Ein generativer Ansatz zur semantischen Beschreibung von Geschäftsdokumenten

Matthias Ferdinand
6ferdina@informatik.uni-hamburg.de
Universität Hamburg, Fachbereich Informatik
Verteilte Systeme und Informationssysteme

Betreuer der Arbeit: Prof. Dr. W. Lamersdorf und Dr. C. Eschenbach
Art der Arbeit: Diplomarbeit
GI-Fachbereich: TI

Zusammenfassung

Im folgenden wird ein allgemein anwendbarer Ansatz beschrieben, mit dem Dokumente und Dokumentbeschreibungen auf XML-Basis auf die semantische Ebene einer Ontologie angehoben werden können. Das Vorgehen besteht dabei im Kern aus zwei automatisierten Sprachabbildungen, die auf Technologien des Semantic Web zurückgreifen. Nach einer Beschreibung der Konzeption und Implementierung wird gezeigt, wie der Ansatz dazu beitragen kann, wesentliche Probleme bei der Vorbereitung und Ausführung XML-basierter Geschäftsprozesse zwischen Unternehmen (B2B) zu lösen.

1. Einführung

Elektronische Geschäftsprozesse zwischen Unternehmen (B2B) werden heute vor allem mit Hilfe XML-basierter Standards wie z. B. RosettaNet [1] durchgeführt. Sie kombinieren die Vorteile des Internet und der XML-Sprachfamilie und bieten so, im Vergleich zu älteren Technologien wie EDI [2], ein großes Potential, um Geschäftsdokumente kostengünstiger auszutauschen und flexibler zu gestalten. Die höhere Flexibilität dieser Standards ist zwar wünschenswert, hat aber auch zur Folge, daß die Unternehmen beim Aufbau neuer Geschäftsbeziehungen dazu gezwungen sind, die Spezifikationen der auszutauschenden Dokumente ihren eigenen Anforderungen anzupassen und untereinander abzustimmen. Dabei handelt es sich oft um einen langwierigen und teuren Prozeß. Von Nachteil ist in diesem Zusammenhang zudem, daß man mit den XML-Sprachen ausschließlich Vorgaben syntaktischer Natur ausdrücken kann und sich somit verschiedene Arten komplexer Bedingungen und Regeln nicht spezifizieren lassen.

Diese Aspekte stellen einen Problemkomplex im B2B-Bereich dar, für den im Rahmen dieser Diplomarbeit ein Lösungsansatz erarbeitet wird. Der Ansatz sieht vor, Dokumente und Dokumentbeschreibungen auf XML-Basis auf die Ebene einer Ontologie anzuheben, wobei auf Technologien des Semantic Web [3] zurückgegriffen wird. Im einzelnen wird dabei XML Schema [4] in die Web Ontology Language (OWL) [5] und XML in das Datenmodell des Resource Description Framework (RDF) [6] abgebildet, was zunächst konzeptionell erarbeitet und anschließend implementiert wird. Auf diesem Weg lassen sich die Möglichkeiten der semantischen Modellierung, die Ausdrucksmächtigkeit dieser Sprachen und die leistungsfähigen Inferenzmechanismen der Beschreibungslogik [7] auf die Dokumente anwenden. Darauf aufbauend wird demonstriert, wie sich semantische Vorgaben für Geschäftsdokumente mit Hilfe von Ontologien spezifizieren und überwachen lassen. Die praktische Umsetzung wird durch ein Inferenzsystem unterstützt.

Im nächsten Abschnitt dieses Abstracts sollen zunächst einige für den Themenkomplex zentrale Begriffe vorgestellt werden. Im dritten Teil wird die grundlegende Vorgehensweise bei der Lösung skizziert, die im Kern aus zwei Sprachabbildungen besteht. In Abschnitt 4 wird auf die allgemeinen und problembezogenen Anwendungsmöglichkeiten eingegangen. Der fünfte Teil faßt schließlich die Ergebnisse zusammen.

2. Grundlagen

XML Schema bildet die Basis für viele aktuelle B2B-Standards und spielt daher eine zentrale Rolle für diese Arbeit. Diese vom World Wide Web Consortium (W3C) entwickelte Sprache dient dazu, Syntax und Struktur für eine Menge von XML-Dokumenten zu definieren. Mit Hilfe von verschiedenen Arten von Typen und Regeln lassen sich mit dieser Sprache syntaktische Einschränkungen spezifizieren, die für eine Klasse von XML-Instanzen gelten sollen. Ontologien, Beschreibungslogik und das Semantic Web sind die Grundlagen für den gewählten Lösungsweg und werden im folgenden kurz charakterisiert.

