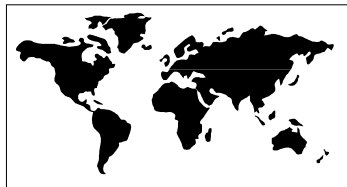


WWW - Grundlagen und Technologie

Berichte und Trends – WWW10 und XML Europe 2001



Erik Wilde
TIK – ETH Zürich
Sommersemester 2001

Übersicht

- WWWx Konferenzen
- Trends und Schwerpunkte
- Schema-Sprachen für XML
 - XML Schema als "DTD++"
 - weitere Schema-Sprachen
- XML Kompressionsmethoden
 - bisher kein Standard
 - Anwendungsszenarien

WWW10

- WWWx Konferenzen vom IW3C2 veranstaltet
 - weltweit die wichtigste generelle Web-Konferenz
 - Genf, Chicago, Darmstadt, Boston, Paris, Santa Clara, Brisbane, Toronto, Amsterdam, Hong Kong, Honolulu, Budapest, ...
- verschiedene parallele Programme
 - Refereed Papers (20% Acceptance Rate)
 - W3C Track (Standardisierungsaktivitäten)
 - Culture & Society Track (spezifisch untechnisch)
 - Web & Industry Track (Industrieprojekte)
- Tutorials & Developer's Day

XML und DTDs

- DTDs sind der traditionelle Weg
 - definiert im SGML- und im XML-Standard
 - momentan der einzige etablierte Standard für Schemadefinitionen
- DTDs haben einige Nachteile
 - die Syntax ist nicht XML-Dokumentensyntax
 - die Strukturierungsmechanismen sind einfach (insbesondere ist keine Vererbung möglich)
 - sie kennen keine Datentypen (schlecht für B2B Szenarien)

Neue Schema-Sprachen für XML

- DTDs kommen aus der Dokumentenwelt
 - XML ist populär in anderen Szenarien
- weitergehende Ansätze
 - verwenden meistens XML-Syntax
 - bieten weitergehende Modellierungsmethoden
 - unterstützen Datentypen
- Problem der Ersetzung der DTDs
 - DTDs sind integraler Bestandteil von XML
 - jeder XML Prozessor muss DTDs unterstützen
 - neue Schema-Sprachen müssen sich etablieren

Die Schema Begriffsverwirrung...

- eine Schema beschreibt eine Dokumentenklasse
 - Beschreibung von Element- und Attributtypen
 - Beschreibung ihrer erlaubten Verwendung
 - Kombinationsmöglichkeiten
 - erlaubte Datentypen in den Instanzen
- DTDs sind eine mögliche Schema-Sprache
 - speziell, weil im XML-Standard selbst definiert
- XML Schema vom W3C definiert
 - extrem unglückliche Namensgebung
- keine relevanten anderen Dialekte

W3C XML Schema

- zweigeteilter Standard
 - ein Standard für die Strukturierung
 - ein Standard für die Datentypen
- DTDs decken nur den ersten Teil ab
 - zweiter Teil sehr minimal (z.B. `CDATA/ID/IDREF`)
- erste Implementierungen vorhanden
 - Microsofts *XML-Data* ist ein proprietärer Ansatz
 - *Document Content Description (DCD)*
- W3C *Recommendation* seit der WWW10 (05/00)

XML Schema Part 1: Structures

- Reformulierung der DTD-Mechanismen in XML
 - Elemente zur Elementbeschreibung
 - Elemente zur Elementdefinition (+ Content Model)
 - Elemente zur Attributdefinition
- Tools zur Konvertierung von DTDs nach Schema
 - zu wenig Infos (Schema ist mächtiger als DTD)
 - Rückrichtung möglich, aber nur mit Verlusten
- Modellierung immer mit Blick auf Schema
 - DTD als Zwischenlösung, Schema als Grundlage
- Tools unterstützen manchmal nur DTDs
 - Frage: Validierung gemäss DTD oder Schema?

