimensionaler Datenstrukturen durch Fachexperten lässt sich Datenqualität in Organisationen proaktiv verbessern. Wird bei der Modellierung des Informationsbedarfs explizit ausformuliert, welche Daten zulässig und welche unzulässig sind, können Inkonsistenzen von vornherein vermieden werden. Bereits bei der Entstehung der Daten greifen diese Konsistenzbedingungen, so dass Anwender diese Mängel direkt bereinigen können. Wurden Datenqualitätsmängel bislang erst bei der Datenintegration beispielsweise in ETL-Prozessen entdeckt und aufwändig korrigiert, so definiert der hier vorgestellte Ansatz bereits für die Datenentstehung hohe Qualitätsanforderungen.

Der hier vorgestellte Prototyp wird zurzeit erweitert. Besonderes Augenmerk wird auf die Widerspruchsfreiheit der modellierten Bedingungen gelegt. Dazu wird die Struktur in eine ontologiebasierte Repräsentation überführt. Mit Hilfe geeigneter Reasoner (z.B. Racer [HM03]) lässt sich diese Struktur dann validieren.

Im Kontext modellgetriebener Integration kann Datenqualitätsmanagement ebenfalls explizit berücksichtigt werden. Im Sinne eines Top-Down-Ansatzes können Konsistenzbedingungen bereits bei der Modellierung der Integrationsservices definiert werden. Diese werden dann bei der Datenintegration direkt genutzt.

8 Literaturverzeichnis

[BA07] Brüggemann, S.; Aden, T: Ontology Based Data Validation and Cleaning: Restructuring operations for ontology maintenance. In (Hitzler, P.; Sure, Y.; Hrsg.), GI Proceedings 109, Band 1, Jgg. 4 of LNI. GI, 2007.

[BM99] Bothner, U.; Meissner, F. W.: Data Mining und Data Warehouse: Wissen aus medizinischen Datenbanken nutzen. Deutsches Ärzteblatt, 96, 1999.

[BQS08] BQS Bundesgeschäftsstelle Qualitätssicherung gGmbh Online. http://www.bqs-online.de, 2008. zuletzt besucht am 12.02.2008.

[CG06] Chamoni, P.; Gluchowski, P., Hrsg.: Analytische Informationssysteme. Springer, 2006.

[Cor08] Microsoft Corporation: Domain Specific Language Tools. Internet: http://msdn2.microsoft.com/en-us/vstudio/aa718368.aspx, 2008. Zuletzt besucht am 12. März 2008.

[FH76] Fellegi, I. P.; Holt, D: A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Assocation, 71:17–35, 1976.

[Gru93] Gruber, T. R: A Translation Approach to Portable Ontologies. Knowledge Acquisition, 5(2):199–220, 1993.

[HM03] Haarslev, V.; Möller, R: Racer: A Core Inference Engine for the Semantic Web. In 2nd International Workshop on Evaluation of Ontology-based Tools, Seiten 27–36, 2003.

[HSW07] Herzog, T. N.; Scheuren, F. J.; Winkler, W. E: Data Quality and Record Linkage Techniques. Springer, 2007.

[ICD] ICD 10: International Statistical Classification of Diseases and Related Health Problems, 10th edition. Geneva: American Psychiatric Association; Jun 1, 1992. World Health Organization.

[KM06] Koch, S.; Meister, J.: Adaptives Performance Management mit Annotierten Strategy Maps. In (Schelp, J.; Winter, R.; Frank, U.; Rieger, B.; Turowski, K., Hrsg.): Integration, Informationslogistik und Architektur. Proceedings der DW 2006, Jgg. 90, Seiten 13–33, 2006.

[Leh03] Lehner, W.: Datenbanktechnologie für Data-Warehouse-Systeme : Konzepte und Methoden. Heidelberg : dpunkt, 2003.

[LKM06] Lenz, H-J.; Koppen, V.; Muller, R.M.: Edits – Data Cleansing at the Data Entry to assert semantic Consistency of metric Data. ssdbm, 0:235–240, 2006.

[LKT04] Luoma, J., Kelly, S.; Tolvanen, J.-P.: Defining Domain Specific Modeling Languages: Collected Experiences. In Proceedings of the 4th OOPSLA Workshop on Domain-Specific Modeling, 2004.

[LN07] Leser, U.; Naumann, F.: Informationsintegration. dpunkt.verlag, 2007.

[MF03] Müller, H.; Freytag, J.-C.: Problems, Methods, and Challenges in Comprehensive Data Cleansing. Bericht, Humboldt Universität Berlin, 2003.

[MN07] Martin, W.; Nußdorfer, R.: CPM – Corporate Performance Management, Kompendium: Analytische Services in einer SOA, Teil 1: Herstellerunabhängige Beschreibung und Referenzarchitektur. August 2007. http://www.competence-site.de/soa.nsf/3B1B7B84328E667AC12573960074221A/$File/cpm_analytische_services_soa.pdf (zuletzt besucht am 30.11.2007).

[MR01] Manzari, A.; Reale, A.: Towards a new method of edit and imputation of the Italian Census: a Comparison with the Canadian Nearest-Neighbour Methodology. Presented at the International Statistical Institute Meeting in Seoul, Korea,, 2001.

[Nau02] Naumann, F.: Quality-Driven Query Answering for Integrated Information System. LNCS 2261, 2002.

[Org78] World Health Organization: International Classification of Procedures in Medicine. 1978.

[OWS+08] Otto, B.; Wende, K.; Schmidt, A.; Kai Hüner, K.; Vogel, T.: Integrierte Informationslogistik, Kapitel Unternehmensweites Datenmanagement, Seiten 211–230. Springer, 2008.

[RD00] Rahm, E.; Do, H. H: Data Cleaning: Problems and Current Approaches. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 23(4):3–13, 2000.

[SMB05] Scannapieco, M.; Missier, P.; Batini, C.: Data Quality at a Glance. Datenbank Spektrum, 14:6–14, 2005.

[Wie00] Wietek, F.: Intelligente Analyse multidimensionaler Daten in einer visuellen Programmierumgebung und deren Anwendung in der Krebsepidemiologie. Dissertation, Universität Oldenburg, 2000.

[Win04] Winkler, W. E.: Methods for Evaluating and Creating Data Quality. Information Systems, 29, 2004.

[WP97] Winkler, W. E.; Petkunas, T.: Statistical Data Editing, Kapitel The DISCRETE Edit System, Seiten 56–62. U.N. Economic Commission for Europe, Geneva, 1997.

[WQ03] DeWaal, T.; Quere, R.: A fast and simple algorithm for automatic editing of mixed data. J. Official Statist, 19, 2003.

[WSDB08] Winter, R.; Schmaltz, M.; Dinter, B.; Tobias Bucher. Integrierte Informationslogistik, Kapitel Das St. Galler Konzept der Informationslogistik, Seiten 1–16. Springer, 2008.