Ontologien sind ein Begriff aus der Künstlichen Intelligenz und stellen – vereinfacht ausgedrückt – abstrakte, semantische Modelle eines Ausschnitts aus der realen Welt dar. Mit ihnen läßt sich das Wissen über ein Anwendungsgebiet auf formale Weise u.a. mit Hilfe von Konzepten, Eigenschaften, Beziehungen, Werterestriktionen und Regeln formal beschreiben, so daß es von Maschinen automatisch verarbeitet und kommuniziert werden kann [8].

Beschreibungslogik (Description Logics, DL) stellt eine Familie von speziellen Fragmenten der Prädikatenlogik erster Ordnung dar, die sich aufgrund ihrer Eigenschaften ideal zur Wissensrepräsentation eignen und die formale Grundlage für eine ganze Reihe von Sprachen zur Definition von Ontologien bilden, so auch für OWL. Insbesondere wird durch die formale Syntax und deklarative Semantik der Beschreibungslogiken ermöglicht, durch Schlußfolgerungen aus bestehendem Wissen neues Wissen abzuleiten (logische Inferenz). Implementiert wird dies durch sog. DL-Inferenzsysteme wie z.B. RACER [9].

Das Semantic Web stellt eine Vision für ein World Wide Web der zweiten Generation dar, bei dem die verfügbaren Informationen mit einer maschinenlesbaren Semantik versehen werden sollen. Im Mittelpunkt steht dabei das bereits erwähnte OWL, mit dem die Semantik von Web-Ressourcen auf eine standardisierte Weise beschrieben werden kann. RDF dient in diesem Kontext dazu, die einzelnen Instanzen der in einer OWL-Ontologie definierten Konzepte zu beschreiben.

3. Lösungsansatz und Implementierung

Der grundlegende Lösungsansatz für den in der Einführung skizzierten Problembereich besteht darin, sowohl XML Schema-Spezifikationen als auch XML-Dokumente von ihrer syntaktischen Ebene auf eine semantische Ebene anzuheben, indem man sie in eine Ontologie auf Beschreibungslogik-Basis abbildet. Ausgangspunkt hierfür ist die Annahme, daß sich letztere ideal dazu eignen, Dokumente auf einer semantischen, abstrakteren Ebene zu beschreiben.

Das Vorgehen besteht erstens darin, eine Abbildung von XML Schema zu OWL zu entwickeln. Zweitens werden auf der Ebene der Instanzen XML-Dokumente, die konform zu einem XML Schema sind, in RDF-Graphen abgebildet.

3.1. Abbildung von XML Schema zu OWL

XML Schema und OWL sind für zwei unterschiedliche Aufgabenbereiche konzipiert: XML Schema dient allgemein der syntaktischen Beschreibung und Einschränkung von XML-Dokumenten, während OWL die semantischen Zusammenhänge eines Wissensbereichs modelliert. Dennoch haben sie eine Gemeinsamkeit: Sie basieren beide auf objektorientierten Modellierungsprinzipien. Diese gemeinsame konzeptionelle Wurzel und die daraus resultierenden Ähnlichkeiten werden in der Arbeit dazu herangezogen, die Abbildung zwischen den beiden Sprachen zu motivieren, detailliert zu entwickeln und zu begründen.

Da ein XML Schema i.d.R. von mehreren Parteien als Referenz für die genaue syntaktische Beschreibung von XML-Dokumenten herangezogen wird, ist es unabdingbar, daß die darin enthaltenen Informationen exakt und möglichst verlustfrei abgebildet werden. Insbesondere müssen die grobe Struktur eines Schemas mit seinen Typen, ihren einzelnen Strukturelementen und die Beziehungen zwischen diesen Komponenten übertragen werden.

Auf die Einzelheiten der Abbildung kann hier aus Platzgründen nicht eingegangen werden. Sie wird jedoch in der Diplomarbeit sowohl in formaler als auch in syntaktischer Hinsicht exakt definiert und begründet. Es sei hervorgehoben, daß die Abbildung keinerlei Annahmen über Herkunft und Aufbau der Dokumente trifft. Sie kann also nicht nur auf den B2B-Bereich, sondern grundsätzlich auf alle XML Schema-Dokumente angewendet werden. Das Resultat des Abbildungsprozesses ist stets eine zur OWL-Spezifikation vollständig konforme Ontologie. Sie dient im Hinblick auf das weitere Vorgehen als eine Rohversion, die von einem Benutzer i.d.R. um die erforderlichen Konzepte, Einschränkungen, Regeln etc. ergänzt wird.