XML Schema Part 2: Datatypes

- definiert ein Typensystem für XML Schema
 - einige Grundtypen (Zahlen, Strings, Datum, ...)
 - benutzerdefinierte Typen
- Datentypen werden charakterisiert
 - Einschränkungen auf Wertebereiche
 - Einschränkungen auf lexikalische Werte
 - Verwendung von *Regular Expressions*
- Implementierungen bisher keine!
 - grosser Aufwand (recht komplexer Standard)
 - grosser Nutzen (hohe Qualität der Dokumente)
- Standard verfolgen und als Ziel sehen

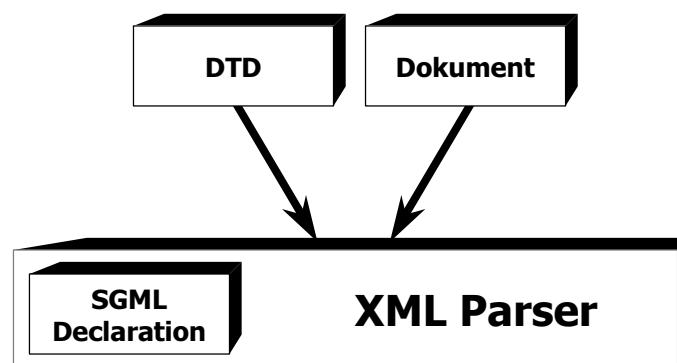
Valid und schema-valid XML

- XML unterscheidet zwischen zwei "Levels"
 - *well-formed* gehorchen dem XML-Standard
 - *valid* sind *well-formed* und gehorchen einer DTD
- *well-formed* und *valid* Konzepte
 - sind direkt im XML Standard definiert
 - können mit DTD und Dokument verifiziert werden
- *schema-valid* Dokumente
 - müssen gemäss eines XML Schema validiert werden
 - gibt es nur mit XML Schema Applikationen
 - haben mehr Randbedingungen als *valid* Dokumente
 - sollten kontrolliert importiert/exportiert werden

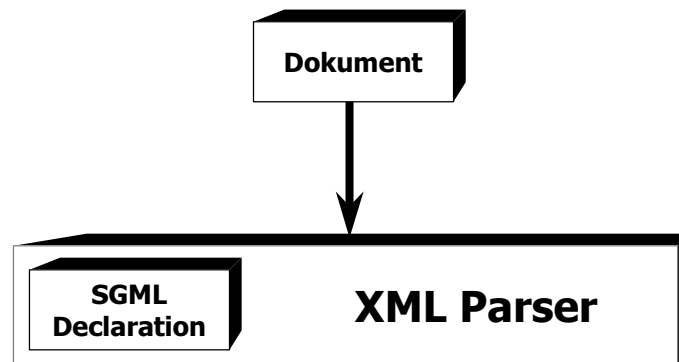
Well-formed und valid XML

- XML unterscheidet zwischen zwei "Levels"
 - *well-formed* gehorchen dem XML-Standard
 - *valid* sind well-formed und gehorchen einer DTD
- *well-formed* Dokumente
 - falls keine DTD vorhanden (nicht immer nötig!)
 - falls DTD nicht verfügbar
 - falls keine Weiterverarbeitung notwendig
- *valid* Dokumente
 - Validierung anhand einer DTD
 - notwendig zur Weiterverarbeitung
 - im B2B Umfeld wohl ausnahmslos valid XML

XML Parser (Validierung)



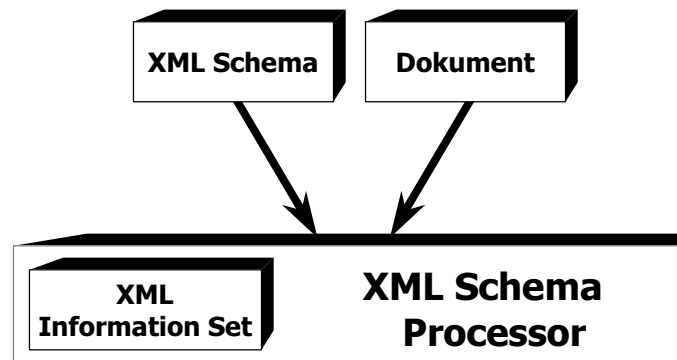
XML Parser (Well-formedness)



