XML Vorlesung FH Aargau, SS 2006

Organisatorisches, Überblick und Einführung

Erik Wilde

13.3.2006

http://dret.net/lectures/xml-fhnw-ss06/

Übersicht

- Inhalt und Ziel der Vorlesung
  - Vorstellung der behandelten Themen
- Herkunft und Motivation für XML
  - wofür sollte XML ursprünglich benutzt werden
  - wofür wird XML heute grösstenteils benutzt
- Thema der Übungen

Inhalt und Ziel der Vorlesung

- Inhalt: XML als Format für strukturierte Daten
  - XML als hierarchisches Informationsmodell
  - XML Schema als XML-Standard der Zukunft
  - XSLT als Transformationssprache für XML
  - Speicherung von XML in XML Datenbanken
- Ziel: Umgang mit XML lernen
  - XML primär als Informationsmodell begreifen
  - Tools für den Umgang mit XML kennenlernen
  - XML als Grundlage für viele Anwendungen sehen
- "XML is ASCII for the 21st century"

Vorlesungsüberblick

13.3.2006	Überblick und Einführung
20.3.2006	XML Grundlagen (XML Dokumente und DTDs)
27.3.2006	
3.4.2006	XML Informationsmodelle und XML Namespaces
10.4.2006	XML Path Language (XPath)
24.4.2006	XML Schema
8.5.2006	
15.5.2006	
5.6.2006	XML Schemasprachen
12.6.2006	XSL Transformations (XSLT)
19.6.2006	
26.6.2006	
3.7.2006	
10.7.2006	Ausblick XPath 2.0, XSLT 2.0 und XQuery 1.0

Herkunft und Motivation für XML

- XML im Vergleich zu HTML
- Motivation für die Einführung von XML
  - Geschichte des Web
  - HTML als präsentationsorientierte Sprache
  - SGML als erprobte Grundlage von HTML
- Aufbau einer Architektur mit XML
  - XML als Basis für viele andere Komponenten
  - XML ist nur der Anfang der Entwicklung
  - XML wird in verschiedenen Bereichen verwendet
    - Web, B2B, XML Protocol, Prototyping

XML und HTML

- sieht "so ungefähr" aus wie HTML
  - gleiche Basis (*Standard Generalized Markup Language, SGML*)
  - *proven success* (SGML und HTML sind Erfolge)
  - geringere Hemmschwelle für Umsteiger
- funktioniert ähnlich wie HTML
  - gleiche Strukturierungsverfahren (Grammatiken)
  - rein textorientiertes Format (keine Binärdaten!)
- andere Zielgruppe als HTML
  - weiterverarbeitbare Information (B2B)
  - anwendungsabhängige Datenstrukturen
  - etabliertes Umfeld (EDIFACT, SGML, proprietär)

Probleme mit HTML

- HTML als Standard des "prä-XML Web"
- HTML ist präsentationsorientiert
  - erste Probleme schon seit Beginn des "Browser War"
  - trotz Verbesserungen (CSS) bleibt der Grundansatz
  - HTML basiert auf reiner Präsentationssemantik
- gewünscht: Wiederverwendbarkeit von Inhalten
  - HTML als Dokumentenformat ungeeignet
  - HTML nur eine mögliche Form der Präsentation
- Content Management wird immer wichtiger
  - fehlende Freiheit, eigene Strukturen zu verwenden
  - fehlende Flexibilität, auf Neues zu reagieren (WAP *Wireless Markup Language, WML*)

SGML

- *Standard Generalized Markup Language*
  - definiert in ISO Standard 8879 (1986)
- Markup Language trennt Struktur und Text
  - Struktur ist syntaktisch identifizierbar
- Vorteile einer Markup Language
  - einfache Repräsentation (Erzeugung "von Hand")
  - Austauschbarkeit als rein textbasierte Dokumente
- Definition beliebiger Dokumenttypen
  - anwendungsspezifische Dokumenttypen
  - gemeinsamer zugrundeliegender Mechanismus
  - gemeinsam verwendbare Software

SGML als Design-Philosophie

- GML als erstes inhaltsorientiertes Format
  - Dokumentformatierung als getrennter Schritt
  - Definition von Dokumentenklassen (Datentypen)
- SGML kombiniert Struktur und Text
  - zeichenbasiertes Format (keine binäre Form)
  - Strukturinformation in der Form von "Markup"
- internationaler Standard (ISO 8879, 1986)
  - nicht unter Kontrolle eines Herstellers
  - unabhängig von spezifischer Software
- Optimierungen für manuelle Eingabe
  - gedacht als Erleichterungen für Menschen
  - aber Hindernisse und Anforderungen für Maschinen

