Ein Plädoyer für die Berücksichtigung von Semantik beim Stammdaten-Alignment – Vorgehensmodell und prototypische Anwendung im Einzelhandel

Axel Winkelmann, Martin Matzner, Oliver Müller, Jörg Becker

Institut für Wirtschaftsinformatik
Westfälische Wilhelms-Universität Münster
Leonardo-Campus 3
48149 Münster
{axel.winkelmann | martin.matzner | oliver.mueller | joerg.becker}@ercis.uni-muenster.de

Abstract: Eine Automatisierung von Prozessen mithilfe einer durchgehenden informationssystematischen Unterstützung stellt hohe Anforderungen an die Qualität der zugrundeliegenden Stammdaten. In der Domäne „Handel" adressieren Händler und Produzenten mangelnde Stammdatenqualität mit Projekten zur Vermeidung syntaktischer Datenfehler. Auch die Forschung widmet sich vornehmlich Methoden und Techniken zur Entdeckung, Vermeidung und Korrektur syntaktischer Datenfehler (z. B. fehlender Werte, Tippfehler). Semantische Belange wurden und werden bislang vernachlässigt. Unzureichende semantische Datenqualität kommt in Abweichungen zwischen relevanten Realweltentitäten, deren Attributen und Beziehungen einerseits und ihrer Repräsentation durch Daten andererseits zum Ausdruck. Semantische Datenqualität bezieht sich primär auf die gegenwärtige (fachliche) Verwendung gespeicherter Daten. Sie ist jedoch nicht auf die aktuelle Nutzung der Daten und ihren Einfluss auf bestehende Geschäftsprozesse beschränkt, sondern beinhaltet darüber hinaus auch mögliche künftige Verwendungsformen der Daten. Dieser Beitrag untersucht, inwiefern Ontologien ein geeignetes Instrument zur Identifikation und Bewertung semantischer Stammdatenfehler darstellen. Ein konzeptueller Ansatz und ein Vorgehensmodell zur Nutzung von Ontologien zur Erhöhung semantischer Stammdatenqualität werden entwickelt. Als beispielhaftes Anwendungsszenario dient das automatische Coupon-Clearing an der Kasse im Einzelhandel.

1 Motivation

Unternehmen in Industrie und Handel sind in verstärktem Maße auf korrekte und zuverlässige Stammdaten angewiesen. Im Handel werden Artikelstammdaten begleitend zu den Produkten unmittelbar von den Herstellern bereitgestellt. Die Produzenten wiederum benötigen Daten vom Einzelhandel, um z. B. eine kooperationsweite Optimierung des Güterflusses innerhalb der Lieferkette unter Berücksichtigung aktueller Kundenanforderungen durchführen zu können. Um die Datenqualität zu verbessern, haben Produzenten

und Handel begonnen, Stammdaten zu harmonisieren und zentral zu organisieren [Lo05]. Außerdem haben elektronische Marktplätze, kollaborative Führungs- und Planungskonzepte wie ECR (Efficient Consumer Response) und CPFR (Collaborative Planning, Forecasting and Replenishment) und die Einführung von zentralen Stammdatenpools die Unternehmen dazu gezwungen, die Qualität ihrer Stammdaten zu überprüfen und zu verbessern. Dennoch bleiben ein unzureichender Strukturierungsgrad der Daten und eine sehr variable und häufig unzureichende Qualität der Datenobjekte selbst bedeutende Herausforderungen im Rahmen des Stammdatenmanagements. Die Qualität der Daten ist dabei nicht als „isolierte Funktion" eines oder mehrerer betrieblicher Informationssysteme zu verstehen, sondern ist ein inhärenter und integraler Bestandteil des gesamten Business Managements [En99].

Das Problem mangelnder Datenqualität betrifft vor allem die durch die Produzenten bereitgestellten Stammdaten. Hier mangelt es derzeit noch an standardisierten Richtlinien für Struktur und inhaltliche Beschreibung [BD07]. Beispielsweise werden identische Artikel oft inkonsistent in Datenbanken gespeichert, Schlüssel werden verschiedenartig gebildet und zahlreiche Abkürzungen verwendet. Mangelhafte Datenqualität zieht dann eine Vielzahl an Fehlern bei der Ausführung von Geschäftsprozessen nach sich und führt im Ergebnis zu einer ineffizienten Nutzung von Ressourcen in Form von Personen, Geld, Material und Maschinen [En99]. Im Handel ergibt sich als eine Konsequenz, dass bislang die Einführung von flexiblen Datenstrukturen, die über eine generische Klassifikation von Waren hinausgehen, nicht möglich ist. Aufgrund dieses Mangels an Struktur und Qualität, können Artikelstammdaten selten unmittelbar in hoch automatisierten Geschäftsprozessen verwendet werden. Vielmehr ist das Mitwirken von Menschen zum manuellen Interpretieren, Aktualisieren und Restrukturieren der Daten erforderlich [Ca04].

Als wissenschaftliche Disziplin befasst sich die Semiotik mit der zeichenbasierten Repräsentation von Realweltentitäten. Mit Rückgriff auf die semiotischen Teildisziplinen schlagen PRICE and SHANKS [PS05] syntaktische, semantische und pragmatische Kriterien für die Evaluation von Datenqualität im Allgemeinen vor. Andere Forscher zeigen Methoden und Techniken zur Beseitigung so entdeckter syntaktischer und pragmatischer Qualitätsdefizite auf [Re96]. Nur wenige Forscher diskutieren jedoch konkrete Ansätze oder praktische Richtlinien für semantische Aspekte der Stammdatenqualität [CFP03]. Ferner betrachten bisherige Untersuchungen Datenqualität vornehmlich auf der aktuellen Nutzung der Daten. Datenqualität ist jedoch nicht auf die aktuelle Verwendung von Daten und deren aktuellen Einfluss auf Geschäftsprozesse beschränkt, sondern beinhaltet auch die zukünftige Nutzung der Daten [En99]. Schließlich sind die Stammdaten eines Unternehmens die Grundlage für künftige Geschäftsideen und insbesondere Dienstleistungen, die bei der Konzeption derzeitig genutzter Informationssysteme noch keine Berücksichtigung finden konnten.

In diesem Artikel diskutieren wir den Beitrag von Ontologien zur Identifikation und Bewertung von Stammdatenqualitätproblemen auf semantischer Ebene. Wir erläutern unseren Ansatz mit Rückgriff auf das Szenario des automatischen Coupon-Clearings in Einzelhandelsunternehmen. Bei manuellen, wenig automatisierten Coupon-Clearing-Prozessen an der Kasse sind Einzelhändler nicht auf die Nutzung von Artikelstammdaten ange-

wiesen. Das Kassenpersonal prüft Coupons durch Inaugenscheinnahme manuell und entnimmt die durch die Promotion eingeräumten Kaufkonditionen einem Aufdruck auf dem Coupon. Die Einführung von Self-Checkout-Systemen und organisierte Coupon-Betrügereien zwingen den Handel jedoch, Coupon-Clearing-Prozesse zu automatisieren. Es entstehen dadurch neue Anforderungen an die Qualität der Artikelstammdaten [Wi06].

