XML

Grundlagen und Übungen

Gliederung

Übersicht
XML 1.0 - die Spezifikation
XML Namespaces, XML Schema
XPath, XSLT
Darstellungsfragen - XSL-FO, CSS2

Übersicht

Vorstellung
XML - ein Überblick
(Herkunft, Zweck, Bestandteile)
Die Spezifikationen

Vorstellung

Ihr Dozent
Methodik
Vorgeschichte zum Kurs, Ausblick
Erwartungen der Teilnehmer
Ziele des Kurses
Literaturhinweise

Ihr Dozent

```
<Dozent>
   <Name>
        <Vorname>Heinz</Vorname>
        <Nachname>Werntges</Nachname>
   </Name>
</Dozent>
```


Ihr Dozent

```
<Dozent>
   <Name>
        <Vorname MI= 'W'>Heinz</Vorname>
        <Nachname Titel="Dr">Werntges</Nachname>
   </Name>
   <Beschäftigungsverhältnis Art="LB"/>
</Dozent>
```

Einschub: Erste Eindrücke zu XML

- XML sieht fast aus wie HTML
  - Auszeichnungssprache
  - gemeinsame Herkunft: SGML
- Allgemeine Unterschiede
  - Klein-/Groß-Schrift unterscheiden!
- Elemente
  - anscheinend beliebige Elementtyp-Namen möglich
  - neu: *empty elements*
  - *immer zu schließen!*
- Attribute
  - *single or double quotes, but always quotes*

… und mehr zum Dozenten (aber ohne XML):

- Persönliche Angaben
  - Jahrgang 1959
  - Diplom-Physiker, Dr. rer. nat.
  - Verheiratet, 3 Kinder
  - Hobby: Marathon / Langstreckenlauf
- Fachlicher Werdegang - mit einigen Wendepunkten…
  - Mein erster „PC“: 1979 (Z80-System)
  - Studentenzeit: Mathematik, theoretische Physik (Simulationen, DV), Digitaltechnik
  - Diplomarbeit: Physikalische Biologie (DNA)
  - Promotion: (a) MRI, (b) Neuronale Netze (Robotikumfeld)
    - In dieser Zeit Erfahrungen als Open Source-Entwickler („hp2xx“)
    - Diverse Publishing-Experimente im LaTeX-Umfeld
  - 1993: SW-Entwickler & Berater für EDI-Anwendungen
    - C auf Unix / C++ auf Windows
  - 1994: Projektleiter EDI bei der Braun AG (Gillette), Kronberg
  - Seit 1996: Leiter EDI & EC-Technologien
    - Umfeld: Unix / C / Perl / HTML / DB-gestützte Web-Applikationen
  - 2003: Professor für Angewandte Informatik der FH Wiesbaden

Methodik

- Induktiv
- Deduktiv
- Anekdotisch
- Wenn möglich - interaktiv!

Zum Kurs

- Grundlagen-Teil
  - zu einer möglichen Reihe von Veranstaltungen
  - für diverse XML-basierte Technologien
- Thema noch im Aufbau!
  - Neuer Dozent, neues Thema
  - WS 2002/03: Noch geringe Kapazität, Dozent = Lehrbeauftragter (noch!)
  - Neues Lehrgebiet „Angewandte Informatik“
  - Volle Kapazität im SS 2003, bei Bedarf bereits dann eine Wiederholung des Kurses
- Ausblick
  - Folgekurs mit Schwerpunkt „Anwendung“
    - Web Services, APIs, XSLT

Ihre Erwartungen ...

Ziele des Kurses

- Allgemeine Ziele
  - Solide Grundlagen für weiterführende Arbeiten auf XML-Basis
  - Kenntnis und sicheres Beherrschen der wichtigsten Standards
  - Fähigkeit zu eigener Recherche von Detailfragen in diesen W3C-Dokumenten
  - Entscheidungsgrundlagen für Designfragen

Ziele des Kurses

- Publishing-Seite von XML („POP“)
  - Eigene Dokumente auf der Basis vorhandener *document type definitions* (DTDs) bzw. Schemata erstellen und validieren
  - Verständnis von DTDs/Schemata
    - Fähigkeit zur Erweiterung vorhandener DTDs/Schemata
    - Fähigkeit zum Entwurf eigener - einfacher - DTDs.
  - Verarbeitung von XML-Quellen mit Standard-Tools z.B.
    - zu HTML
    - zu PDF
  - „Nagelprobe“:
    Was wird das Quellformat Ihrer Diplomarbeit sein?
    - Noch WORD‘s DOC-Format - oder besser gleich XML? ...

