

Netzwerkbasierte Modellierung der Semantik von XML-Strukturen*

Henning Lobin

Zusammenfassung

XML-Dokumentgrammatiken, die als DTDs oder neuerdings als XML-Schemata spezifiziert werden, spezifizieren zwar die syntaktischen Eigenschaften einer Klasse von Dokumenten, für sie existiert aber normalerweise kein formales semantisches Modell des Gegenstandsbereichs, auf das Dokumentstrukturen abgebildet werden können. Der Beitrag zeigt am Beispiel der Tabelle, wie semantische Netze für diese Aufgabe herangezogen werden können. Die konkrete Umsetzung geschieht dabei auf der Grundlage des Topic-Map-Standards in Verbindung mit XPath-Ausdrücken, die aus dem semantischen Netz in die Dokumentinstanz bzw. in ein XML-Schema verweisen.

13.1. Problemstellung

Die Verwendung der *Extensible Markup Language* (XML, vgl. z. B. Lobin, 2000) zur Informationsauszeichnung stellt einen qualitativen Schritt in Richtung auf einen semantischen Umgang mit Informationen dar. Nicht mehr die grafische Gestaltung der Daten oder ihre programmtechnische Einbettung steht dadurch im Mittelpunkt, sondern ausschließlich ihre inneren strukturellen Gesetzmäßigkeiten. Den Kern dieser Entwicklung bildet die Eigenschaft von XML, Benutzern die Definition eigener Auszeichnungssysteme zu erlauben. Die Definition derartiger Auszeichnungssysteme geschieht bislang durch sog. Dokumenttyp-Definitionen (DTDs), die kontextfreien Grammatiken mit zusätzlichen Attribut-Spezifikationen entsprechen. In allen Phasen des *Document Lifecycle* spielt die DTD eine zentrale Rolle bei der Automatisierung der beteiligten Verarbeitungsprozesse. Die Datenerfassung geschieht z. B. in XML-Editoren, die die DTD nutzen, um den Benutzer durch die möglichen Datenstrukturen zu leiten. Bei der Verarbeitung derartiger Dokumente können durch die Festlegung möglicher Dokumentinstanzen in der DTD sehr komplexe Transformationen generisch definiert werden (vgl. Lindén, 1997, Lobin und Reinsch, 1999).

Die Erstellung von DTDs wird zwar vereinzelt von Autorensystemen unterstützt, ist aber im wesentlichen eine von Menschen durchzuführende intellektuelle Tätigkeit geblieben, da Nützlichkeitsabwägungen in ihr Design einfließen (vgl. Maler und Andaloussi, 1996). In der DTD schlägt sich das Wissen über die Struktur des Dokumenttyps nieder, sie enthält also die Syntax

* Erschienen in: *Proceedings der GLDV-Frühjahrstagung 2001*, Henning Lobin (Hrsg.), Universität Gießen, 28.–30. März 2001, Seite 141–150. <http://www.uni-giessen.de/fb09/ascl/gldv2001/>

	Hamburg	Berlin
Bielefeld	253,6	384,4
Gießen	442,0	468,2

Tabelle 13.1.: Eine exemplarische Tabelle

der Dokumentstruktur. Die semantische Beschreibung allgemeiner Dokumentstrukturen erfolgt bislang auf informelle Weise ähnlich der Kommentierung und Dokumentation von Programmen (besonders ausgereift etwa bei den TEI-Richtlinien, Sperberg-McQueen und Burnard, 1994). Allerdings finden sich auch in anderen Informationsstrukturen, die auf die DTD Bezug nehmen, wichtige Hinweise auf die semantische Funktion ihrer Teile, etwa in Konvertierungsskripten, Viewing-Informationen, Verarbeitungsprogrammen usw. Bislang wird die Semantik von DTDs – außer durch die Namengebung von Elementen und Attributen – also explizit nur in der Dokumentation der DTD spezifiziert, implizit hingegen in Daten oder Programmen, die auf die DTD Bezug nehmen (vgl. Loeffen, 1996).