3.2. Abbildung von XML zu RDF

Auf der Ebene der Instanzen werden XML-Dokumente, die die im Rahmen eines B2B-Prozesses ausgetauschten Geschäftsdaten enthalten, in Instanzen der zuvor erzeugten OWL-Ontologie transformiert. Das Resultat ist stets ein RDF-Graph. Im folgenden wird zunächst auf die Unterschiede zwischen den beiden Sprachen eingegangen.

XML ist eine Sprache, die eine generische Syntax zur Speicherung und zum Austausch von Dokumenten mit Hilfe einer Baumstruktur definiert [10]. Zwar besitzt RDF eine XML-basierte Syntax, die beiden Sprachen dienen jedoch unterschiedlichen Zwecken und wurden unabhängig voneinander entwickelt. Dies führte zu unterschiedlichen Datenmodellen: XML basiert auf einem Baummodell, bei dem nur die Knoten beschriftet sind und die von einem Knoten ausgehenden Kanten eine Ordnung aufweisen. Es weist eine enge Verbindung zu semistrukturierten Daten auf. Grundlage von RDF ist hingegen ein gerichteter Graph, bei dem die Kanten beschriftet und ungeordnet sind.

Um die Lücke zwischen diesen beiden Formen der Datenrepräsentation zu schließen, wurde in der Arbeit ein Algorithmus zur Abbildung von XML zu RDF entwickelt, der auf der rekursiven Traversierung eines XML-Baums und der Analyse und Transformation der einzelnen Komponenten basiert. Die Besonderheit dabei ist, daß die Typinformationen des dazugehörigen XML Schemas (welches zuvor zu OWL abgebildet worden sein muß) mit einbezogen werden. So können die einzelnen Teile des resultierenden RDF-Graphs den Konzepten der OWL-Ontologie zugeordnet werden, sie sind also typisiert. Abschließend sei noch erwähnt, daß diese Abbildung selbstverständlich kompatibel zu der XML Schema-Abbildung ist, so daß auch nach der Abbildung die Beziehung zwischen Dokumentbeschreibung und Dokumentinstanz erhalten bleibt.

3.3. Implementierung und Integration

Die beiden beschriebenen Abbildungen wurden in Form zweier plattformunabhängiger Softwarewerkzeuge mit der Sprache Java implementiert. Die Werkzeuge lassen sich mit bestehenden Inferenzsystemen integrieren. Praktisch untersucht wird dies in der Arbeit anhand des RACER-Systems, das sowohl RDF- als auch OWL-Daten verarbeiten kann.

4. Anwendung

Nachdem nun die beiden Abbildungen erläutert worden sind, soll im folgenden auf deren Anwendung eingegangen werden. Da sich die Abbildungwerkzeuge auf jegliche Arten von Dokumenten anwenden lassen, ergeben sich dabei mehrere grundsätzliche Einsatzmöglichkeiten, die zunächst skizziert werden. Anschließend werden diese im Hinblick auf den in der Einführung genannten B2B-Problembereich zu einer Gesamtlösung kombiniert.

4.1. Grundlegende Nutzungsmöglichkeiten

Die Abbildungen von XML und XML Schema ermöglichen in Zusammenhang mit einem DL-Inferenzsystem mehrere potentielle Anwendungsmöglichkeiten. Diese lassen sich gliedern in Anwendungen zur Entwicklungszeit und zur Laufzeit.

Zur Entwicklungszeit: Die XML Schema-Abbildung kann Entwickler dabei unterstützen, Ontologien zu kreieren. Anstatt eine Ontologie komplett neu zu definieren, wird mit der Abbildung eines bestehenden XML Schemas zunächst eine Skelett-Ontologie erzeugt, die dann die Basis für die weitere Arbeit bildet. Dieser Ansatz kann dazu beitragen, Zeit und Kosten für die Entwicklung einer Ontologie zu reduzieren und komplette Neuentwicklungen zu vermeiden.

Ein andere Nutzung besteht darin, mehrere bestehende Schemata zu vergleichen. Logische Widersprüche zwischen ihren Definitionen und implizite Subtypen-Beziehungen lassen sich so automatisch auffinden. Dies ist insbesondere bei großen und unübersichtlichen XML Schemata sinnvoll und erleichert die Wiederverwendung von existierenden Komponenten. Dazu bildet man die Schemata jeweils in OWL-Ontologien ab und vergleicht und analysiert diese anschließend mit Hilfe eines DL-Inferenzsystems.