Ziele des Kurses

- Messaging-Seite von XML („MOM“)
  - Grundlagen für Web Services-Technologien erwerben (SOAP, SWA; WSDL, UDDI, …)
  - Erarbeitung eines realen B2B-Dokumenttyps (z.B. Bestelldatenaustausch)
  - Präzise Typisierung mittels XML Schema

Literaturhinweise

- Jonathan Pinnock et al.: **Beginning XML** 2nd ed.
  Wrox Press Ltd., Birmigham, UK, 2001.
  - Zum Teil Leitfaden / Quelle von Beispielen in dieser Vorlesung
  - Didaktisch gut, recht vollständig, tiefergehend als der Titel suggeriert.
  - Reihenfolge der Themen offenbar durch Autorenrollen geprägt...
- Charles Goldfarb, Paul Prescod: **Charles F. Goldfarb's XML Handbook** - 4th ed. Prentice Hall, Upper Saddle River, NJ, 2002.
  - Von den SGML- und XML-Vätern selbst
  - Erklärt besonders gut die "warum"-Fragen
  - Große Sammlung von Anwendungsfällen und Ausgangspunkten für weiterführende Themen sowie von Web-Ressourcen.
- Elizabeth Castro: **XML for the World Wide Web.**
  PeachPit Press, Berkeley, 2001.
  - Kochbuchartig, kompakt, preiswert
  - Fokus auf Web publishing mit XML
- XML Spezifikationen (http://www.w3c.org) - mehr dazu später...

XML - ein Überblick

Herkunft
Motive der Urheber
Bestandteile

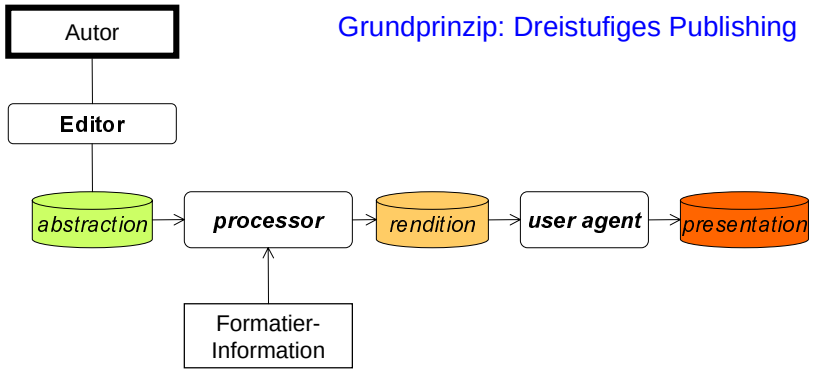
Herkunft: *Information interchange*

- Austausch zwischen Computersystemen
  - ohne Informationsverlust, d.h. hochstrukturiert
  - von einfachen Zahlen bis zu hochkomplexen Datenstrukturen einerseits ...
  - … und von menschenlesbaren Dokumenten andererseits, bis hin zu ganzen Buchreihen
  - … mit der Option zur vollautomatischen Weiterverarbeitbarkeit ...
  - … wie auch der zur auszugsweisen Übermittlung und Referenzierung bei längeren Dokumenten …
  - … sowie der Option zur semantischen, inhaltsbezogenen Suche
- Hier bereits erkennbar: POP & MOM
  - Zwei Enden eines Spektrums der Möglichkeiten

POP: *People-Oriented Publishing*

- Die *text processing*-Tradition!
- Der Schlüssel zum Erfolg:
  Konsequente Trennung der Dokumenterzeugung in
  - *Abstraction*
    - Die logische Struktur eines Dokuments
    - Beispiele: LaTeX- oder SGML-Quelldaten
  - *Rendition*
    - Die darstellungsorientierte Aufbereitung des Dokuments
    - Beispiele: HTML, troff, RTF, (La)TeX, PS, PDF
  - *Presentation*
    - Das abgelieferte Ergebnis
    - Ausdruck, angezeigte Seite, Tonspur, ...