Allen Ansätzen der expliziten Dokumentation ist gemein, dass sie sich auf die Spezifikation einer Informationseinheit in ihrem durch die DTD vorgegebenen hierarchischen Zusammenhang beschränken. Dieses ist insbesondere dann von Nachteil, wenn die Beziehungen innerhalb der beschriebenen Daten über die von Baumstrukturen hinausgehen. Ein Beispiel dafür sind Tabellen wie etwa die in Tab. 13.1 dargestellte Entfernungstabelle.

Jede der vier Datenzellen weist Bezüge sowohl zu den Spaltenköpfen als auch zu den Zeilenköpfen auf, da die Entfernung nur als Relation zwischen zwei Orten angegeben werden kann. In einer XML-Umsetzung (hier zugleich die kanonische HTML-Repräsentation) wird eine Tabelle jedoch nur zeilenweise repräsentiert, die Datenzellen also allenfalls mit den Zeilenköpfen in einen unmittelbaren Zusammenhang gebracht:

```
<table>
  <tr>
    <td></td><td>Hamburg</td><td>Berlin</td>
  </tr>
  <tr>
    <td>Bielefeld</td><td>253,6</td><td>384,4</td>
  </tr>
  <tr>
    <td>Gießen</td><td>442,0</td><td>468,2</td>
  </tr>
</table>
```

Die Repräsentation einer Tabelle unterscheidet sich also in nichts von der Repräsentation nicht-tabellarischer, rein hierarchisch gegliederter Information wie etwa einer verschachtelten Liste. Der Bezug der Datenzellen in der Tabelle zu den nirgends als solchen repräsentierten Spalten-

köpfen wird also nicht zum Ausdruck gebracht, obwohl dieses für automatische Dokumentverarbeitungsprozesse von großer Bedeutung ist. Auch andere Beispiele für Abhängigkeiten zwischen Informationseinheiten, die trotz Dokumentation implizit bleiben, zeigen, dass in einer DTD nur in sehr eingeschränkter Weise ausgedrückt werden kann, was eine bestimmte Informationseinheit im semantischen Sinne überhaupt darstellt, mithin also nur sehr unzulängliche Theorien über den Gegenstandsbereich der beschriebenen Datenstrukturen ausgedrückt werden können.

13.2. Semantisches Netzwerk zur Formalisierung von Dokumentsemantik

Eine mögliche einfache DTD zur Validierung der oben gezeigten XML-Dokument-Instanz sieht etwa folgendermaßen¹ aus:

```
<!ELEMENT table (tr*)>
<!ELEMENT tr (td*)>
<!ELEMENT td (#PCDATA)>
```

Um zu spezifizieren, was diese Deklarationen bedeuten sollen, wird gewöhnlich der Weg der Dokumentation gewählt. In der Dokumentation wird jedes Element mit Konzepten des durch die DTD strukturierten Gegenstandsbereich in Verbindung gebracht, hier der Bereich des Renderings textueller Daten für die Darstellung auf Computer-Monitoren. „table“ wird dabei mit der gesamten Tabelle identifiziert, „tr“ mit einzelnen Tabellenzeilen, „td“ mit den in der Tabelle enthaltenen Zellen. In welchem Verhältnis diese Begriffe zueinander sowie zu anderen relevanten Begriffen wie dem der Spalte, des Zeilenkopfes usw. stehen, bleibt dem analytischen Verstand des Benutzers (ggfs. des Programmierers) überlassen, dessen Kenntnis dieses Gegenstandsbereichs vorausgesetzt wird.

Um die Semantik von XML-Strukturen so zu formalisieren, dass automatisierte Verarbeitungsprozesse wie etwa Inferenzziehung oder strukturelle Anreicherungen möglich werden, müssen dagegen etablierte Repräsentationsverfahren der Computerlinguistik und der Künstliche-Intelligenz-Forschung herangezogen werden. Die semantischen Zusammenhänge lassen sich etwa durch ein semantisches Netzwerk wie in Abb. 13.1 (S. 144) formalisieren.²