Zur Laufzeit: In RDF-Graphen transformierte XML-Instanzen können zusammen mit der dazugehörigen OWL-Ontologie in ein DL-Inferenzsystem geladen werden. Damit ist es dann beispielsweise möglich, die enthaltenen Anwendungsdaten semantisch gegenüber den Konzepten der Ontologie zu klassifizieren. Ein Beispiel dazu wäre die Lösung der Fragestellung: „Zu welcher Produktkategorie gehört der vorliegende Artikel anhand seiner Eigenschaften?“. Vor allem aber läßt sich mit diesem Vorgehen zur Laufzeit prüfen, ob die Instanzdaten den in einer Ontologie (z.B. eine erweiterte Skelett-Ontologie) formulierten semantischen Regeln und Vorgaben gehorchen.

4.2. Kombinierter Einsatz im B2B-Bereich

Abb. 1 zeigt eine Methodologie, die die bisher beschriebenen Abbildungen und Einsatzmöglichkeiten zu einer für den B2B-Bereich relevanten Gesamtlösung integriert. Sie ermöglicht es, die zwischen zwei Geschäftspartnern ausgetauschten XML-Dokumente in einem zweistufigen Prozeß zunächst syntaktisch gegenüber einem XML Schema und anschließend semantisch gegenüber einer vom Benutzer bearbeiteten OWL-Ontologie zu validieren.

Zur Entwicklungszeit wird dabei zunächst ein XML Schema in eine Skelett-Ontologie transformiert. Diese wird vom Benutzer um die gewünschten semantischen Einschränkungen und Geschäftsregeln ergänzt, welche sich i.d.R. nicht mittels XML Schema ausdrücken lassen. Ein Beispiel hierfür ist z.B. die Regel „Wenn ein Kunde unbekannt ist, muß die Zahlung per Vorkasse erfolgen.“.

Zur Laufzeit werden die von den Geschäftspartnern empfangenen XML-Dokumente dann zunächst syntaktisch gegenüber dem XML Schema validiert. Dies kann mit einem gewöhnlichen XML-Parser wie z.B. Xerces [11] geschehen. Ist dies erfolgreich, wird das Dokument in einem zweiten Schritt gegenüber den zuvor formulierten semantischen Vorgaben überprüft. Dies wird erreicht, indem das XML-Dokument zunächst nach RDF transformiert und zusammen mit der erweiterten OWL-Ontologie in ein DL-Inferenz-system geladen wird. Hier wird dann mit geeigneten Inferenzdiensten (z.B. per Erfüllbarkeitstest) die Validierung durchgeführt.

In der Diplomarbeit wird die erfolgreiche Durchführung der Methodologie anhand eines Beispielszenarios demonstriert. In Zusammenhang mit einem übergeordneten Anwendungssystem wird außerdem gezeigt, wie zwei Partner bei Aufbau und Verwaltung neuer Geschäftsbeziehungen Vorgaben für die Dokumente des B2B-Prozesses untereinander austauschen und abstimmen können. Dies ist ein Ansatz, um den in der Einführung genannten Aufwand bei der Einrichtung von elektronischen Geschäftsbeziehungen zu reduzieren.