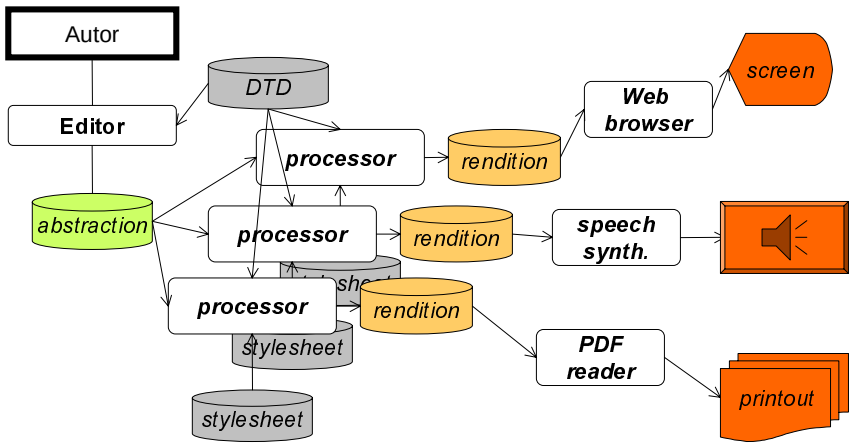
POP: *People-Oriented Publishing*



Grundprinzip: Dreistufiges Publishing

POP: *People-Oriented Publishing*

Vision: Mehrfach-Verwertung eines abstrakten Dokuments



POP: People-Oriented Publishing

- Bemerkungen & Warnungen (auch provokative)
  - Die gedankliche Trennung von logischer Ebene und Darstellungsebene erfordert Disziplin und zunächst Mehraufwand, ggf. Verhaltensumstellungen.
  - Contra WYSIWYG - *What you see is all you get*.
  - Populäre Publishing/DTP-Werkzeuge und *text processors* trennen gerade nicht bzw. erschweren eine saubere Trennung der logischen von der Darstellungsebene
  - Wer WORD als Schreibmaschinenersatz kennenlernte, wird diese Trennung nur schwer nachvollziehen.
  - Wer beim Verfassen eines Textes schon an‘s spätere Aussehen denkt, hat die Methode nicht verinnerlicht.
  - Denke „genauso“ über die logische Struktur Deines Dokumentes nach wie über seinen Inhalt - am besten sogar zuerst.

MOM: *Machine-Oriented Messaging*

- *EDI, EAI, ATA, B2B, IEC, …*
  - *EDI: Electronic Data Interchange*
  - *EAI: Enterprise Application Integration*
  - *ATA: Application-to-application*
  - *B2B: Business-to-business …*
- *IEC: Integrated E-Commerce*
  - Goldfarb‘s Versuch eines Oberbegriffs
  - Betonung auf vollautomatische Weiterverabeitung, typisch:
    - Kleine Dokumente, dafür zahlreich, kurzlebig
    - I.d.R. einfache Strukturen (na ja…)
    - Kein / kaum menschlicher Eingriff in die Verarbeitungskette
    - Oft hohe Folgekosten bei Fehldeutungen (produktionskritisch)
    - Daher höchste Präzision notwendig.

XML im Spannungsfeld der Erwartungen

- XML: Gemeinsame Grundlage für das breite Spektrum zwischen POP- and MOM-Anwendungen
  - Skeptiker: Noch eine eierlegende Wollmilchsau...
  - Optimisten: Endlich - das „TCP/IP der Datenformate“
- Pole der Diskussion
  - Überall anwendbar („Es gibt kaum etwas, das sich nicht mit XML machen läßt“) versus
  - Nirgends wirklich nötig („Praktisch alle XML-Lösungen wurden schon mit anderen Mitteln gelöst“)
- Versachlichung
  - Mehraufwand jetzt vs. Langfristnutzen ?
  - Entsteht Gewinn durch Vereinheitlichung des Methoden- und Werkzeug-Pools?
  - Ausmaß der Wiederverwertbarkeit semantisch sauber strukturierter Dokumente?

Herkunft: SGML

- … oder: Der Weg zum universellen Datenformat
- 1969: Markup, GML
  - Charles Goldfarb, Ed Mosher, Ray Lorie (IBM)
  - GML = *Generalized Markup Language*
  - Prinzipien:
    - Einheitliche *representation* als *markup*
    - Erweiterbarkeit der *Markup*-Sprache
    - Formale Definition & Beschreibung von Dokumenttypen
- 1974: SGML-Geburtsstunde
  - Erster validierender Parser
- 1986: ISO 8879 (SGML)
  - Ausgereifter, komplexer Industriestandard