Dieses semantische Netzwerk erlaubt es, die wesentlichen in der DTD und den annotierten Daten nur implizit enthaltenen Zusammenhänge explizit auszudrücken. Neben den Konzepten, auf die durch die DTD referiert wird (Tabelle, Zeile, Zelle) führt das Netzwerk eine Reihe weiterer Konzepte in die semantische Struktur ein: Spalte, Zeilen- und Spaltenkopf, Datenzelle (in Abgrenzung zur Tabellenzelle überhaupt) sowie Zeilen- und Spalterkörper. Diese Konzepte werden im wesentlichen durch Assoziationen vom Typ „enthält“ und „ist-ein“ miteinander in Verbindung gebracht. Der wesentlichen Zusammenhang, der durch eine Tabelle zum Ausdruck gebracht wird, manifestiert sich durch die Assoziierung von Konzepten vom Typ Datenzelle mit den Konzepten Zeilenkopf bzw. Spaltenkopf durch eine „bezieht sich auf“-Beziehung. Für Zeilenköpfe wird zugleich durch Eigenschaften festgelegt, dass sie jeweils in der ersten Spalte einer Tabelle erscheinen, nicht aber in der ersten Zeile, umkehrt entsprechend auch für Spaltenköpfe. Für diese drei Konzepte sind dabei Instanzen definiert (Zelle-1, Zelle-2 und Zelle-3), als deren Zeilenposition der Wert i und als deren Spaltenposition der Wert j vermerkt ist. Zusammen mit

¹ Hierbei handelt es sich um vereinfachte Deklarationen aus der HTML-DTD.

² Zu semantischen Netzwerken vgl. z. B. Reimer (1991).