Dieser Artikel ist wie folgt strukturiert: In Kapitel 2 diskutieren wir den Stand der Forschungsbemühungen zu Stammdatenqualität und erläutern die verschiedenen adressierten Problemdimensionen. In Kapitel 3 beschreiben wir einen Ansatz zur Vermeidung semantischer Stammdatenqualitätsprobleme bei der Verwaltung von Artikelstammdaten durch die Nutzung von Ontologien. In Kapitel 4 erläutern wir den Beitrag unserer Arbeit für Theorie und Praxis, illustrieren Beschränkungen des Konzepts und geben einen Ausblick auf künftige Forschungsaktivitäten in diesem Bereich.

2 Status-quo der Forschung zur Qualität von Stammdaten

2.1 Stammdaten

Stammdaten umfassen diejenigen Informationsobjekte, die benötigt werden, um ein unternehmensweites „Berichtssystem“ für grundlegende betriebliche Funktionen zu schaffen und aufrecht zu halten, um betriebliche Transaktionen zu erfassen und die Ergebnisse dieser betrieblichen Funktionen zu messen [BD07] [Gr05]. Im Handel werden verschiedenartige Stammdaten benötigt und verwaltet. Händler benötigen Stammdaten einerseits über Kunden und Lieferanten wie etwa Name, Adresse, Bankverbindungen etc. und andererseits über die verwalteten Artikel. Artikelstammdaten setzen sich aus Grunddaten (z. B. Artikelnummer, UPC, EAN), Listungsdaten (Listungszeitraum, Zuordnung zu Sortimenten), Bezugsdaten (z. B. lieferantenabhängige Verfügbarkeiten), Logistik-, Absatz- und Kassendaten [BUV01] zusammen.

Automatisierte Kassenprozesse greifen auf Artikelstammdaten und Promotionsstammdaten zurück. Artikelstammdaten beinhalten Informationen über Produktname, Größe, Preis etc. Zur Identifikation von Artikelstammsätzen wird auf eindeutige Bezeichner wie die UPC (Unique Product Code (USA)), die EAN (European-/International Article Number) oder die JAN (Japanese Article Number) zurückgegriffen. Promotionsstammdaten beinhalten neben relevanten Informationen für Promotionen wie dem Wert des Promotionsangebotes, Gültigkeitsbeschränkungen u. ä. auch eine Referenz zum beworbenen Produkt. Zum Beispiel benötigt der Händler zur automatisierten Abwicklung einer BOGOF-Promotion (buy one, get one free) Promotionsstammdaten über den Wert der Promotion („ein Artikel umsonst“), die Laufzeit der Promotion („bis zum 31. Juli“), und das beworbene Produkt beziehungsweise den Produktcode des Produkts (z. B. „Coca-Cola 0,5 Liter“, EAN: 4023786889760, EAN: 4046758456739, EAN:…).

2.2 Stammdatenqualität

Zahlreiche Forschungsaktivitäten ergründeten Kriterien, in denen Daten- und Informationsqualität zum Ausdruck kommt.[1] Mehrere Autoren bilanzierten in den vergangenen Jahren eine tendenziell unzureichende Stammdatenqualität in Industrie und Handel. KUIPERS berichtet zum Beispiel, Nestlé vermute, dass zu über 50% der gelisteten Produkte inkorrekte oder redundante Stammdateneinträge existierten [Ku04]. Im Rahmen einer 2005 in den USA durchgeführten Studie wurde darüber hinaus erhoben, dass 30% der Data-Warehouse-Betreiber mit massiven, durch unzureichende Datenqualität bedingten, Problemen kämpften. Für eine generellere Einschätzung zum Thema „Datenqualität" sei auf die Arbeiten von [Ag05] sowie [BW08] verwiesen. Eine aktuelle Untersuchung zu den Defiziten bei der Datenqualität von Verbundgruppen im Handel findet sich bei [BGW08].

Eine Quintessenz der Forschungsaktivitäten der vergangenen zwei Jahrzehnte ist, dass es sich bei Datenqualität um ein „multidimensionales Konzept" handelt; vgl. [PLW02] und die dort genannten Quellen. Es lassen sich unterscheiden: Eine (a) *subjektiv* empfundene Datenqualität, die in der Einschätzung der Datennutzer zum Ausdruck kommt, und eine (b) *objektive* Bewertung der betrachteten Datensätze anhand akzeptierter Qualitätskriterien. Andere Autoren beschreiben diese Phänomene mit den Ausdrücken *pragmatische* (a) und *inhärente* (b) Datenqualität [En99]. Qualität auf pragmatischer Ebene bezeichnet Korrektheit im Sinne einer korrekten Repräsentation (a.I) „richtiger" Fakten [En99]. Das Hauptaugenmerk liegt hier auf der Bereitstellung und schließlichen Nutzung der Datenobjekte durch den Anwender. Qualität im Sinne inhärenter Datenqualität ist beschränkt auf die Korrektheit (b.II) der gespeicherten Fakten und eine klare Definition (b.III) und damit Bedeutung der Daten. Die korrespondierenden Subkategorien der semiotischen Lehre sind *Syntax* und *Semantik* [PS05]. Für alle drei Komponenten (Definition, Korrektheit der Fakten und Repräsentation) können voneinander losgelöst Maßnahmen zur Erhöhung der Datenqualität durchgeführt werden [En99]. Eine separate Messung der Qualitätskomponenten wird durch Interrelationen hingegen verhindert [LSK02]. Der folgende Abschnitt adressiert die semantischen Aspekte von Datenqualität und fokussiert dabei die Datendefinitionskomponente (b.III) semantischer Stammdatenqualität.

2.3 Semantische Stammdatenqualität

Semantische Stammdatenqualität ist ein Ausdruck für das Maß an Übereinstimmung zwischen den Daten, die in einer Datenbank abgelegt sind, und der Summe an Eigenschaften externer Realweltphänomene, die sie repräsentieren sollen und die für den mit der Datenspeicherung verbundenen Zweck relevant sind [PS05]. Eine etablierte Methode zur Bewertung *semantischer Datenqualität* ist das Random Sampling dieser Abbildungen. Datenqualität auf semantischer Ebene in diesem Sinne ist ein Ausdruck der Über-

[1] Die Begriffe Daten- und Informationsqualität werden hier synonym verwendet; vgl. dazu [HLW99] [KPM04]).

einstimmung einer epistemologischen Sicht auf die Datenobjekte mit der korrespondieren ontologischen Beschreibung der externen Phänomene [HKL95].