Von SGML zu XML

- 1989: Tim Berners-Lee
  - gründet HTML auf der Basis von SGML (eigentlich nur GML)
  - aber ohne Erweiterbarkeit und formale Überprüfung
- 199x: Das drohende HTML-Chaos
  - Proprietäre Erweiterungen, inkompatible Browser
  - *W3C-Reaktionen:*
    - *Style sheets* (CSS) - Übernahme eines weiteren GML-Konzepts
    - Erste Ansätze zur standardisierten Erweiterbarkeit von HTML
- 1996: XML Working Group
  - Chair: Jon Bosak, Sun
- 1998-02-10: XML 1.0 - endlich der „große Wurf“?
  - Übernahme auch des dritten Leitgedankens von SGML:
    - Strenge Dokumenttyp-Definitionen und deren Überprüfung
    - Allgemeine Erweiterbarkeit

Von SGML zu XML

- 1998-02-10: XML 1.0
  - Autoren (allesamt langjährige Markup-Verfechter):
    - Tim Bray (Netscape),
    - Jean Paoli (Microsoft),
    - C.M. Sperberg-McQueen (TEI / W3C)
  - Übernahme auch des dritten Leitgedankens von SGML:
    - Strenge Dokumenttyp-Definitionen und deren Überprüfung
    - Allgemeine Erweiterbarkeit
  - Endlich der „große Wurf“?
- 2000-10-06: XML 1.0 (SE)
  - Inhaltlich unverändert, nur „*errata*“ berücksichtigt

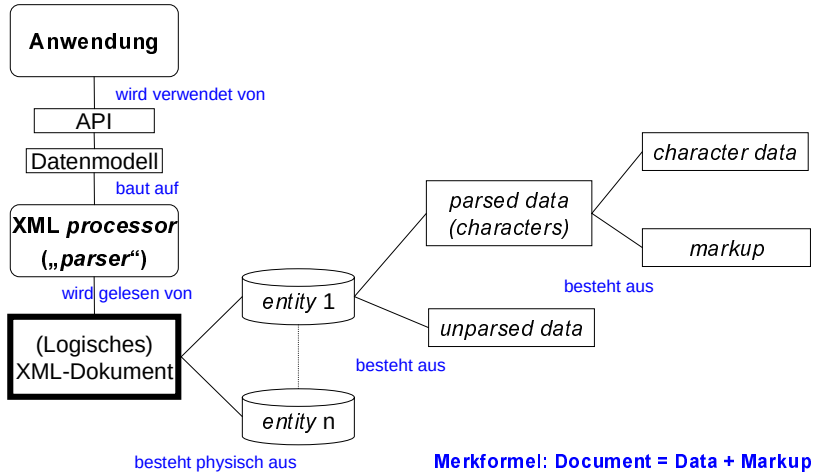
XML: Ein SGML-Subset

- Beibehaltung der wichtigsten Vorzüge von SGML
  - Jedes gültige XML-Dokument ist auch ein gültiges SGML-Dokument
  - 15 Jahre Industriepraxis von SGML werden geerbt
  - Abwärtskompatibilität führt manchmal zu nicht-intuitiven Erweiterungen
- Vereinfachungen für Web-Zwecke
  - minimalistische Tradition
- Weitere Anleihen
  - *Extensible Style Language* (XSL): abgeleitet von
    - CSS des Web einerseits und
    - ISO's DSSSL (*Document Style Semantics and Specification Language*, sprich „dissel") andererseits
  - *Extensible Linking Language* (XLink): abgeleitet von
    - HyTime (ISO Standard zum Verlinken von SGML-Dokumenten)
    - TEI (*Text Encoding Initiative*)-Regeln (akadem. SGML-Umfeld)
  - Unicode (http://www.unicode.org), ISO 10646
  - RFC 1766 (*language ID tags*), ISO 639 (*language name codes*), ISO 3166 (*country name codes*)

Die 10 Design-Ziele laut XML 1.0

- *XML shall be straightforwardly usable over the Internet.*
- *XML shall support a wide variety of applications.*
- *XML shall be compatible with SGML.*
- *It shall be easy to write programs which process XML documents.*
- *The number of optional features in XML is to be kept to the absolute minimum, ideally zero.*
- *XML documents should be human-legible and reasonably clear.*
- *The XML design should be prepared quickly.*
- *The design of XML shall be formal and concise.*
- *XML documents shall be easy to create.*
- *Terseness in XML markup is of minimal importance.*

