in: Proceedings der 5. Paderborner Frühjahrstagung "Innovationen im E-Business", 10.04.2003, Paderborn, S. 159-168.

Behandlung von Informationsdefiziten und -verlusten bei der Transformation von XML-Geschäftsdaten

Michael Beul, Christian Bittscheidt, Jörg Leukel, Thorsten Spies
Universität Essen

Inhaltsverzeichnis

1. Einleitung

Im Zuge der schnellen Etablierung von XML (Extensible Markup Language) als Metasprache für die Definition von Datenformaten sind zahllose Sprachen für Geschäftsdokumente entstanden, die von proprietären Lösungen bis zu internationalen Standards reichen. Für den zwischenbetrieblichen Geschäftsdatenaustausch und die engere Kopplung von Anwendungssystemen ist XML damit zu einer Basistechnologie geworden. Angesichts des Fehlens von gefestigten, branchenübergreifenden Standards, die sich bereits eine langfristige Verbreitung gesichert hätten, stehen viele Unternehmen vor der Aufgabe, sowohl beschaffungs- als auch vertriebsseitig XML-Daten in unterschiedlichsten Formaten zu verarbeiten. Für elektronische Marktplätze ist die effiziente und kundenorientierte Bewältigung dieser Aufgabe sogar zu einem kritischen Erfolgsfaktor geworden.

Die skizzierte Situation erfordert sowohl technische als auch fachlich-inhaltliche Fähigkeiten, Geschäftsdaten zu verarbeiten und von einem Quellformat in eines oder mehrere Zielformate zu überführen. Der Beitrag adressiert diese Problematik, indem ein fortgeschrittenes Transformationskonzept entwickelt, implementiert und anhand einer Feldstudie aus dem B2B-Bereich evaluiert wird. Grundlagen sind eine Klassifikation von Transformationsfällen und die explizite Berücksichtigung des divergierenden inhaltlichen Umfanges verschiedener XML-Formate, der in der Regel Informationsdefizite und –verluste zur Folge hat.

2. XML-Geschäftsdaten

Die hohe Flexibilität und die individuelle Erweiterbarkeit von XML verleiteten viele Unternehmen und Organisationen dazu, eigene Sprachen und Strukturen für Geschäftsdaten und -dokumente festzulegen und diese einzusetzen. Auch werden vielfach standardisierte Geschäftssprachen bilateral modifiziert und erweitert, um unternehmensspezifische Anforderungen zu erfüllen. In der Folge wird die Standardsprache verletzt und es entstehen Sprachdialekte. Daher ist die ursprüngliche Zielsetzung, Anwendungssysteme enger miteinander zu koppeln, nicht bereits mit der Verwendung von XML erreicht. Vielmehr wird das Standardisierungs- und Integrationsproblem auf die höheren Kommunikationsebenen des Vokabulars und der Nachrichtentypen verlagert.

Damit nun Informationssysteme und menschliche Aufgabenträger beider Parteien die Daten und Dokumente verarbeiten können, müssen diese einerseits inhaltlich erfasst und verstanden werden und andererseits in Übereinstimmung mit den betrieblichen Datenstrukturen und –konzepten gebracht werden. Hierzu weisen XML-basierte Geschäftsdaten im Gegensatz zu anderen Datenformaten (z.B. Komma-separierte Daten, CSV; EDI-basierte Formate) eine Reihe von Vorteilen auf, die insgesamt die zwischenbetriebliche Verarbeitung erleichtern.

Als ein Vorteil von XML wird oftmals angeführt, dass XML-Daten selbstbeschreibend sind, indem alle Datenwerte durch die zugehörigen Datenelementbezeichner ausgezeichnet werden. Dadurch erschließt sich die Bedeutung und die Struktur der Daten für den Nutzer, insbesondere für den Menschen, besser. Anzumerken ist, dass auch CSV-Dateien die Datenelementbezeichner zu Beginn nennen. Die Wertauszeichnung findet sich bei EDI ebenfalls wieder, die Bezeichner sind jedoch codiert und daher höchstens rudimentär selbstbeschreibend. Dagegen kommt der formalen Spezifikation des Datenformates eine wesentlich größere Bedeutung zu.