Im Gegensatz zu diesem umfassenden Verständnis, beschränken viele Autoren Aspekte semantischer Qualität auf die Vermeidung von Inkonsistenzen, die durch redundante Speicherung von Informationselementen auftreten (z. B. auch CODD in seinem wegweisenden Beitrag [Co70]), vernachlässigen jedoch (a) das korrekte Mapping von Daten zu den Realweltentitäten während des gesamten Lebenszyklus von Realweltentitäten und Daten, und (b) die Eignung der Daten zur Verwendung in künftigen neuen fachlichen Kontexten.

(a) Korrektes Mapping: Das Postulat des korrekten Mappings korrespondiert mit dem Konzept des „Information as a Product" [Pi04] [Wa98] [WYP98]. Datenqualität kann hier einerseits durch die Anwendung von Methoden sichergestellt werden, die die Konformität der Daten zu einem initial spezifizierten Anforderungskatalog und weiteren Integritätsregeln garantieren. Alternativ können externe Phänomene z. B. mithilfe von Ontologien [PS05] fachlich beschrieben und die Entsprechung mit den gespeicherten Daten geprüft werden. Ebenso wie physikalische Produkte folgen Daten einem Lebenszyklus. Dabei ist die Veränderung der Daten im Zeitablauf eine Funktion der Informationssysteme, die die Unternehmensdaten erstellen und ändern [HLW99]. Der Datenlebenszyklus beschreibt Erstellung, Veränderung und Löschung der Datenobjekte.

(b) Nutzung im neuen Kontext: Nicht nur die Stammdaten, auch die abgebildeten Realweltentitäten, die Geschäftsregeln, die diese miteinander verbinden, und der umgebende fachliche Anwendungskontext sind dynamisch. Folglich kann die spätere operative Nutzung der Daten, bedingt durch nicht berücksichtigte, unvorhergesehene oder veränderte fachliche Anforderungen, sich grundlegend von derjenigen unterscheiden, die während der Systementwicklung spezifiziert wurde [PS05]. Diese Aspekte sollen unter dem Begriff *kontextuelle Datenanforderungen* subsumiert werden. Beispiele im Handel sind umfassende Dokumentations- und Berichtspflichten, die z. B. die Verwendung von Inhaltsstoffen (z. B. genetisch veränderte Inhaltsstoffe), Gefahrgütern (z. B. REACH, ROHS oder WEEE) oder artikelspezifischen Informationen zur Erstellung finanzwirtschaftlicher Berichte betreffen. Gerade hier sehen sich Unternehmen mit zahlreichen neuen Gesetzen und Verordnungen (z. B. Sarbanes-Oxley, Basel II) konfrontiert, die sie zur Bereitstellung, Nutzung und zum Reporting aktueller, verifizierbarer und relevanter Daten über die allgemeine betriebswirtschaftliche Leistungsfähigkeit und signifikante Ereignisse zwingen [BD07].

Zum Verständnis von kontextabhängigen Aspekten der Stammdatenqualität sind die Metadaten der Datenobjekte zu berücksichtigen. Die Metadaten beinhalten Definitionen und Dokumentationen, die einerseits die Domäne der fachlichen Nutzung in der Realwelt und andererseits das Datenmodell beschreiben [PS05]. Sie adressieren somit gezielt die zuvor identifizierte Lücke zwischen Realweltbeschreibung und Datenbankinhalt.

2.4 Forschungsmethodischer Ansatz

Unsere Arbeit untersucht das Konstrukt „Ontologiebasiertes semantisches Stammdatenmanagement“. Als qualitative Forschungsmethode wurde „Design Science“ zur Führung des Forschungsprozesses gewählt. Nach HEVNER U. A. versucht Design-Science-Forschung, IT-Artefakte zu erstellen und zu evaluieren, um identifizierte organisatorische Probleme zu lösen [HMP04]. Für die Ermittlung und Beschreibung dieser Probleme gilt es, im Rahmen eines Design-orientierten Forschungsprozesses insbesondere, Bedeutung und Relevanz der adressierten Problembereiche für die untersuchte Domäne zu belegen. Dazu kann auf eine Vielzahl von Forschungsmethoden wie Literaturanalyse, fallstudienbasierte Untersuchungen und Experteninterviews [Ja91, Mi97, Sc91] oder eine pluralistische Kombination dieser Ansätze zurückgegriffen werden [Pe07]. Die in diesem Artikel dargelegten Erkenntnisse basieren auf praktischen Erfahrungen aus zahlreichen Experteninterviews und sechs Stammdatenmanagementprojekten, die in den vergangenen Jahren am Institut begleitet wurden. Der Artikel beschreibt die Weiterentwicklung des in Zusammenarbeit mit einem Clearingunternehmen entwickelten ontologischen Stammdatenmanagementansatzes, der auf den Erfahrungen aus den genannten Projekten (vgl. Tabelle 1) aufsetzt (vgl. auch [BJ07a, BJ07b, BJ07c, Wi06]).

Fallstudie	**Unternehmen**	**Durchgeführte Projekte**
Handelsunternehmen	Großhändler mit 400 Mitarbeitern, 50 Großmärkten und annähernd 30.000 Artikeln.	Prozessanalyse, Stammdatenreorganisation sowie Auswahl und Analyse eines neuen ERP-Systems.
Handelsunternehmen	Juwelier- und Uhrmacherhandelskette mit 200 Geschäften und 2.000 Mitarbeitern in ganz Deutschland.	Reorganisation des Controlling-Systems mit Vereinheitlichung des Stammdaten- und Kennzahlenkonzepts. Einführung eines Data-Warehouse.
Gebäudeserviceunternehmen	Komplettserviceanbieter für Immobilienverwaltung. Management von mehr als 35.000 Objekten, 6.800 Mitarbeiter.	Einführung eines neuen Stammdatenkonzepts.
Industrie	Kleines metallverarbeitendes Unternehmen mit 125 Mitarbeitern. Entwicklung von Produkten mit einem hohen Grad an manueller Fertigung.	Prozessautomatisierung und Stammdatenreorganisation. Einführung eines Produktionsplanungssystems.
Betreiber eines europäischen Artikelstammdatenpools	Hersteller von Lösungen für das Management von strukturierten und unstrukturierten Daten, multimedialen Inhalten und Transaktionsdaten von weit verteilten Quellen.	Analyse des existierenden Stammdatenpools und seines syntaktischen Qualitätsmaßes.
Deutsches Coupon-Clearing Unternehmen	Zehn Mitarbeiter, vereint ca. den halben deutschen und große Teile des europäischen Marktes für automatische Coupon-Prozess-Lösungen im Handel auf sich.	Entwicklung einer in-store Coupon-Prozess-Lösung. Entwicklung eines zentralisierten Stammdatenpools für Promotionsstammdaten.

Tabelle 1: Berücksichtigte Fallstudien und Projekte zum Stammdatenmanagement

3 Ontologie-basiertes, semantisches Stammdatenmanagement

Dieser Abschnitt beschreibt einen Ansatz zum Management semantischer Stammdatenqualität. Wie bereits dargelegt, kann nach PRICE UND SHANKS [PS05] syntaktische Datenqualität durch Integritätsüberprüfung, semantische Datenqualität durch Random Sampling und pragmatische Datenqualität – da sie entschieden von der Wahrnehmung des Benutzers abhängt – nur durch empirische Methoden, wie z. B. Befragungen oder Interviews, überprüft werden. Im Folgenden stellen wir einen systematischen Ansatz zur Messung der semantischen Datenqualität vor, der über die zufällige Stichprobengenerierung des Random Sampling hinaus geht.