Eine erste Analyse



XML - Die wichtigsten Spezifikationen

Namen, Quellangaben und Kurzbeschreibungen

XML-Spezifikationen

- Hinweis zu deutschsprachiger Dokumentation
  - Generell sind nur die englischsprachigen W3C-Originaldateien maßgebend. Diese werden auf den folgenden Seiten im Detail benannt.
  - Es gibt ein Projekt, das (nach und nach) diese W3C-Dateien ins Deutsche übersetzt und veröffentlicht. Es besitzt dazu die Genehmigung vom W3C - und ist in dieser Hinsicht die einzige deuschsprachige „offizielle“ Quelle dieser Art.
  - Übersetzte Dateien finden sich unter dem gewohnten URL, wenn man den *domain*-Teil wie folgt systematisch abändert. Beispiel:
    - Original: http://www.**w3c.org**/TR/REC-xml
    - Deutsche Version: http://www.**edition-w3c.de**/TR/REC-xml
  - Nur wenige Dokumente sind bereits übersetzt. Was übersetzt ist, ist selten ein sehr aktuelles Dokument. Über die Qualität lassen sich noch keine Aussagen treffen. Daher: Besser gleich das englische Original lesen!

XML-Spezifikationen

- XML: Extensible Markup Language 1.0
  - *Autoren: Tim Bray*, *Jean Paoli, C. M. Sperberg-McQueen*
  - Quelle: http://www.w3c.org/TR/REC-xml
  - Kommentare:
    - DIE Grundlage schlechthin
    - Präzise, kompakt, vollständig, allgemeingültig.
    - Erfreulich klar geschrieben und gut lesbar - nach einer Eingewöhnung...

XML-Spezifikationen

- Namespaces in XML
  - *Autoren:* *Tim Bray*, *Dave Hollander, Andrew Layman*
  - Quelle: http://www.w3c.org/TR/REC-xml-names
  - Kommentare:
    - Eine Konvention zur Vermeidung von Namenskollisionen beim Mischen von XML-Dokumentteilen unterschiedlicher Herkunft.
    - Grundlage weiterer Spezifikationen, insb. von
      - XML Schema
      - XPath
      - XSLT
      - XPointer, XLink

XML-Spezifikationen

- CSS2: Cascading Style Sheets - level 2
  - *Autoren:* *Håkon Wium Lie, Bert Bos*
  - Quelle: http://css.nu/pointers/
  - Kommentare:
    - Dieser *style sheet*-Standard ist bereits über HTML populär geworden und wird bereits von gängigen Browsern unterstützt.
    - Er gestattet auch die „direkte“ Gestaltung von XML-Dokumenten bei deren Anzeige in heutigen Browsern
    - Im XML-Kontext ist CSS noch wichtiger als im HTML-Kontext, da die Browser zwar für HTML verwendbare *default style*-Einstellungen aufweisen, aber XML-Dokumente - wenn überhaupt - nur in einer pauschalen, am hierarchischen Aufbau orientierten Weise darstellen.
    - XML + CSS2 kann in vielen einfachen Fällen bereits befriedigende Ergebnisse auf der Darstellungsseite liefern.
    - Ein viel allgemeingültigerer Ansatz verwendet XSL (XSLT, FO).

XML-Spezifikationen

- XSL: eXtensible Stylesheet Language
  - *Autoren:* *Sharon Adler, Anders Berglund, Jeff Caruso, Stephen Deach*, *Paul Grosso, Eduardo Gutentag, Alex Milowski, Scott Parnell*, *Jeremy Richman, Steve Zilles*
  - Quelle: http://www.w3c.org/TR/XSL/
  - Kommentare:
    - Diese Spezifikation beschreibt, wie das Formatieren von XML-Dokumenten generell spezifiert werden soll.
    - Die Umsetzung dieser Formatierungen bleibt XSLT überlassen.

XML-Spezifikationen

- Associating stylesheets with XML documents
  - *Autoren:* *James Clark*
  - Quelle: http://www.w3c.org/TR/xml-stylesheet
  - Kommentare:
    - Dieser - sehr einfache - Standard beschreibt, wie *stylesheets* aus einem XML-Dokument heraus angesprochen (referenziert) werden.
    - Die Methode versteht sich als unabhängig von der benutzten Anwendung wie auch unabhängig von der verwendeten *stylesheet*-Sprache.

XML-Spezifikationen

- XPath: XML Path Language
  - *Autoren:* *James Clark*, *Steve DeRose*
  - Quelle: http://www.w3c.org/TR/xpath
  - Kommentare:
    - Mit dieser leistungsfähigen und erweiterbaren Sprache können XML-Dokumente durchsucht und gefiltert werden.
    - Sie dient insbesondere zur Adressierung bestimmter Bestandteile von XML-Dokumenten und zur Berechnung von Werten aus diesen.
    - Ein sehr vielseitiger und nützlicher Standard
    - Wird selten isoliert verwendet, sondern bildet die Grundlage weiterer Standards, insb. von XSLT und XPointer sowie XQuery.