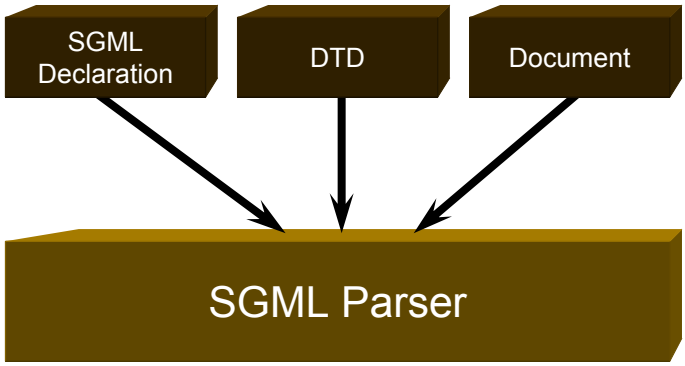
Markup Beispiel (XML ⇒ SGML)

```
<?xml version="1.0" ?>
<!DOCTYPE kurs SYSTEM "kurs.dtd">

<kurs>
<titel kurz="XML">XML - Grundlagen und Umfeld</titel>

<referent email="xml@dret.net"
          homepage="http://dret.net/">
  <vorname>Erik</vorname>
  <name>Wilde</name>
  <organisation homepage="http://www.tik.ee.ethz.ch/">ETH
  Zürich</organisation>
</referent>

<referent> ... </referent>
<inhalt> ... </inhalt> </kurs>
```

Aufbau von SGML

- die *SGML Declaration*
  - konkrete Syntax (Zeichen mit spezieller Bedeutung)
  - SGML Features (wie *Markup Minimization*)
- die *Document Type Definition (DTD)*
  - definiert erlaubte Elemente
  - definiert Kombination der erlaubten Elemente
  - definiert Attribute der erlaubten Elemente
  - definiert *Entities*, z.B. für Sonderzeichen
- das *Document* selber
  - Struktur des Inhaltes gemäss der DTD
  - Inhalt der Struktur (Text)

SGML Parser



SGML Declaration

- bezieht sich auf mehrere Dokumente
- Festlegung der konkreten Syntax
  - SGML selber verwendet abstrakte Syntax
  - Zeichen mit Sonderbedeutung
  - HTML: `< > </ <! <? = " & ;`
- Festlegung von Zeichensätzen
- Festlegung von Kapazitäten
  - Länge von Namen, Schachtelungstiefen, ...
- SGML Features (von HTML bzw. XML verwendet)
  - *tag omission* (Weglassen von Tags)
  - *short tags* (Abkürzen von Tags)

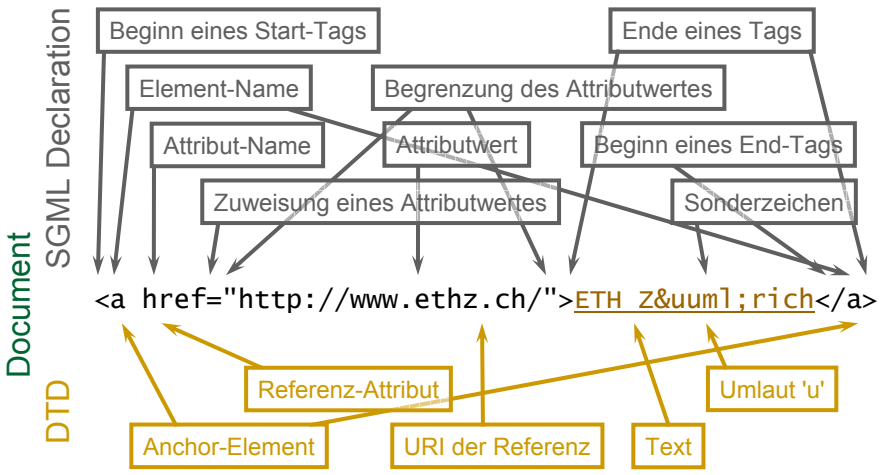
SGML Document Type Definition

- Festlegung einer Grammatik
  - bestimmt die *Wörter* einer Sprache
  - bestimmt die *Regeln zur Satzbildung*
- Definition der Elemente
  - Elementnamen (frei wählbar)
  - Attributnamen, -typen und -werte
- Definition zur Kombination der Elemente
  - Vorkommen der Elemente in einem Dokument
  - *Model Groups*: `<!ELEMENT UL (LI)+ >`
  - nur HTML: *Exceptions (Inclusions/Exclusions)*

Document

- Instanz eines bestimmten Dokumententyps
  - Kennzeichnung zu Beginn des Dokuments
  - kann nur mit Hilfe der DTD interpretiert werden
- Einhaltung der Regeln der SGML Declaration
- Einhaltung der Regeln der DTD
- stellt einen *document tree* dar
- Erstellung eines SGML Dokumentes
  - als Textdatei (ursprüngliches Modell)
  - mit SGML-Tools (zunehmend verbreitetes Modell)