Die Grundidee ist, ein Informationssystem als Repräsentation eines Realweltsystems zu verstehen [We97]. Ein Zustand des Realweltsystems zu einer bestimmten Zeit wird von den Daten repräsentiert, die im Informationssystem gespeichert sind [WW96]. Da Informationssysteme in der Regel nur einen unvollständigen und vereinfachten Ausschnitt der realen Welt abbilden, stellen sich zwei Fragen:

- Frage 1: Welche Realweltphänomene sollen im Informationssystem repräsentiert werden (d. h. welche Entitäten sind relevant)?
- Frage 2: Wie sollten relevante Entitäten im Informationssystem repräsentiert werden (d. h. welche Attribute und Beziehungen sind relevant)?

Die Beantwortung dieser Fragen ist eine notwendige (jedoch nicht hinreichende) Bedingung zur Beurteilung der semantischen Datenqualität eines gegebenen Informationssystems. Um dieses Problem anzugehen, schlagen wir vor, zunächst ein konzeptionelles Model zu erstellen, das die Semantik der realen Welt, die wir repräsentieren wollen, dokumentiert. In darauf folgenden Schritten wird dieses konzeptionelle Modell mit dem physischen Datenmodell des Informationssystems und dessen Datenbestand verglichen.

Ein möglicher Ansatz zur Abbildung der Semantik einer Domäne ist die Erstellung einer Ontologie. Der Begriff Ontologie entstammt der Philosophie und bezieht sich auf die Studie des Seins. Die zentrale Fragestellung dieser Disziplin lautet: „Was existiert?“ oder um präziser zu sein [So00] „Welche Kategorien von Dingen existieren?“. Eine Domänenontologie spezifiziert folglich die fundamentalen Typen von Entitäten innerhalb einer wohl abgegrenzten Domäne. Im Kontext von Informationssystemen wird eine Ontologie häufig als ein abstrakter, vereinfachter Blick auf die Welt, die wir für einen bestimmten Zweck repräsentieren möchten, definiert [Gr93]. Hier besteht eine Ontologie aus einer Menge von Konstrukten (representational primitives), mit denen eine Domäne modelliert werden kann (vgl. im Folgenden [Gr08]). Bei diesen Konstrukten handelt es sich typischerweise um Konzepte (häufig auch Klassen genannt), Attribute dieser Konzepte (häufig auch Eigenschaften genannt) und Beziehungen zwischen Konzepten. Die Definitionen der grundlegenden Konstrukte umfasst Informationen über ihre Bedeutung und Bedingungen ihrer logisch konsistenten Anwendung. Im Kontext von Datenbanksystemen können Ontologien als ein Detaillierungsgrad, der von Datenmodellen und Implementierungsfragen abstrahiert, betrachtet werden.

Liegt eine Ontologie für eine relevante Domäne der Realwelt vor, kann dieses konzeptionelle Modell sowohl mit dem physischen Datenmodell des Informationssystems als auch mit dem genutzten aktuellen Datenbestand verglichen werden. Abweichungen, die bei dieser Analyse offengelegt werden, können als mögliche Indikatoren für mangelnde Datenqualität interpretiert werden. Abbildung 1 detailliert den geschilderten Ansatz, der in die vier Phasen Analyse, Modellierung, Mapping und Anwendung unterteilt werden kann. Jede Phase beinhaltet Aufgaben, die durch Artefakte verbunden sind. Die einzelnen Phasen sind Gegenstand der folgenden Unterabschnitte. Die abgebildeten Modellelemente und deren Beziehungen werden dort explizit adressiert.

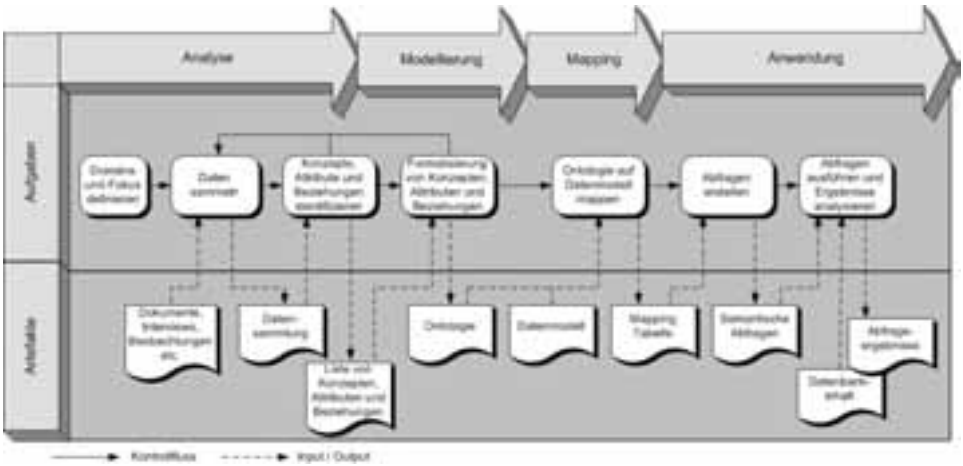

Abbildung 1: Vorgehensmodell für ein semantisches Datenqualitätsmanagement

3.1 Analyse

Der erste Schritt beinhaltet die Definition sowie klare Abgrenzung der zu behandelnden Domäne. Beispielhafte Schwerpunkte im Handel sind u. a. IT-unterstützte Geschäftsstrategien (z. B. E-Commerce), Geschäftsprozesse, die auf hohe Datenqualität angewiesen sind, (z. B. CPFR, ECR, RFID, Promotionen) oder gesetzliche Richtlinien, die Voraussetzungen für das Datenmanagement definieren (z. B. SOX, Basel II).

Im Folgenden werden wir das sogenannte Couponing als ein durchgehendes Beispiel verwenden. Coupons sind ein Marketinginstrument in Form eines gedruckten oder elektronischen Gutscheins, der dem Besitzer entweder direkt oder indirekt Rabatt zugesteht, wenn bestimmte Einlösebedingungen erfüllt sind [Wi06]. In der Vergangenheit wurden Coupons manuell am Point-of-Sale (POS) verrechnet. Heute wird automatisches Coupon-Clearing immer wichtiger, nicht zuletzt, da es viele der Nachteile des manuellen oder halbautomatischen Clearings vermeidet (insb. Zeitaufwand, Fehleranfälligkeit und Betrugsrisiko) und eine automatische Verarbeitung am POS ohne manuelle Eingriffe ermöglicht (z. B. an Self-Checkout-Automat). Das gleiche gilt für das Ausstellen von Coupons an der Kasse. Heute ist es möglich, Coupons gemeinsam mit dem Kassenbon zu drucken, wenn bestimmte Bedingungen erfüllt sind. Kauft ein Kunde zum Beispiel

einen Artikel eines bestimmten Herstellers, erhält er einen Coupon, der ihm beim nächsten Einkauf Rabatt auf andere Artikel (z. B. Neueinführungen) des gleichen Herstellers gewährt. Automatisches Couponing hängt stark von einer effektiven und effizienten IT-Unterstützung ab. Insbesondere die Qualität der benötigten Stammdaten ist eine notwendige Voraussetzung für ein automatisches Abwickeln von Coupon-Aktionen. Das Zusammenspiel von Coupon und Artikelstammdaten illustriert Abbildung 2.