XML Schema Validierung

- well-formed und valid sind XML Konzepte
 - basieren auf der XML Syntax
 - arbeiten direkt auf einem XML 1.0 Dokument
- XML Schema basiert auf dem XML Infoset
 - damit ist well-formedness sozusagen garantiert
 - Validierung des XML Infoset
 - Annotierung des XML Infoset
- XML Schema definiert neue *Information Items*
 - Kennzeichnung validierter Teile des Baumes
 - operiert komplett auf der abstrakten Ebene

XML Schema Parser



WWW (SS2001) - WWW10 & XML Europe 2001

15

Die Zukunft von XML Schema

- XML Schema löst viele Probleme
 - entwickelt aufgrund Benutzeranforderungen
 - sinnvoller Schritt weg von der Syntax-Ebene
- andere Schemaentwicklungen laufen parallel
 - "Semantic Web" mit dem *Resource Description Framework (RDF)*
 - Topic Maps als RDF-ähnliche Sprache
- Entwicklung noch nicht 100%ig klar
 - XML Schema wird sicher akzeptiert
 - parallele Entwicklungen noch nicht genau absehbar

WWW (SS2001) - WWW10 & XML Europe 2001

16

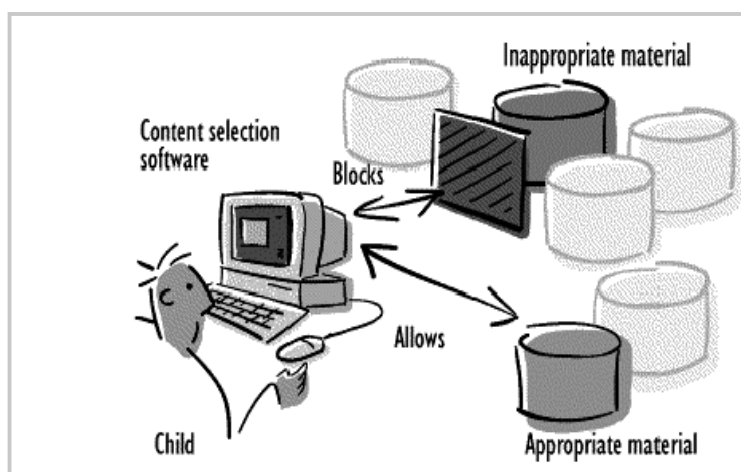
Platform for Internet Content Selection

- gedacht als Plattform für die Selbstkontrolle
- drei verschiedene Arten, Labels zu verwenden
 - in einem HTML-Dokument
 - mit einem HTML-Dokument (in HTTP Headers)
 - Sammlungen von Labels unabhängig vom Server
- *Rating Systems* und *Rating Services*
 - Rating System: Menge an möglichen Labels
 - Rating Service: erzeugt und verteilt Labels
- PICS definiert nur den Austausch von Labels
 - keine Definition der Erzeugung von Labels
 - keine Definition der Software zur Auswertung

WWW (SS2001) - WWW10 & XML Europe 2001

17

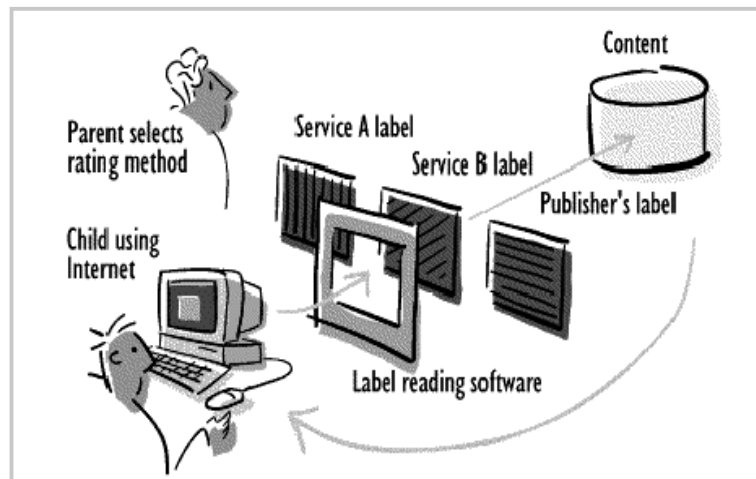
Content Selection (PICS)