Die formale Spezifikation unterstützt die Erstellung und Verarbeitung von XML-Daten, indem die Syntax der Daten einfach und dennoch weitgehend beschrieben werden kann. Durch Vergleich von XML-Dokumenten mit der zugehörigen formalen Spezifikation kann die Gültigkeit der Daten und damit die Konformität zum definierten Format überprüft werden. So ist sichergestellt, dass keine syntaktisch falschen Daten verarbeitet werden (z.B. falsche Datentypen, falsche Elementreihenfolgen, fehlende oder nicht definierte Datenelemente, usw.). Diese Möglichkeit ist bei CSV- und EDI-Formaten nicht gegeben und muss gegebenenfalls durch individuelle Maßnahmen sichergestellt werden. Zweitens wird die Handhabung von XML-Daten durch hierarchische und standardisierte Zugriffsmodelle erleichtert (z.B. XPath, DOM, SAX), so dass im Gegensatz zu EDI die Unterstützung durch Softwarewerkzeuge und Programmierbibliotheken nahezu umfassend ist.

3. Transformationskonzepte

Die Verarbeitung von XML-Geschäftsdaten und deren Überführung von einem Quellformat in ein oder mehrere Zielformate erfordert sowohl technische als auch fachlich-inhaltliche Fähigkeiten. Zum einen benötigt man entsprechende Transformationstechniken, die die Möglichkeit bieten, generisch XML-Dokumente zu verarbeiten. Andererseits müssen Dokumente semantisch erfasst werden, d.h. die gespeicherten Daten müssen inhaltlich identifiziert werden. Erst wenn die Bedeutung der Elemente und ihrer enthaltenen Daten festgelegt ist, besteht die Möglichkeit, mit Hilfe entsprechender Transformationsverfahren, die Überführung eines Formats in ein anderes zu vollziehen. Datentransformationen sind jedoch keine neue Erscheinung der XML-Welt, sondern bilden einen Kernbereich der Forschung im Bereich Datenmodelle und Datenbanken. Einen Schwerpunkt stellen Arbeiten zur Transformationen unterschiedlicher, insbesondere relationaler Schemata dar.

Der erste Baustein von Transformationskonzepten befasst mit der Definition so genannter Mappings zwischen dem Quell- und Zielformat. Der Begriff Mapping bezeichnet die Zuordnung von Datenelementen eines Quelldokumentes zu Datenelementen eines Zieldokumentes. Die Erarbeitung der Mapping-Definitionen basiert auf der Analyse der formalen und deskriptiven Formatspezifikationen, soweit diese verfügbar und hinreichend sind.

Zunächst muss für jedes Element übergeprüft werden, welche Informationen es enthält und ob diese in dem Zieldokument relevant sind. Die se-

mantische Zuordnung erfordert genaue Kenntnisse der betroffenen Dokumenttypen. Zu beachten sind hierbei besonders die unterschiedlichen Interpretationsformen einzelner Bereiche. Fehlinterpretationen führen in den meisten Fällen zu einer Verfälschung der Ergebnisse und damit zu einer Qualitätsminderung des Zieldokumentes. Zu vielen Formaten kann zwar die zugehörige Spezifikation Hilfe bei der Klärung von Elementinhalten bieten, jedoch ist die Bedeutung oft nur mit fachlichen Hintergrundwissen eruierbar, d.h. es ist Wissen über die jeweilige Domäne (z.B. E-Procurement, Logistik) notwendig.

Nachdem die semantisch zueinander passenden Elemente identifiziert worden sind, muss die Syntax der Elemente überprüft und gegebenenfalls angepasst werden. Dies umfasst die Überprüfung und den Vergleich der Datentypen der zu überführenden Elemente, die Anpassung der Datengranularität sowie die formale Beschreibung der Zuordnung mit Hilfe einer XML-verarbeitenden Sprache.

Weichen Datentypen voneinander ab, so müssen Regeln für die Anpassung definiert werden. Beispielsweise werden Preise in dem einen Format als beliebige Fließkommazahlen dargestellt, während sie im anderen Format eine feste Anzahl von Nachkommastellen besitzen, die zudem durch einen Punkt von den Vorkommastellen getrennt werden. Des Weiteren besteht die Möglichkeit, dass (verschiedene) Code-Standards verwendet werden und entsprechend angepasst werden müssen. Beispielsweise kann bei einem Dokument die Angabe des Landes nach den in ISO 3166-1 definierten Codes erfolgen (Deutschland = DE, Österreich = AT,…), in dem anderen wird die Angabe des vollständigen Landesnamen gefordert. Allgemein sind Unterschiede in der Repräsentation ansonsten gleicher Inhalte zu überbrücken.