Es ist ratsam, die relevante Domäne nicht nur zu identifizieren, sondern auch sorgfältig abzugrenzen. Eine bekannte Technik ist die Verwendung von so genannten motivierenden Szenarien und Kompetenzfragen ([GF95], [Us96], [UG96]). Die motivierenden Szenarien haben die Form von Geschichten, Beispielen oder Problembeschreibungen, enthalten aber auch mögliche Lösungsansätze. Diese Lösungsansätze liefern oft eine erste Idee für die beabsichtigte Semantik der Ontologie. Aus den motivierenden Szenarien kann eine Menge von Fragen, die so genannten Kompetenzfragen, abgeleitet werden. Die Kompetenzfragen decken Anforderungen bezüglich der zu entwickelnden Ontologie auf. Ihr primärer Zweck ist jedoch nicht die Generierung von ontologischen Verpflichtungen; sie sind vielmehr dazu bestimmt als eine Art Nagelprobe für die fortlaufende Evaluation der Ontologie zu dienen [NM08], z. B.: Enthält die Ontologie genug Informationen, um alle Kompetenzfragen zu beantworten? Ist das Detaillierungslevel passend, um die Kompetenzfragen effektiv und effizient zu beantworten?

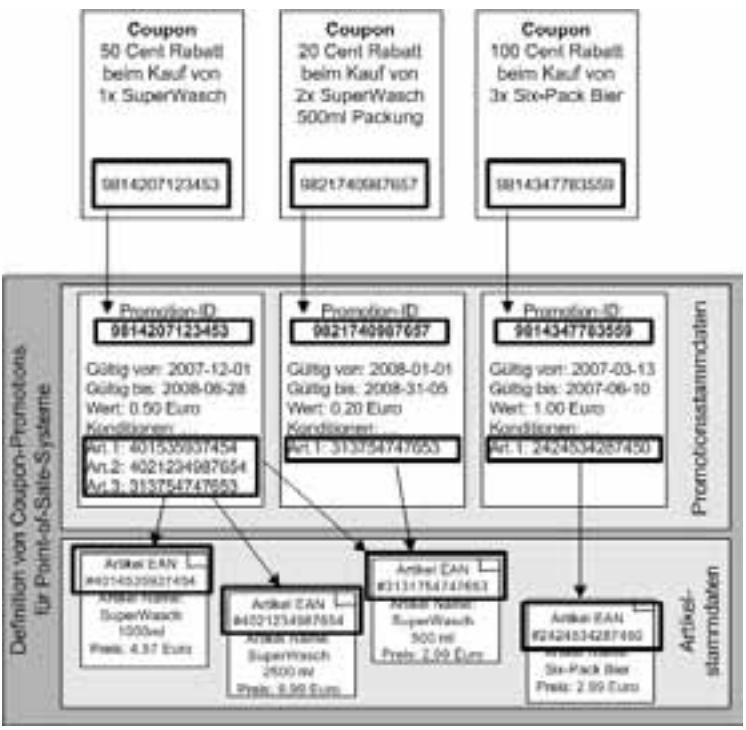

Abbildung 2: Zusammenspiel von Coupon, Barcode und Artikelstammdaten im automatischen Coupon-Clearing

Im Folgenden wird ein exemplarisches motivierendes Szenario für unser begleitendes Beispiel vorgestellt: Ein Händler will ein neues Fleckensalz bewerben. Er umwirbt Kunden, die Artikel mit fleckenverursachenden Inhaltsstoffen, wie z. B. Tomaten, kaufen. Es soll automatisch ein Coupon für eine kostenlose Packung Fleckensalz an der Kasse ausgedruckt werden, wenn ein Kunde ein Produkt kauft, das Tomaten enthält, bspw. Ketchup oder Pastasauce. Die Aktion ist auf zwei Wochen begrenzt.

Beispielhafte Kompetenzfragen:

- Welche Artikel sind aus Tomaten gemacht?
- Welche EANs (Europäische Artikelnummer) gehören zu Artikeln mit Tomaten?
- Welche EANs gehören zum Fleckensalz (oft hat ein Artikel mehr als eine EAN – abhängig vom Produktionsort etc.)?
- Läuft die Werbeaktion noch?

Ist die Domäne definiert und abgegrenzt, so folgt die Datensammlung. Eine Vielzahl von qualitativen empirischen Methoden kann dafür benutzt werden. Wichtige Datenquellen, die als Input für diese Aufgabe dienen können, sind Dokumentanalysen (z. B. von Gesetzen, Richtlinien, Geschäftsdokumenten, Geschäftsberichten, Mitschriften von Meetings), strukturierte, offene oder fokussierte Interviews (z. B. mit Domänenexperten, Mitarbeiten) und direkte oder teilnehmende Beobachtungen (z. B. von Mitarbeitern). Der Output dieser Aktivität ist eine konsolidierte Datensammlung, z. B. in Form eines zentralen Berichtes, der als Input für alle weiteren Phasen der Analyse dient. Auf der Basis dieser Datensammlung können zentrale Konzepte, Attribute und Beziehungen abgeleitet und detailliert werden. Dabei ist es hilfreich, zunächst durch Brainstorming zentrale Begriffe der Domäne zu identifizieren und sich nicht zu sehr über Klassifikationen und Überschneidungen Gedanken zu machen [Us96]. In nachfolgenden Schritten werden die gesammelten Begriffe als Konzepte, Attribute oder Beziehungen klassifiziert und ausdifferenziert, um Überschneidungen zu reduzieren. Während dieser Aufgabe kann es notwendig sein, zu vorgelagerten Schritten zurückzugehen, wenn Informationen über wichtige Aspekte fehlen oder wenn man identifizierte Konzepte, Attribute und Beziehungen evaluieren möchte. Der finale Output der Analysephase ist eine semi-strukturierte Liste von zentralen Konzepten, Attributen und Beziehungen der analysierten Domäne.

3.2 Modellierung

Der Zweck dieser Phase ist die Formalisierung der Ergebnisse der vorangegangenen qualitativen Analyse durch die Konstruktion der eigentlichen Domänenontologie. Dies bezieht typischerweise die Kodierung der identifizierten Konzepte, Attribute und Beziehungen mittels einer formalen Sprache, z. B. OWL (Web Ontology Language), RDF (Ressource Description Framework) oder UML (Unified Modelling Language) und eines Modellierungswerkzeugs (z. B. Protégé) ein.