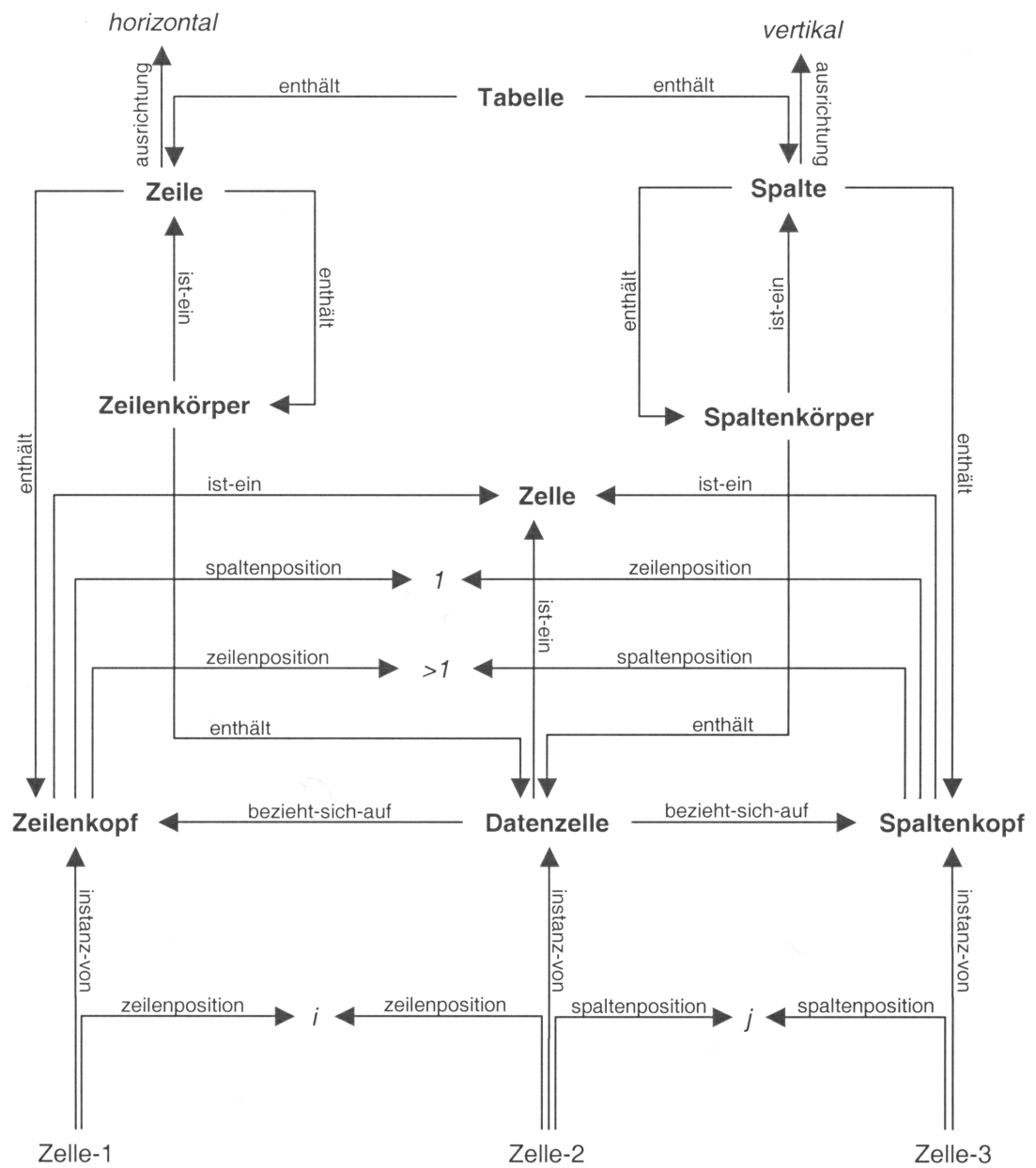


Abbildung 13.1.: Ein semantisches Netzwerk zur Beschreibung von Tabellen

der durch Eigenschaften erfolgenden Festlegung, dass die Zeilenposition von Zeilenköpfen und die Spaltenposition von Spaltenköpf größer Null ist, wird damit die Verbindung einer Datenzelle an der Position (i, j) in der Tabelle mit dem Kopf der Zeile i und dem Kopf der Spalte j beschrieben.

13.3. Verbindung von Semantischem Netzwerk und XML-Strukturen mittels XPath-Ausdrücken

Die Verbindung des semantischen Netzwerks mit der XML-Dokument-Instanz geschieht durch XPath-Ausdrücke. Die Datenzelle „442,0“ in unserer Beispieltabelle kann etwa durch den XPath-Ausdruck `table/tr[position()=3]/td[position()=2]` adressiert werden. Jedes der im semantischen Netzwerk enthaltenen Konzepte kann auf diese Weise durch einen XPath-Ausdruck adressiert werden (i : Zeilenposition, j : Spaltenposition):

Tabelle:	<code>table</code>
Zelle:	<code>table/tr/td</code>
Datenzelle allgemein:	<code>table/tr[position()>1 and position()=<last()]/td[position()>1 and position()=<last()]</code>
Datenzelle speziell:	<code>table/tr[position()=i]/td[position()=j]</code>
Zeilenkopf:	<code>table/tr[position()=i]/td[position()=1]</code>
Spaltenkopf:	<code>table/tr[position()=1]/td[position()=j]</code>
Zeile:	<code>table/tr[position()=i]/td</code>
Spalte:	<code>table/tr/td[position()=j]</code>
Zeilenkörper:	<code>table/tr[position()=i]/td[position()>1 and position()=<last()]</code>
Spaltenkörper:	<code>table/tr[position()>1 and position()=<last()]/td[position()=j]</code>

Die inferenziellen Eigenschaften des semantischen Netzes machen es damit z. B. möglich, von der Datenzelle „442,0“ (XPath: `table/tr[position()=3]/td[position()=2]`) auf den Zeilenkopf „Gießen“ (`table/tr[position()=3]/td[position()=1]`), den Spaltenkopf „Hamburg“ (`table/tr[position()=1]/td[position()=2]`) sowie entsprechend auch auf die übrigen Daten im Zeilen- und Spaltenkörper zu schließen.

Das semantische Netz kann natürlich auch als eine formalisierte Dokumentation der DTD aufgefasst werden. Die drei darin deklarierten Elemente „table“, „tr“ und „td“ werden mit den im semantischen Netz enthaltenen Konzepten „Tabelle“, „Zeile“ und „Zelle“ assoziiert. Werden diese drei Elemente wie oben dargestellt in der traditionellen Syntax deklariert, ist es nicht möglich, über XPath-Ausdrücke den Bezug zu formalisieren, da die DTD-Syntax selbst nicht XML-konform ist. Wird stattdessen zur Spezifikation der Dokumentgrammatik der XML-konforme Standard XML Schema herangezogen, können die Element-Deklaration folgendermaßen adressiert werden:

```
Deklaration des Elements „Tabelle“:  xsd:element[@name="table"]
Deklaration des Elements „Zeile“:    xsd:element[@name="tr"]
Deklaration des Elements „Zelle“:    xsd:element[@name="td"]
```