Dokumente unterschiedlicher Herkunft unterscheiden sich meistens in der Struktur der gespeicherten Informationen. Es können vier Mapping-Grundtypen identifiziert werden, die beschreiben, wie viele Datenelemente am Mapping beteiligt sind. Die Typisierung basiert also auf der Kardinalität der Elemente. Wird der Inhalt eines Datenelementes in genau ein anderes Element überführt, spricht man hierbei von einem 1:1-Mapping. Es kommt aber auch vor, dass mehrere Elemente aus einem Quellformat in ein Element des Zielformats konkateniert werden müssen. Dieses bezeichnet man als N:1-Mapping. Diese zwei Mappingvarianten sind mit relativ geringem Aufwand durchzuführen. Ist die Granularität invertiert, spricht man von einem 1:N-Mapping. Die Implementierung dieser Mapping-Variante ist komplexer und schwieriger, da die Daten eines Elementes verteilt werden müssen. Dazu müssen Regeln definiert werden, die beschreiben, nach welchen Kriterien, Merkmalen oder Regeln die Daten separiert werden. Der vierte Fall ist das N:M-Mapping, bei dem in beiden Formaten die Informationen auf mehrere Elemente unterschiedlich aufgeteilt sind. Der Aufwand für ein solches Mapping hängt erstens von der inhaltlichen Überschneidung und zweitens von der hierarchischen Positionierung der Elemente in den Dokumentstrukturen ab. Beispielsweise kann es vorkommen, dass die Daten von zwei Elementen des Quellformats jeweils kreuzweise auf zwei/drei Elemente eines Zielformats aufgeteilt werden müssen.

Um derartige Regeln zu definieren und Transformationen durchführen zu lassen, steht die zur XML-Standardfamilie gehörende Extensible Stylesheet Language (XSL) zur Verfügung, wobei für Datentansformationen nur die

Untermenge XSLT (XSL Transformations) von Relevanz ist. Mit diesen Sprachen lassen sich Transformationsregeln definieren. Dabei entspricht jede Regel einem Muster im Quelldokument. Die Regeln werden durch XSLT-Prozessoren abgearbeitet, die somit die Transformation ausführen und als Ergebnis die Zieldokumente erstellen.

4. Informationsdefizite und -verluste

Voraussetzung für die verlustfreie Transformation von XML-Geschäftsdaten ist, dass Quell- und Zielformat inhaltlich übereinstimmen, d.h. alle Inhalte, die gemäß der Formatdefinition im Quellformat repräsentiert werden, finden sich unter Verwendung einer gleichen oder verschiedenen Syntax im Zielformat wieder (Omelayenko/Fensel 2001) (Wüstner/Hotzel/Buxmann 2002). In diesem Fall reicht die einmalige Mapping-Definition aus, um Dokumenttransformationen automatisch ohne Benutzereingriffe ausführen zu lassen. Jedoch ist diese Anforderung angesichts der hohen Anzahl von XML E-Business-Standards sehr restriktiv und schränkt den Anwendungsbereich deutlich ein. Deshalb ist es notwendig, auch für die weitergehenden Fälle, in denen sich Quell- und Zielformat hinsichtlich ihrer inhaltlichen Überdeckung unterscheiden, Transformationskonzepte zu entwickeln. Ein Ansatzpunkt ist, die Formate differenziert nach ihrem Informationsgehalt gegenüberzustellen. Verallgemeinert man dieses Vorgehen, so entstehen, wie in der Abbildung 1 zusammengefasst, insgesamt neun Fälle. Differenzierungsmerkmal ist jeweils, ob es sich bei dem relevanten Informationsgehalt, der sich auf XML-Datenelemente zurückführen lässt, um Pflicht-, optionale oder nicht vorhandene Datenelemente handelt.

		Zielformat		
		Pflicht	**Optional**	**Nicht vorhanden**
Quellformat	**Pflicht**	Mapping	Mapping	Informationsverlust
	Optional	Informationsdefizit	Mapping	Informationsverlust
	Nicht vorhanden	Informationsdefizit	-	-

Abb. 1: Typisierung von Datentransformationen nach dem Informationsgehalt

Drei Aussagen lassen sich aus der entstehenden Typisierung von Datentransformationen nach dem Informationsgehalt ableiten. Erstens werden die Fälle, in denen im Quellformat enthaltene Informationen ebenfalls Bestandteil des Zielformates sind, vollständig durch Mapping-Definitionen abgedeckt. Zweitens entsteht dann ein Informationsverlust, wenn Informationen des Quellformates nicht im Zielformat wiedergegeben werden können. Besonders gravierend ist der Verlust dann, wenn es sich um Pflichtelemente handelt. Hier ist eine Konvertierung überhaupt nicht sinnvoll, da bereits Basisinformationen verloren gehen. Weniger schwer wiegt der Verlust bei optionalen Informationen, da wenigstens die Pflichtbestandteile des Quellformates verlustfrei transformiert werden können. Die Konvertierung

muss akzeptieren, dass das Zielformat weniger mächtig als das Quellformat ist.