Ein sinnvoller Ansatz ist es, die Liste der identifizierten Konzepte zunächst in einer hierarchischen Taxonomie anzuordnen. Danach wird die interne Struktur der Konzepte durch die Definition beschreibender Attribute (und ihrer Wertebereiche) spezifiziert. Der

nächste Schritt ist die Definition von Beziehungen (inklusive Kardinalitäten) unter den Konzepten. Der letzte (optionale) Schritt umfasst die Erstellung von Instanzen der Konzepte, das Ausfüllen von Attributwerten sowie die Definition konkreter Beziehungen zu anderen Instanzen.

Abbildung 3 zeigt die Ontologie für unser begleitendes Beispiels in OWL-Notation. Die zentralen Konzepte sind Coupons, Promotionen, Konditionen, Rabatte (Ermäßigungen und Naturalrabatte), Artikel und Inhaltsstoffe. Zusätzlich sind Beziehungen zwischen Konzepten enthalten, z. B. Coupons gehören zu einer Promotion, Artikel beinhalten Inhaltsstoffe und Artikel sind zu EANs zugeordnet. Die Ontologie enthält auch Instanzen von Konzepten. Zum Beispiel beinhaltet der Artikel „Mama Mia Pasta Sauce“ den Inhaltsstoff „Tomate“ und die EAN „300819814711“. Analog hat der Coupon „Coupon_4711“ die Kondition „Regel_Tomate-Fleckensalz“, welche die Prämisse Inhaltsstoff = „Tomate“ und die Konklusion „GetOneFree“ des Artikels „FleckWeg_Fleckensalz“ hat.

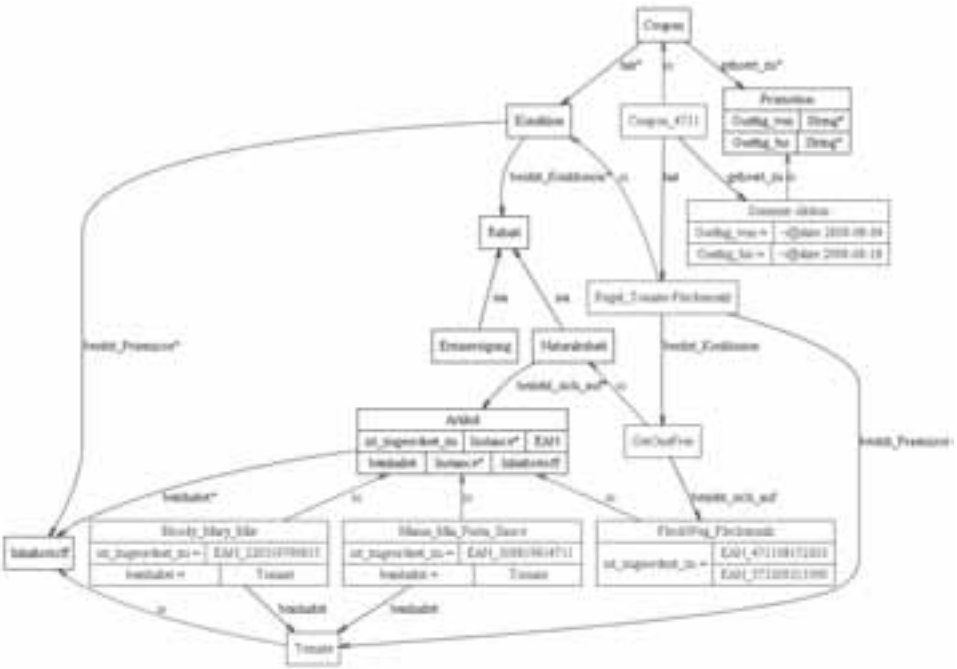

Abbildung 3: Exemplarische Ontologie (inkl. Instanzen) für Couponig-Aktionen im Einzelhandel

3.3 Mapping

Um die semantische Datenqualität eines bestimmten Informationssystems tatsächlich zu analysieren, muss ein Mapping zwischen der Domänenontologie und der zu betrachtenden Datenbank durchgeführt werden (Abbildung 4). Dieses Mapping verbindet die Semantik der Ontologie mit dem physischen Datenmodell und dem aktuellen Datenstand

der Datenbank. Für jedes Konzept (inkl. der Attribute) und alle Beziehungen innerhalb der Ontologie muss ein Gegenstück im vorliegenden Datenmodell identifiziert werden. NECIB und FREYTAG [NF05] haben in diesem Zusammenhang drei Typen von Mappings identifiziert: Zunächst Mappings zwischen Ontologiekonzepten (inkl. Attributen) und Datenbankrelationen (d. h. Tabellen und Views). Das Ontologiekonzept „Artikel“ könnte z. B. auf die Datenbankrelation „Item“ abgebildet werden. Desweiteren kann ein Mapping von Ontologiebeziehungen auf Datenbankrelationen erfolgen. Zum Beispiel könnte die Ontologiebeziehung „beinhaltet“ auf die Datenbankrelation „Rezept“ abgebildet werden, die Inhaltsstoffe von „items“ enthält. Dieser Link könnte eine Relation (im Falle einer many-to-many-Kardinalität) oder auch nur eine Foreign-Key-Bedingung (im Fall einer one-to-many-Kardinalität) sein. Abschließend können Mappings zwischen Ontologiekonzepten (inkl. Attributen) oder deren Instanzen und Datenbankattributwerten definiert werden. Zum Beispiel könnte die Ontologieinstanz „Tomate“ zum Wert „true“ des Attributes „tomatenhaltig“ in der Datenbankrelation „items“ gemappt werden.

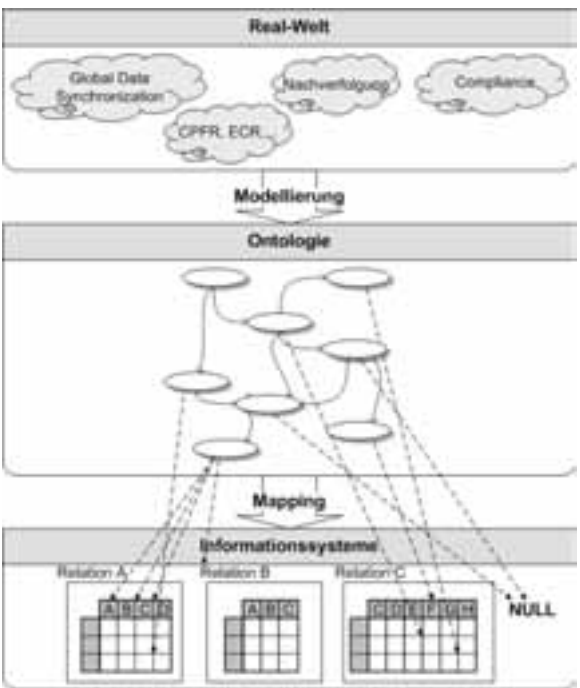

Abbildung 4: Präsentierter Ansatz zum semantischen Datenqualitätsmanagement

In einigen Fällen wird es nicht möglich sein, eine Entsprechung für ein Ontologiekonstrukt im Datenbankschema sofort zu finden. Dies kann der der Fall sein, wenn ein Phänomen der realen Welt nicht im Informationssystem repräsentiert ist. Zum Beispiel besitzen Standard-ERP-Systeme nicht alle Relationen und Attribute, die im Handel benötigt werden (z. B. Promotionen oder Coupons). In diesem Fall liegt ein Defekt des Datenmodells (und nicht des Datenbestandes) vor, welcher nur durch Modifizierung oder Erweiterung des Datenmodells behoben werden kann [Sh99]. Zur schließlichen Ausführung existieren Konzepte zur Erweiterung der SQL-Funktionalität [Prud06].

