

Dokumentgrammatiken und die Semantik generischer Dokumentstrukturen

Henning Lobin
Justus-Liebig-Universität Gießen

Jüngste Entwicklungen im Bereich von XML-Schemasprachen haben erneut das Interesse an den formalgrammatischen Grundlagen strukturierter Dokumente entfacht. Neben der inzwischen verabschiedeten ersten Version des W3C-Standards *XML Schema* sind diverse Vorschläge gemacht worden, die teils aus konkreten Anforderungserfordernissen heraus, teils aber auch mit dem ausdrücklichen Ziel entwickelt worden sind, den formalen Spielraum von Markup-Sprachen besser auszunutzen. Murata et al. (2001) betrachten die Klasse der regulären Baum-Sprachen und –Grammatiken und drei Einschränkungen davon mit absteigender Mächtigkeit: *Restrained-Competition*, *Single-Type* und *Local Tree Grammars*. Dabei besitzen die herkömmlichen DTDs die geringste Mächtigkeit (lokale Baum-Grammatiken), XML Schema ist auf der Ebene der *Single-Type* Baum-Grammatiken anzusiedeln, während der aus der Fuji-Entwicklung RELAX und James Clarks' TREX hervorgegangene Vorschlag RELAX NG ohne Einschränkungen Baum-Sprachen zu modellieren erlaubt.

Das Interessante an der Betrachtung der formalen Eigenschaft von Schema-Sprachen besteht darin, dass jede dieser Klassen mit bestimmten Operationen in Verbindung gebracht werden können, unter der die entsprechende Klasse geschlossen ist. So sind DTDs lediglich geschlossen für den Schnitt, nicht aber für Vereinigung oder Differenz. Dieses bedeutet, dass aus zwei DTDs algorithmisch eine DTD abgeleitet werden kann, die alle Dokumente beschreibt, die gleichzeitig für beide DTDs validierbar sind. Baum-Grammatiken lassen sich hingegen auch vereinigen und subtrahieren, ohne das mit der resultierenden Grammatik die Klasse der Baum-Grammatiken verlassen wird. Dieses eröffnet eine Vielzahl von Möglichkeiten für neue Verarbeitungstools, so z.B. Editoren, die aus partiellen Dokumenten partielle Differenzgrammatiken ableiten, die für den verbleibenden Eingabeprozess zu nutzen sind.

Einen weiteren wichtigen Aspekt bildet die semantische Datenmodellierung, für die es Ansätze unter Verwendung von UML, Entity-Relationship-Modellen und Semantischen Netzwerken bzw. Ontologien gibt. Allen diesen Ansätzen ist gemein, durch semantische Modellierungsverfahren die in der grammatischen Datenmodellierung durch DTDs, XML Schema oder andere Schema-Sprachen impliziten Abhängigkeiten explizit zu machen, um dadurch weitergehende Retrieval- und Transformationsprozesse unterhalb der Ebene der semantischen Interpretation der Daten zu ermöglichen. Die dabei verfolgten Ansätze lassen sich unterscheiden in eine Gruppe, bei denen das in der Dokument-Grammatik kodierte Strukturwissen präzisiert wird, wobei zugleich von Reihenfolgebeziehungen und anderen klar auszudrückenden Aspekten abstrahiert wird. Andere Ansätze versuchen das in der Dokument-Grammatik enthaltene Wissen insgesamt zu modellieren, wobei auch Aspekte berücksichtigt werden, die in der Grammatik überhaupt nicht ausgedrückt werden.

Beide Herangehensweisen – sowohl die Untersuchung der formalen Eigenschaften von Dokument-Grammatiken als auch ihre semantische Modellierung – eröffnen einen Bereich für die maschinelle Bearbeitung von XML-Dokumenten in einer Weise, wie sie bislang nur menschlichen Bearbeitern zugänglich war. Insbesondere automatisierte Transformationsprozesse on demand befreien die Informationsmodellierung von der engen Bindung an bestimmte Auszeichnungsvokabularien und lenken das Interesse auf die eigentlich grundlegenden Prinzipien der Dokument-Strukturierung.