WWW (SS2001) - WWW10 & XML Europe 2001

18

Unterschiedliche PICS Labels



WWW (SS2001) - WWW10 & XML Europe 2001

19

PICS Architektur

- *PICS Services and Ratings*
 - beschreibt *rating services*
 - wichtigster Teil: *rating system*
 - Dimensionen und Wertebereiche
- *PICS Label Distribution*
 - eingebettet im Dokument (HTML **<META>** Element)
 - zusammen mit dem Dokument (HTTP Header für Anforderung eines Ratings und Label)
 - *label bureau*: spezialisierter HTTP-Server
- *PICSRules*
 - Definition von Profilen (Service, Label, Verhalten)
 - ermöglicht den Austausch von Profilen

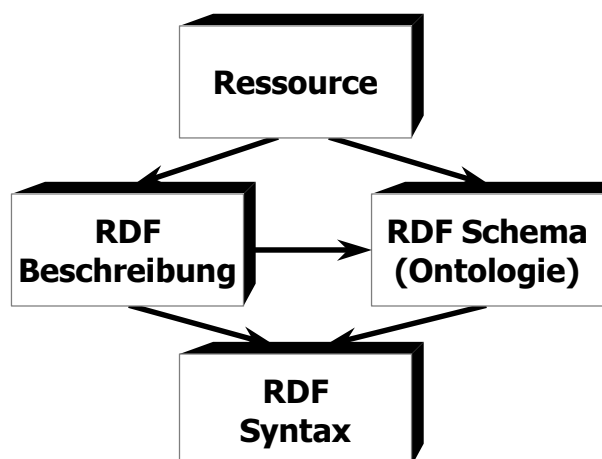
WWW (SS2001) - WWW10 & XML Europe 2001

20

Resource Description Framework

- Entwicklung aus verschiedenen Bereichen
 - XML als Beschreibung strukturierter Dokumente
 - XML *Web Collections* als Microsofts Vorschlag
 - XML *Meta Content Framework (MCF)* von Netscape
- RDF Schemas ähnlich den PICS Rating Systems
 - allgemeiner, keine Festlegung von konkreten Werten
 - orientiert an Datenbanken und mehr noch Knowledge Representation Languages (wie z.B. KIF)
- erst im Anfangs-Stadium, Entwicklung noch unklar (Unterstützung der Search Engines)

RDF Metadata



XML Kompressionsmethoden

- XML ist auf dem Syntax-Level standardisiert
 - Austausch von Daten auf Zeichen-Ebene
 - Unicode UTF-8 als Standard-Zeichensatz
- XML ist keine kompakte Syntax
 - besonders deutlich bei stark strukturierten Daten
 - stark strukturierte Daten der Default im B2B Fall
- kompakte Codierung von XML
 - Reduzierung des Datenvolumens
 - keine Kompatibilität mit dem XML 1.0 Standard

Zwei Wege der XML Kompression

- XML Dokument komprimieren
 - muss einen eigenen Parser benutzen
 - komprimiert wirklich alles (z.B. wichtig bei steganographischen Dokumenten)
 - konzeptionell eher nicht die beste Lösung
- XML Infoset komprimieren
 - arbeitet direkt auf dem abstrakten XML
 - Parser ist eine u.U. notwendige Vorstufe
 - komprimiert nur die "relevanten" Teile
 - geht sehr stark in die Richtung von ASN.1
- bisher nur Ansätze der ersten Art

Topic Maps

- Zweigleisigkeit der Entwicklung
 - RDF aus dem W3C-Umfeld, TimBL als grösster Fan
 - Topic Maps aus dem ISO-Umfeld
- RDF als Grundbaustein des "Semantic Web"
 - beschreibt Ressourcen auf semantischem Level
- Topic Maps als semantisches Netz
 - beschreibt Ressourcen durch Zuordnung zu Topics
- beide Standards haben viele Überschneidungen
 - erste Ansätze zur gemeinsamen Betrachtung
 - bisher nur sehr bescheidene Bestrebungen

Zusammenfassung

- Konferenzen als Wissensquelle
 - es ist alles schon mal dagewesen...
 - übliches Web-Problem: Informationen finden
- interessantes Gebiet, viele Aktivitäten
 - Möglichkeit für SA/DA und mehr!
 - bei Interesse Mail an net.dret@dret.net
 - ...oder einfach mal vorbeikommen (ETZ D97.7)