Das Ergebnis der Mapping-Phase ist eine Mapping-Tabelle, die auf der einen Seite Ontologiekonstrukte enthält und andererseits Relationen und Attributwerte des Datenbankschemas ausweist.

3.4 Anwendung

Die letzte Phase hat die resultierende Qualitätsbewertung zum Gegenstand, die den Entwurf von semantischen Abfragen, die Ausführung dieser Abfragen und die Analyse der Abfrageergebnisse beinhaltet. Aufgrund der vorangegangenen Konstruktion einer Domäneontologie und des Mappings zwischen Ontologiekonstrukten und dem physischen Datenmodell, können Analysten Abfragen auf einem semantischen Level entwerfen ohne sich dabei mit Details des physischen Datenbankmodells zu beschäftigen. Diese semantischen Abfragen werden dann basierend auf der Mapping-Tabelle in SQL-Abfragen übersetzt. Nach Ausführung der Abfragen kann beantwortet werden, ob alle Attribute und Beziehungen, die in einem spezifischen Kontext, wie z. B. dem Couponing, benötigt werden, gefüllt sind. Im Gegensatz zur Beantwortung solcher Fragestellungen durch reine SQL-Abfragen ohne übergeordnete Semantik erlaubt ein ontologiebasierter Ansatz zudem die Überwindung von typischen Heterogenitätsproblemen, z. B. inkonsistente Benennung von Relationen, Attributen und Attributwerten.
Abbildung 5 zeigt einen Screenshot eines ersten Prototyps, der den präsentierten Ansatz zur semantischen Bewertung von Stammdatenqualität im Bereich Couponing mit einbezieht. Die grünen Balken auf der rechten Seite des Screenshots repräsentieren den Grad der Übereinstimmung zwischen der Ontologie und dem physischen Datenmodell und Datenbestand. Das Qualitätsmaß in diesem ersten Prototyp ist relativ einfach gehalten und baut auf einem linearen Modell auf, das Inkonsistenzen und falsche Bezeichnungen berücksichtigt.

Abbildung 5: Beispielhafte Anwendung

4 Zusammenfassung und weiterer Forschungsbedarf

Fragestellungen bezüglich der syntaktischen Aspekte von Datenqualität sind in der Forschung weitgehend durchdrungen (auch wenn sie in der Praxis noch nicht zufriedenstellend gelöst sind). Praktische Ansätze bezüglich der semantischen Aspekte von Datenqualität existieren hingegen kaum.

Der präsentierte Ansatz soll zur Diskussion über Möglichkeiten der Verwendung von Ontologien für das semantisches Stammdatenmanagement beitragen. Der Ansatz adressiert eine Reihe von Fragen auf dem Gebiet der semantischen Datenqualität:

(A) Ebene des Datenmodells

- Die Modellierung von Domänen der Realwelt in einer Ontologie und das Mapping dieser Ontologie auf das physische Modell eines Informationssystems tragen zu einer verbesserten Abstimmung zwischen Entitäten der Realwelt und Datenbankstrukturen bei.
- Ontologiekonzepte können gleichzeitig auf verschiedene Informationssysteme (mit heterogenen Datenbankschemata) abgebildet werden. Dies trägt zu einer besseren Abstimmung von Datenbankstrukturen verteilter betrieblicher Informationssysteme bei.

(B) Ebene des Datenbestands

- Technisch weniger versierte Benutzer können semantische Abfragen, die von Implementierungsdetails abstrahieren, erstellen (Abfragegenerierung).
- Der Nutzen von Ontologien zur Erfassung der Semantik der vorliegenden Datenbank erlaubt die Analyse der Datenqualität ohne sich mit technischen Fragen zu beschäftigen (Analyseunterstützung).

Eine konkrete weiterführende Anwendung unsers Beitrags liegt beispielsweise im Auditing von Informationssystemen bezüglich der Eignung zur Unterstützung neuer Geschäftsmodelle oder der Erfüllung der Anforderungen neuer gesetzlicher Regelungen. Durch die Modellierung von Konzepten, die heute vielleicht noch nicht benötigt werden, aber in Zukunft von Interesse sein werden, können eventuell in der Zukunft auftretende Datendefizite frühzeitig erkannt werden.

Beschränkungen

Die Ergebnisse in dieser Arbeit basieren auf einer begrenzten Anzahl von Fallstudien und Experteninterviews. Zudem erlaubt der interpretivistische Forschungsansatz keine allgemeingültigen Aussagen. Wir verstehen unseren Ansatz deshalb nur als Startpunkt auf dem Weg zu einem semantischen Datenqualitätsmanagement.

Zukünftige Forschung

Wir schlagen vor, das in diesem Artikel eingeführte Vorgehensmodell und das Werkzeug bei der Durchführung zukünftiger Forschung auf dem Gebiet der Datenqualität zu

verwenden. Außerdem sollte das Konzept zum Zwecke der Evaluation und Weiterentwicklung auf weitere Domänen ausgeweitet werden. Wir glauben, dass semantischen Aspekten des Datenqualitätsmanagement in Zukunft sowohl in Theorie als auch in der Praxis eine zunehmend wichtige Rolle zukommen wird.

5 Danksagung

Dieser Beitrag wurde durch die Förderung des BMBF-Projektes „ManKIP“ (Management kreativitätsintensiver Prozesse, Förderkennzeichen 01FM07061) im Rahmen des Förderprogramms „Hightech-Strategie für die moderne Arbeitswelt“ ermöglicht. Wir danken an dieser Stelle dem Projektträger Deutsches Zentrum Luft- und Raumfahrt (DLR) für die Unterstützung.

6 Literaturverzeichnis

[Ag05] Agosta, L.: Trends in Data Quality. In: DM Review, 15 (2) 2005, S. 34-35.

[BD07] Berson, A.; Dubov, L.: Master Data Management and Customer Data Integration for a Global Enterprise. McGraw-Hill, New York, 2007.

[BGW08] Becker, J.; Glaser, J.; Winkelmann, A.: Auszug aus der aktuellen IT-Verbundgruppenstudie. In: Retail Technology Journal (03) 2008.