Die Verweise sowohl auf Abschnitte des XML-Dokuments als auch in ein XML-Schema können als Eigenschaften der Konzepte des semantischen Netzwerks aufgefasst werden. Jedes Konzept besitzt somit die Eigenschaft, einen XPath-Ausdruck zur Lokalisierung einer Konzeptinstanz im XML-Dokument mit sich zu führen, einige Konzepte besitzen darüberhinaus noch die Eigenschaft, über einen XPath-Ausdruck in ein XML-Schema zu verweisen, wo die syntaktischen Eigenschaften des Konzepts definiert werden.

13.4. Semantisches Netzwerk als Topic Map

Für die Implementation von Semantischen Netzwerken in SGML/XML liegt mit *Topic Maps* (ISO/IEC13250, 1999; vgl. z. B. Pepper, 1999, Rath, 2000). ein idealer Repräsentationsstandard vor (Schmidt und Müller, 2000). Generische Instanzen für Datenzelle, Zeilen- und Spaltenkopf mit speziellen Positionseigenschaften lassen sich als *Topics* auffassen, die *Topic Types* wie „Datenzelle“, „Spaltenkopf“, „Zeilenkopf“ usw. zugeordnet sind. Für die Bezugnahme mit der Dokument-Instanz und einem XML-Schema kann das Topic-Map-Konzept der *Occurrence Role* herangezogen werden, durch die die XPath-Adressen den *Topics* und *Topic Types* zugeordnet werden können:³

```
<topic id="Zeile" type="Konzeptklasse">
  <name><basename>Zeile</basename></name>
  <occur occrole="XML-pattern">
    table/tr[position()=i]/td
  </occur>
  <occur occrole="XML-Deklaration">
    xsd:element[@name="tr"]
  </occur>
</topic>
```

Die Beziehungen zwischen den Netzwerk-Konzepten werden durch *Topic Associations* dargestellt:

```
<assoc type="bezieht-sich-auf">
  <assocrl anchrole="Datenzelle">Zeilenkopf</assocrl>
  <assocrl anchrole="Datenzelle">Spaltenkopf</assocrl>
</assoc>
```

Hier werden die „bezieht-sich-auf“-Beziehungen zwischen dem Konzept „Datenzelle“ und „Zeilenkopf“ bzw. „Spaltenkopf“ repräsentiert; obwohl die Assoziationen im Topic-Map-Standard nicht ausdrücklich gerichtet sind, können sie für die Zwecke der Repräsentation von semantischen Netzwerken als gerichtet interpretiert werden.

Es ist zu betonen, dass Topic Maps nicht von vornherein als semantische Netzwerke aufgefasst werden können. Der Topic-Map-Standard beschreibt spezielle Datenstrukturen, die für die Zwecke der Navigation, der Indexierung und des Retrievals in großen Datenbeständen konzipiert sind. Die im Zusammenhang mit semantischen Netzwerken entwickelte Vererbungs- und Inferenzfunktionalität ist bei Topic Maps ebensowenig berücksichtigt worden wie die Terminologie semantischer Netze.

³ Für einen umfangreicheren Ausschnitt aus der Topic Map für das semantische Netz s. Anhang.

13.5. Ausblick

Das in diesem Beitrag dargestellte Konzept der Nutzung semantischer Netzwerke für die Semantik von Dokumentstrukturen lässt sich in die gegenwärtig zu verzeichnenden Aktivitäten zur Schaffung eines *Semantic Web* einordnen. Die von Tim Berners-Lee entwickelte Vision des *Semantic Web* (vgl. Berners-Lee, 1999, S. 177) soll durch die intensive Verwendung von semantischen Metadaten die Verfügbarkeit strukturierter Daten im *World Wide Web* in einem viel weitergehenden Sinne als bisher gewährleisten.