Drittens liegt dann ein Informationsdefizit vor, wenn Pflichtelemente des Zielformates nur optional oder überhaupt nicht im Quellformat vorhanden sind. Um eine Dokumentkonvertierung dennoch zu ermöglichen, ist es notwendig, die fehlenden Informationen in den Transformationsprozess zu integrieren. Dazu ist in diesen einzugreifen, indem die benötigten Informationen manuell hinzugefügt werden.

5. Feldstudie zur Transformation von XML-Katalogdaten

5.1. XML-Katalogstandards

Das beschriebene Transformationskonzept, das sich an der konventionellen Typisierung nach der Kardinalität orientiert und zusätzlich Informationsdefizite und –verluste berücksichtigt, ist anhand einer realen Transformationsproblematik implementiert und evaluiert worden. Gegenstand der Transformationen sind elektronische Produktkataloge, die in dem Format BMEcat vorliegen und in die drei Zielformate cXML, eCX und xCBL zu konvertieren sind. Bei den vier Formaten handelt es sich um die für Katalogdaten wichtigsten XML-Standards. Die Domäne Produktkataloge bietet sich für Transformationsuntersuchungen an, da Katalogdaten strukturell und inhaltlich komplexer als andere Geschäftsdokumente sind (Leukel/Schmitz/Dorloff 2002). Insbesondere ist zu erwarten, dass Informationsdefizite und –verluste häufig anzutreffen sind.

Der Katalogstandard **BMEcat** definiert ein umfangreiches Vokabular, auf dessen Basis drei unterschiedliche Dokumenttypen für vollständige Kataloge sowie Artikel- und Preisaktualisierungen zusammengestellt werden (Schmitz/Kelkar/Pastoors 2001). Jedes BMEcat-Dokument besteht aus einem Kopfbereich, der Metainformationen über die Katalogdaten (u.a. Version, Lieferant, Kunde) enthält, und einem Transaktionsbereich, der in Abhängigkeit von der Transaktion die Katalogdaten im engeren Sinne aufnimmt. BMEcat unterstützt Kataloggruppen- und Klassifikationssysteme.

Als Anbieter von Software zur Errichtung von B2B-Marktplätzen hat Ariba den Standard **cXML** entwickelt (Ariba 2001). cXML sieht für Kataloge die Bereiche Supplier (Stammdaten über den Lieferanten), Index (Artikeldaten) und Contract (flexible Daten) vor. Weder Katalogruppen- noch Klassifikationssysteme werden unterstützt.

Als Anbieter von E-Procurement-Lösungen verwendet Requisite den Standard ***eCX*** (Requisite Technology 2001). Die Bereiche Admin (Allgemeine Daten zum Katalog), Schema (Struktur des Kataloges, Kataloggruppen) und Date (Artikeldaten) beschreiben einen Katalog im eCX-Format. Unterstützt werden Kataloggruppen, jedoch keine Klassifikationssysteme.

Schließlich setzt CommerceOne den eigenen Standard ***xCBL*** ein. Ein xCBL-konformer Katalog hat folgenden Aufbau: CatalogHeader (Infos über

den Katalog, Lieferanten, Käufer,…), CatalogSchema (Struktur des Kataloges) und CatalogData (Artikeldaten) (CommerceOne 2001).