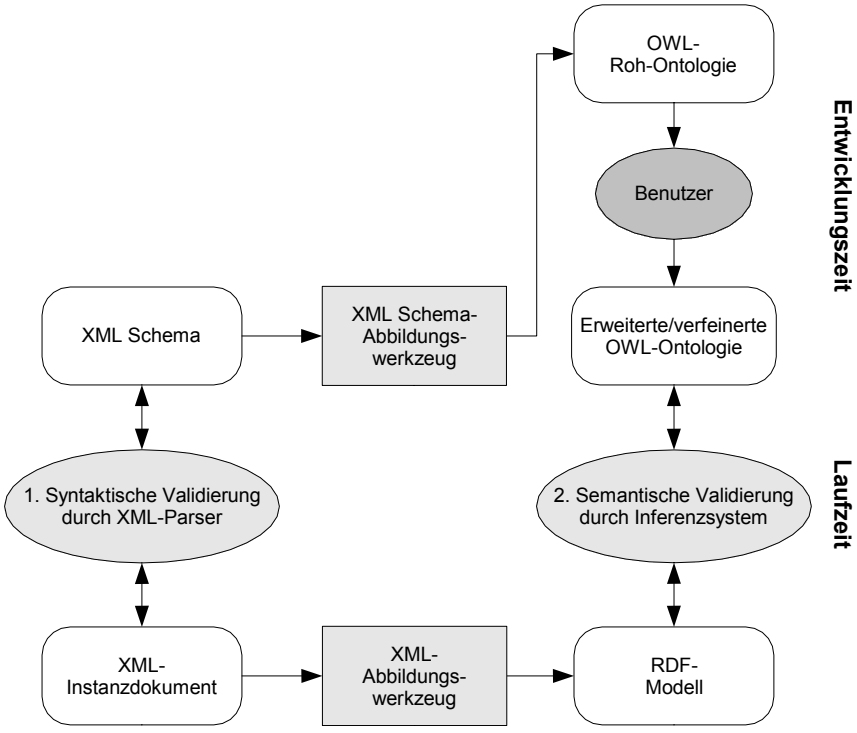

Abb. 1: Zweistufiger XML-Validierungsprozeß

5. Zusammenfassung

Im Rahmen dieser Arbeit wurde eine allgemein anwendbare Lösung entwickelt, mit der sich sowohl XML Schema- als auch XML-Dokumente automatisch auf die Ebene einer Ontologie anheben lassen. Es wurde gezeigt, wie dieser Ansatz in der Praxis dazu beitragen kann, wesentliche Probleme bei der Vorbereitung und Ausführung XML-basierter B2B-Handelsbeziehungen zu lösen. Es lassen sich Einschränkungen und Geschäftsregeln semantischer Natur spezifizieren, die über die Möglichkeiten von XML Schema hinausgehen. Auf dieser Basis wird dann eine semantische Überprüfung und Klassifikation von XML-Daten zur Laufzeit ermöglicht.

Generell sei hervorgehoben, daß mit diesem Vorgehen semistrukturierte Daten den Möglichkeiten der semantischen Modellierung geöffnet werden können. Neben der großen Ausdrucksmächtigkeit der Web Ontology Language lassen sich die Inferenzmechanismen der Beschreibungslogik auf Dokumente für verschiedene Zwecke anwenden. Gegenüber traditionellen Vorgehensweisen, insbesondere der XML-Validierung per ‚festverdrahtetem' Programmcode, werden die semantischen Einschränkungen hier zentral in einer systematischen, einfach wiederverwendbaren und wartbaren Form festgehalten.

Der entwickelte Ansatz trägt dazu bei, eine engere Verbindung zwischen Informationen in XML-Form und dem Semantic Web aufzubauen und zu automatisieren. Er gestattet es, existierende Daten und XML-Anwendungen weiterzuverwenden. Zugleich wird eine konzeptuelle Sicht auf Dokumente bereitgestellt, die von deren Syntax abstrahiert. Parallel kann der Anwender daher zusätzlich von den Vorteilen einer solchen Repräsentation und den neuen Möglichkeiten und Werkzeugen des Semantic Web profitieren.

Literatur

[1] RosettaNet Consortium: RosettaNet. http://www.rosettanet.org/, Abruf am 20.08.2003.

[2] Weitzel, T., T. Harder und P. Buxmann: Electronic Business und EDI mit XML. dpunkt.verlag, Heidelberg, 2001.

[3] Berners-Lee, T., J. Hendler und O. Lassila: The Semantic Web. Scientific American, Mai 2001.

[4] Thompson, H. S. et al.: XML Schema Part 1: Structures. http://www.w3.org/TR/xmlschema-1/, Abruf am 20.08.2003.

[5] Dean, M. et al.: Web Ontology Language (OWL) Reference. http://www.w3.org/TR/owl-ref/, Abruf am 20.08.2003.

[6] Lassila, O. und R. R. Swick: Resource Description Framework (RDF) Model and Syntax Specification. http://www.w3.org/TR/REC-rdf-syntax/, Abruf am 20.08.2003.

[7] Baader, F. et al. (Herausgeber): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, 2003.

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

[9] Haarslev, V. und R. Möller: Description of the RACER System and its Applications. In: McGuinness, D. L. et al. (Herausgeber): Proceedings of the 2001 International Workshop on Description Logics (DL-2001), Stanford, USA. CEUR Workshop Proceedings, Aug. 2001.

[10] Bray, T. et al.: Extensible Markup Language (XML) 1.0 (Second Edition). http://www.w3.org/TR/REC-xml/, Abruf am 20.08.2003.

[11] Apache Software Foundation: Xerces2 Java Parser. http://xml.apache.org/xerces2-j/index.html, Abruf am 20.08.2003.