[BJ07a] Becker, J.; Janiesch, C.; Pfeiffer, D.: Context-Based Modelling. In: Proceedings of the 11th Pacific Asia Conference on Information Systems (PACIS 2007), Auckland, New Zealand, 2007; S.143.

[BJ07b] Becker, J.; Janiesch, C.; Pfeiffer, D.: Reuse Mechanisms in Situational Method Engineering. In: Proceedings of the IFIP WG 8.1 Working Conference on Situational Method Engineering, Geneva, Switzerland, 2007; S. 79-93.

[BJ07c] Becker, J.; Janiesch, C.; Pfeiffer, D.: Towards More Reuse in Conceptual Modeling. In: Proceedings of the 19th International Conference on Advanced Information Systems Engineering (CaiSE 2007), Trondheim, Norway, 2007; S. 81-84.

[BW08] Becker, J.; Winkelmann, A.: Handelscontrolling. Berlin, Heidelberg, New York, 2008.

[BUV01] Becker, J.; Uhr, W.; Vering, O.: Retail Information Systems Based on SAP Products. Springer, Berlin u.a., 2001.

[Ca04] Capgemini: Internal Data Alignment: Learning from Best Practices. Results of an Internal Data Alignment Survey, 2004.

[CFP03] Cappiello, C.; Francalanci, C.; Pernici, B.: Time-Related Factors of Data Quality in Multichannel Information Systems. In: Journal of Management Information Systems, 20 (3) 2003; S. 71-91.

[En99] English, L.P.: Improving Data Warehouse and Business Information Quality. Wiley, New York 1999.

[GF95] Gruninger, M.; Fox, M.S.: Methodology for the Design and Evaluation of Ontologies, Montreal, 1995.

[Gr93] Gruber, T.: A Translation Approach to Portable Ontology Specifications. In: Knowledge Management Acquisition, 5 (2) 1993; S. 199-220.

[Gr05] Griffin, J.: The Master Data Challenge. In: DM Review, 15 (5) 2005; S. 85.

[Gr08] Gruber, T.: Ontology. In (Liu, L.; Tamer Öszu, M. Hrsg.): Encyclopedia of Database Systems, Springer, 2008.

[HKL95] Hirschheim, R.; Klein, H.K.; Lyytinen, K.: Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations. Cambridge University Press, Cambridge, MA, 1995.

[HLW99] Huang, K.-T.; Lee, Y.W.; Wang, R.Y.: Quality information and knowledge, Upper Saddle River, NJ, 1999.

[HMP04] Hevner, A.R.; March, S.T.; Park, J.; Ram, S.: Design Science in Information Systems Research. In: MIS Quarterly, 28 (1) 2004, S. 75-105.

[Ja91] Jackson, M.C.: Systems Methodology for the Management Sciences. Plenum, New York, 1991.

[KPM04] Kahn, B.; Pierce, E.; Melkas, H.: IQ Research Directions. In (Chengalur-Smith, I.; Raschid, L.; Long, J.; Seko, C. Hrsg.): Proceedings of the 9th International Conference on Information Quality, Cambridge, MA, 2004; S. 326-332.

[Ku04] Kuipers, P.: Data in Dire Need of a Spring Clear. In: Elsevier Food International, 3, S. 74-79.

[Lo05] Loshin, D.: Master Data Standards and Data Exchange. In: DM Review, 15 (8) 2005; S. 72-77.

[LSK02] Lee, Y.W.; Strong, D.M.; Kahn, B.K.; Wang, R.Y.: AIMQ: A Methodology for Information Quality Assessment. In: Information & Management, 40 (2) 2002; S. 133-146.

[Mi97] Midgley, G.: Mixing methods: Developing Systemic Intervention. In (Mingers, J.; Gill, A., Hrsg.): Multimethodology: The Theory and Practice of Combining Management Science Methodologies, Wiley, Chichester, 1997.

[NF05] Necib, C.B.; Freytag, J.-C.: Query Processing Using Ontologies. In: Proceedings of the Proceedings of the 17th Conference on Advanced Information Systems Engineering (CAISE'05), Porto, Portugal, 2005.

[NM08] Noy, N.F.; McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. In Stanford, CA, 2008.

[PTR07] Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. In: MIS Quarterly, 24 (3) 2007, S. 45-77.

[Pi04] Pierce, E.M.: Assessing Data Quality with Control Matrices. In: Communications of the ACM, 47 (2) 2004; S. 82-86.

[PLW02] Pipino, L.L.; Lee, Y.W.; Wang, R.Y.: Data Quality Assessment. In: Communications of the ACM, 45 (4) 2002; S. 211-218.

[PS05] Price, R.; Shanks, G.: A Semiotic Information Quality Framework: Development and Comparative Analysis. In: Journal of Information Technology, 20 (2) 2005; S. 88-102.

[Pr 06] Prud'hommeux, Eric: SPASQL: SPARQL Support in MySQL. In: Proceedings of the XTech 2006, Amsterdam, The Netherlands, 2006.

[Re96] Redman, T.C.: Data Quality in the Information Age, Artech House, Boston, MA.

[Sc91] Schecter, D.: Critical Systems Thinking in the 1980s: A Connective Summary. In (Flood, R.L.; Jackson, M.C., Hrsg.): Critical Systems Thinking: Directed Readings, Wiley, Chichester, 1991.

[Sh99] Shanks, G.: Semiotic Approach to Understanding Represntation in Information Systems. In: Proceedings of the Proceedings of the Information Systems Foundations Workshop – Ontology, Semiotics and Practice, 1999.

[So00] Sowa, J.F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole, Pacifc Grove, CA, 2000.

[UG96] Ushold, M.; Gruninger, M.: Ontologies: Principles, methods and applications. In: Knowledge Engineering Review, 11 (2) 1996.

[Us96] Ushold, M.: Building Ontologies: Towards a Unified Methodology. In: Proceedings of the Proceedings of Expert Systems 1996, the 16th Annual Conference of the British Computer Society Specialist Goup on Expert Systems, Cambridge, 1996.

[Wa98] Wang, R.Y.: A Product Perspective on Total Data Quality Management. In: Communications of the ACM, 41 (2) 1998; S. 58-65.

[We97] Weber, R.: Ontological Foundations of Information Systems. Coopers and Lybrand, Melbourne, 1997.

[Wi06] Winkelmann, A.: Integrated Couponing. A Process-Based Framework for In-Store Coupon Promotion Handling in Retail. Logos, Berlin, 2006.

[WW96] Wand, Y.; Wang, R.Y.: Anchoring Data Quality Dimensions in Ontological Foundation. In: Communications of the ACM, 39 (11) 1996; S. 86-95.

[WYP98] Wang, R.Y.; Yang, W.L.; Pipino, L.L.; Strong, D.M.: Manage Your Information as a Product. In: Sloan Management Review, 39 (4) 1998; S. 95-105.