5.2. Transformation von BMEcat nach cXML, eCX und xCBL

Die Mapping-Definition vollzieht sich in zwei Schritten. Zunächst erfolgt die Typisierung, so dass die Kardinalität und ein gegebenenfalls zu berücksichtigendes Informationsdefizit bzw. –verlust feststehen. Anschließend werden Transformationsanweisungen in XSLT formuliert. Eine quantitative Auswertung der Mapping-Definitionen erlaubt Rückschlüsse auf die Transformationskomplexität sowie auf den Abdeckungsgrad zwischen Quell- und Zielformat. Die Abbildung 2 stellt die Anzahl der Mappings differenziert nach den vier Kardinalitäten dar. Im Ergebnis zeigt sich, dass 1:1-Mappings dominieren; ihr Anteil liegt zwischen 87% und 95%. Hinsichtlich der inhaltlichen Reichweite lässt sich anhand der Summe der Mappings je Standard erkennen, daß xCBL die grösste Übereinstimmung mit BMEcat aufweist, wohingegen eCX nur eine kleine Untermenge der BMEcat-Katalogdaten wiedergeben kann.

	cXML	eCX	xCBL
1:1	48	21	64
1:N	6	1	2
N:1	1	0	0
N:M	0	0	3

Abb. 2: Anzahl der Datentransformationen differenziert nach Mapping-Kardinalität

Interessanter erscheint jedoch eine Gegenüberstellung der festgestellten Informationsdefizite, die schließlich in der Transformationsplattform separat behandelt werden müssen. Die Abbildung 3 nennt jene Datenelemente, zu denen in den drei Zielformaten keine äquivalenten BMEcat-Datenelemente existieren. Es handelt sich dabei sowohl um elementare Datenelemente, die Datenwerte aufnehmen, als auch um Container, die weitere Datenelemente enthalten. Daher kann nicht allein aufgrund der Anzahl der genannten Datenelemente auf den Umfang oder den Grad der Informationsdefizite geschlossen werden. Für Katalogdaten sind selbstverständlich die Informationsdefizite kritisch, die produktbezogene Daten betreffen, da sie unter Umständen für jeden Artikel auftreten und jeweils nachgepflegt werden müssten.

	eCX	cXML	xCBL
Informations-defizite	Name [Catalog] Key [Item/Owner]	SupplierID Name Street City Country TelephoneNumber OrderMethods UnitOfMeasure LeadTime	Identifier ListofIdentifiers ContactName ContactNumberValue SchemaName

Abb. 3: Auswertung Informationsdefizite

Aus einem gültigen, minimal ausgestatteten BMEcat-Katalog ist es nahezu möglich, alle Muss-Felder eines eCX-Katalogs zu füllen. Informationsdefizite treten bei einem Mapping aus eCX-Sicht lediglich bezüglich des Katalognamens und bei der Klassenzuordnung auf. Letzter Fall ist jedoch auf eine Modellierungsschwäche in BMEcat zurückzuführen. Die Anzahl der Informationsdefizite bei der Konvertierung nach cXML ist deutlich höher. Dies gilt sowohl für Katalogmetadaten als auch für artikelbezogene Daten. Die Ursache liegt in einer detaillierteren Modellierung und in einer größeren Anzahl von Muss-Feldern. Die Transformation nach xCBL deckt aufgrund der zuvor festgestellten inhaltlichen Überdeckung erwartungsgemäß weniger Informationsdefizite auf, die sich zudem nicht auf die kritischen Artikeldaten beziehen.

5.3. Plattform für Katalogtransformationen

Zur vollständigen Evaluation des Transformationskonzeptes dient ein konfigurierbarer Web-Prototyp für Verarbeitung von Katalogdokumenten und die Ausführung der XSLT-Anweisungen. Die Anwendung befindet sich zurzeit in der Testphase, um anhand von Realdaten die Richtigkeit der Transformationen zu überprüfen. Sie ist unter www.crosscat.de zugreifbar. Ein solcher Dienst kann als Erweiterung von elektronischen Marktplätzen angesehen werden, um eingehende Katalogdaten in benötigte Zielformate zu konvertieren. Außerdem eignet sich dieser als eigenständiger Mehrwertdienst, der es Katalogerstellern ermöglicht, auf der Basis von BMEcat weitere Zielformate zu unterstützen.

Ziel war es zugleich, über Katalogdaten hinaus eine allgemeine Architektur für beliebige Transformationen von XML-Dokumenten zu schaffen. Daher erfolgt eine Trennung der domänenspezifischen Systemeigenschaften von den domänenunabhängigen Komponenten. Neben der Auslagerung der Mapping-Definitionen in physische XSLT-Skripte zählt dazu die Behandlung von Informationsdefiziten, die wiederum durch XML-basierte Konfigurationsdaten beschrieben werden.

Auf der Implementierungsseite nutzt die Plattform Standardtechnologien, die der Dreiteiligung in Präsentations,- Applikations- und Datenschicht folgen: Webserver, der die Benutzerinteraktion steuert (Apache Tomcat); Transformationsserver, der die XSLT-Skripte ausführt (Xalan); Datenbankserver für die Verwaltung von Benutzerprofilen, Konvertierungsaufträgen usw. (SQL Server).