Das generelle Ziel der semantischen Beschreibung von Dokumentstrukturen besteht darin, semantische Informationen zu Dokumentgrammatiken in alle Stadien des *Document Lifecycle* zu integrieren, um auf diese Weise die Verfügbarkeit strukturierter Daten unter Nutzung ihrer semantischen Eigenschaften zu erhöhen. Derartige semantische Metadaten können etwa als *tertium comparationis* zwischen zwei Dokumentgrammatiken herangezogen werden. Werden die durch eine Tabelle wiederzugebenden Daten in zwei partiellen DTDs (bzw. XML-Schemata) auf syntaktisch unterschiedliche Weise strukturiert (z. B. spalten- vs. zeilen-orientiert), so bildet das semantische Netzwerk ein Fundament, beide Strukturen miteinander in Verbindung zu bringen. Die automatische Erstellung von Transformationsskripten tritt damit in den Bereich des Möglichen.

Eine einfachere Möglichkeit, ein solches semantisches Netz nutzbar zu machen, besteht darin, es für Retrieval-Zwecke zu verwenden. Dabei fungiert es als Input für einen XSLT-Prozessor, durch den ein XSLT-Retrieval-Script generiert wird, das zu einer gegebenen Tabellenposition alle dieser Position zugeordneten Informationen ausgibt.

Anhang: Semantisches Netz als Topic Map

Netzwerkkonzepte

```
<topic id="Semantisches-Netzwerk-Konzept">
  <name>
    <basename>Konzept eines semantischen Netzwerkes</basename>
  </name>
</topic>
<topic id="Konzeptklasse" type="Semantisches-Netzwerk-Konzept">
  <name>
    <basename>Konzeptklasse</basename>
  </name>
</topic>
<topic id="Eigenschaft" type="Semantisches-Netzwerk-Konzept">
  <name>
    <basename>Eigenschaft</basename>
  </name>
</topic>
<topic id="Beziehungskante" type="Semantisches-Netzwerk-Konzept">
  <name>
    <basename>Beziehungskante</basename>
  </name>
```

```
</topic>
<topic id="Eigenschaftskante" type="Semantisches-Netzwerk-Konzept">
  <name>
    <basename>Eigenschaftskante</basename>
  </name>
</topic>
```

Occurrence Roles

```
<topic id="XML-pattern">
  <name>
    <basename>XML-Pattern</basename>
  </name>
</topic>
<topic id="XML-deklaration">
  <name>
    <basename>XML-Declaration</basename>
  </name>
</topic>
```

Beispiel für Association Types

```
<topic id="enthaelt" type="Beziehungskante">
  <name>
    <basename>enthaelt</basename>
  </name>
</topic>
```

Beispiel für Eigenschaften

```
<topic id="groesser-1" type="Eigenschaft">
  <name>
    <basename>Zahl groesser 1</basename>
  </name>
</topic>
```

Beispiele für Konzeptklassen

```
<topic id="Tabelle" type="Konzeptklasse">
  <name>
    <basename>Tabelle</basename>
  </name>
  <occur occrole="XML-pattern">table</occur>
  <occur occrole="XML-Deklaration">
    xsd:element[@name="tr"]
  </occur>
</topic>
```

```

<topic id="Datenzelle" type="Konzeptklasse">
  <name>
    <basename>Datenzelle</basename>
  </name>
  <occur occrole="XML-pattern">
    table/tr[position()=i]/td[position()=j]
  </occur>
</topic>
<topic id="Zeilenkopf" type="Konzeptklasse">
  <name>
    <basename>Zeilenkopf</basename>
  </name>
  <occur occrole="XML-pattern">
    table/tr[position()=i]/td[position()=1]
  </occur>
</topic>

```

Beispiel für Instanzen

```

<topic id="Zelle-1" type="Zeilenkopf">
  <name>
    <basename>Instanz eines Zeilenkopfes</basename>
  </name>
  <occur occrole="XML-pattern">
    table/tr[position()=i]/td[position()=1]
  </occur>
</topic>