XML-Spezifikationen

- XSLT: XSL Transformations
  - *Autoren:* *James Clark*
  - Quelle: http://www.w3c.org/TR/xslt
  - Kommentare:
    - XSLT ist eine - in XML verfasste - Transformationssprache zur Steuerung der Umwandlung von XML-Dokumenten in
      - (a) andere XML-Dokumenttypen
      - (b) HTML
      - (c) beliebige Textformate
    - Sie entstand durch Abspaltung aus XSL, als man erkannte, dass die Transformation von XML-Dokumenten eine eigenständige und über *stylesheets* hinausreichende Aufgabe ist.
    - XSLT basiert auf XPath

XML-Spezifikationen

- XML Information Set (infoset)
  - *Autoren:* *John Cowan*, *Richard Tobin*
  - Quelle: http://www.w3c.org/TR/xml-infoset
  - Kommentare:
    - Dieser Standard legt ein formales Datenmodell für XML-Dokumente fest, d.h. definiert die Datenstrukturen, die ein XML *parser* erzeugt.
    - Da grundsätzlich über Parser und demnach über derartige Datenstrukturen auf XML-Daten zugegriffen wird - und eben nicht auf den XML-Quelltext - hat dieser Standard eine viel größere Bedeutung als sein bisher geringer Verbreitungsgrad vermuten lässt.
    - Grundlage für XML APIs

XML-Spezifikationen

- XML Schema Part 0: Primer
  - *Autoren:* *David C. Fallside*
  - Quelle: http://www.w3c.org/TR/xmlschema-0
  - Kommentare:
    - Dies ist *kein* Standard, sondern ein Tutorium.
    - Empfohlene Lektüre vor der Beschäftigung mit den eigentlichen Schema-Spezifikationen.

XML-Spezifikationen

- XML Schema Part 1: Structures
  - *Autoren:* *David Beech, Murray Maloney, Noah Mendelsohn, Henry S. Thompson*
  - Quelle: http://www.w3c.org/TR/xmlschema-1
  - Kommentare:
    - Einer der wichtigsten Standards für anspruchsvolle Entwicklungen auf XML-Basis.
    - Unter Beibehaltung der formalen SGML-Kompatibilität wurde dieser Standard XML beigefügt - und geht dabei weit über SGML hinaus.
    - Ein Schema wird selbst in XML spezifiziert.
    - Schemata ergänzen die traditionen DTDs um zahlreiche Verallgemeinerungen zur präzisen und erweiterten Definition von Elementtypen.
    - Teile 1 & 2 bilden zusammen die vollständige Schema-Spezifikation.

XML-Spezifikationen

- XML Schema Part 2: Datatypes
  - *Autoren:* *Paul V. Biron, Ashok Malhotra*
  - Quelle: http://www.w3c.org/TR/xmlschema-2
  - Kommentare:
    - Ein reicher Satz vordefinierter Datentypen sowie die Möglichkeit, komplexe eigenen Datentypen zu definieren, bilden einen Schwerpunkt der Schema-Spezifikationen.
    - Die Spezifikation der Datentypen wurde als separater Teil formuliert, denn die Schema-Datentypen lassen sich (per Konvention) auch innerhalb von DTDs verwenden! Allerdings unterstützt nicht jede XML-Software diese erst nachträglich eingeführte Erweiterung.

XML-Spezifikationen

- XLink: XML Linking Language
  - *Autoren: Eve Maler, Steve DeRose*
  - Quelle: http://www.w3c.org/TR/xlink
  - Kommentare:
    - XLink legt fest, wie Hyperlinks in XML-Dokumenten definiert werden.
    - XLink geht deutlich über die Möglichkeiten der HTML-Hyperlinks hinaus.

XML-Spezifikationen

- XPointer: XML Pointer Language
  - *Autoren: Ron Daniel Jr.*, *Eve Maler, Steve DeRose*
  - Quelle: http://www.w3c.org/TR/xptr
  - Kommentare:
    - Der XPointer-Standard begleitet XLink und beruht auf XPath
    - Mittels XPointer ist es XLink möglich, Hyperlinks bis tief in XML-Dokumente hinein präzise zu definieren.

XML 1.0 - die Spezifikation

- Erarbeitung des Originals per Browser