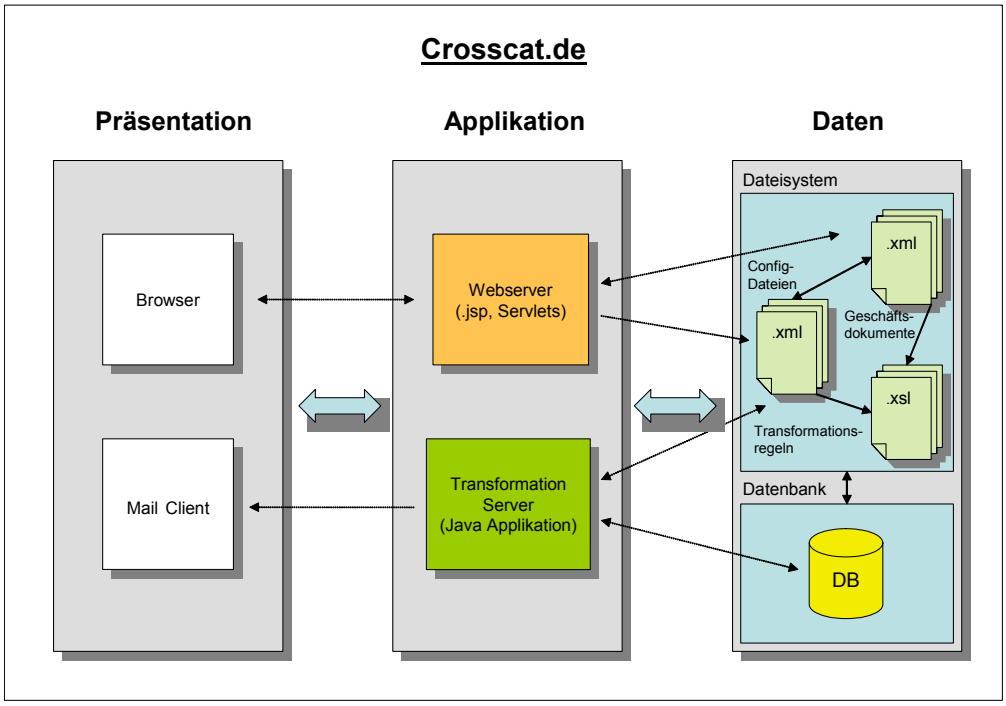

Abb. 3: 3-Schichten-Architektur der Transformationsplattform Crosscat.de

Literatur

Ariba, Inc. (Ariba 2001): cXML 1.2.007. URL: http://xml.cxml.org/current/cXML.zip, 2001.

CommerceOne, Inc. (CommerceOne 2001): XML Common Business Libary (xCBL), Version 3.5. URL: http://www.xcbl.org, 2001.

Leukel, Jörg; Schmitz, Volker; Dorloff, Frank-Dieter (Leukel/Schmitz/Dorloff 2002): Exchange of Catalog Data in B2B Relationships - Analysis and Improvement, in: Proceedings of IADIS International Conference WWW/Internet 2002 (ICWI 2002), Lissabon, Portugal, 13.-15.11.2002, S. 403-410.

Omelayenko, Boris; Fensel Dieter (Omelayenko/Fensel 2001): An Analysis of Integration Problems of XML-Based Catalogs for B2B Electronic Commerce, in: Proceedings of the 9th IFIP 2.6 Working Conference on Database Semantics (DS-9), Hong-Kong, 25.-28.04.2001, S. 232-246.

Requisite Technology (Requisite Technology 2000): Electronic Catalog XML (eCX) Specification, Version 2.0. URL: http://www.ecx-xml.org, 2000.

Schmitz, Volker; Kelkar, Oliver; Pastoors, Thorsten (Schmitz/Kelkar/Pastoors 2001): Spezifikation BMEcat, Version 1.2 URL: http://www.bmecat.org, 2001.

Wüstner, Erik; Hotzel, Thorsten; Buxmann, Peter (Wüstner/Hotzel/Buxmann 2002): Converting Business Documents: A Classification of Problems and Solutions using XML/XSLT, in: Proceedings of the 4th IEEE International Workshop on Advanced Issues of E-Commerce and Web-based Information Systems (WECWIS 2002), Newport Beach, California, USA, 26.-28.06.2002, S. 61-68.