```

Beispiele für Associations

```

<assoc type="bezieht-sich-auf">
  <assocrl anchrole="Datenzelle">Zeilenkopf</assocrl>
  <assocrl anchrole="Datenzelle">Spaltenkopf</assocrl>
</assoc>
<assoc type="enthaelt">
  <assocrl anchrole="Tabelle">Zeile</assocrl>
  <assocrl anchrole="Tabelle">Spalte</assocrl>
  <assocrl anchrole="Zeile">Zeilenkopf</assocrl>
  <assocrl anchrole="Spalte">Spaltenkopf</assocrl>
  <assocrl anchrole="Zeile">Zeilenkoerper</assocrl>
  <assocrl anchrole="Spalte">Spaltenkoerper</assocrl>
  <assocrl anchrole="Zeilenkoerper">Zelle</assocrl>
  <assocrl anchrole="Spaltenkoerper">Zelle</assocrl>
</assoc>

```

Literaturverzeichnis

- BERNERS-LEE, T. (1999): *Weaving the Web – The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco: Harper San Francisco.
- BRAY, T.; PAOLI, J.; SPERBERG-MCQUEEN, C. M. UND MALER, E. (2000): “Extensible Markup Language (XML) 1.0 (Second Edition)”. Technische Spezifikation (Recommendation), World Wide Web Consortium. Online verfügbar: <http://www.w3.org/TR/2000/REC-xml-20001006>.
- CLARK, J. (1999): “XSL Transformations (XSLT) Version 1.0”. Technische Spezifikation (Recommendation), World Wide Web Consortium. Online verfügbar: <http://www.w3.org/TR/xslt>.
- CLARK, J. UND DE ROSE, S. (1999): “XML Path Language (XPath) Version 1.0”. Technische Spezifikation (Recommendation), World Wide Web Consortium. Online verfügbar: <http://www.w3.org/TR/xpath>.
- FALLSIDE, D.; THOMPSON, H. S.; BEECH, D.; MALONEY, M.; MENDELSON, N.; BIRON, P. V. UND MALHOTRA, A. (2000): “XML Schema”. Technische Spezifikation (drei Teile, Candidate Recommendation), World Wide Web Consortium. Online verfügbar: <http://www.w3.org/XML/Schema>.
- ISO 8879 (1986): “Information Processing – Text and Office Information Systems – Standard Generalized Markup Language”. Internationaler Standard, International Organization for Standardization, Genf.
- ISO/IEC 13250 (1999): “Information Technology – Document Description and Processing Languages – Topic Maps”. Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: <http://www.ornl.gov/sgml/wg4/>.
- LINDÉN, G. (1997): *Structured Document Transformations*. Dissertation, University of Helsinki, Helsinki. Report A-1997-2. Department of Computer Science.
- LOBIN, H. (2000): *Informationsmodellierung in XML und SGML*. Berlin, Heidelberg, New York etc.: Springer, 2. Auflage.
- LOBIN, H. UND REINSCH, M. (1999): “Unification of XML Documents”. *InterChange – The Newsletter of The International SGML/XML Users’ Group* 5 (2): S. 31–33.
- LOEFFEN, A. (1996): “Toward Semantic Specifications for SGML Encoded Documents”. Technischer Bericht, Utrecht University, Utrecht.
- MALER, E. UND ANDALOUSSI, J. EL (1996): *Developing SGML DTDs. From Text to Model to Markup*. Upper Saddle River: Prentice Hall.
- PEPPER, S. (1999): “Navigating haystacks and discovering needles. Introducing the new topic map standard”. *Markup Languages* 1 (4): S. 47–74.
- RATH, H. H. (2000): “Topic maps: templates, topology, and type hierarchies”. *Markup Languages* 2 (1): S. 45–64.
- REIMER, U. (1991): *Einführung in die Wissensrepräsentation. Netzartige und schema-basierte Repräsentationsformate*. Stuttgart: Teubner.
- SCHMIDT, I. UND MÜLLER, C. (2000): “Zaubernetz – Inhaltsstrukturen und Topic Maps als Potenzial neuer Informationstechnik”. *iX* 11: S. 100–107.
- SPERBERG-MCQUEEN, C.M. UND BURNARD, L. (Herausgeber) (1994): *Guidelines for Electronic Text Encoding and Interchange*. Chicago, Oxford: Text Encoding Initiative. Zwei Bände.