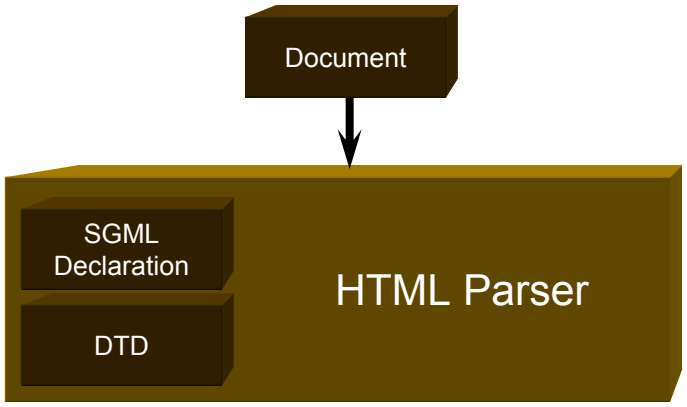
SGML-Teile des HTML Standards



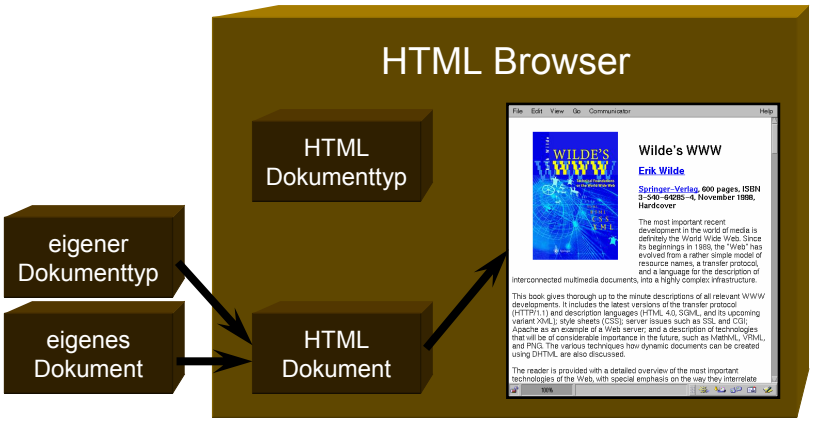
Hypertext Markup Language (HTML)

- Anwendung (*Application*) von SGML
- HTML SGML *Declaration*
  - bestimmt die konkrete Syntax von HTML
  - bestimmt die SGML Features von HTML
- HTML SGML *Document Type Definition (DTD)*
  - drei verschiedene Varianten der DTD
  - Definition der Elemente, Attribute und Grammatik
- zusätzliche Definitionen (kein SGML!)
  - Einschränkungen von Attributwerten (z.B. Zahlen)
  - Bedeutungen von Elementen und Attributen
  - alles in Prosa beschrieben (nicht formal definiert)

HTML Parser



Publishing mit HTML



13.3.2006 XML Vorlesung FHA SS 2006 19

HTML in the Wild

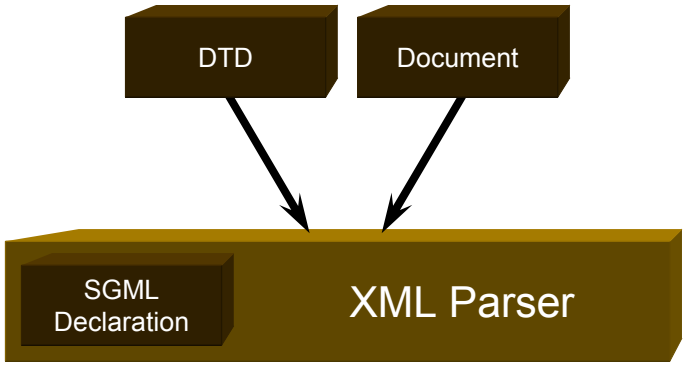
- fast alle HTML-Seiten sind fehlerhaft
- validieren Sie Ihre meistbenutzten Seiten
  - `http://validator.w3.org/`
- gruppieren Sie die Fehlermeldungen
  - mögliche Fehlerklassen
    - SGML Syntaxfehler (z.B. fehlende/falsche Markup-Zeichen)
    - falsch verwendetes HTML (z.B. fehlendes Attribut)
    - schlecht verwendetes HTML (kein formaler Fehler, aber schlecht für die Weiterverarbeitung)
  - eine feinere Einteilung ist möglich
- wieso sind die meisten Seiten fehlerhaft?

13.3.2006 XML Vorlesung FHA SS 2006 20

Extensible Markup Language (XML)

- benutzerdefinierbare Dokumenttypen
- überwindet Einschränkungen von HTML
  - beliebige Dokumenttypen
  - neues Problem: Semantik von Elementen
  - begleitende Mechanismen werden notwendig
- überwindet Komplexität von SGML
  - fest definierte konkrete Syntax (SGML Declaration)
  - keine Markup Minimization (immer volles Markup)
  - reduzierte Zahl an erlaubten Attributtypen
- Ziele sind Einfachheit und Flexibilität

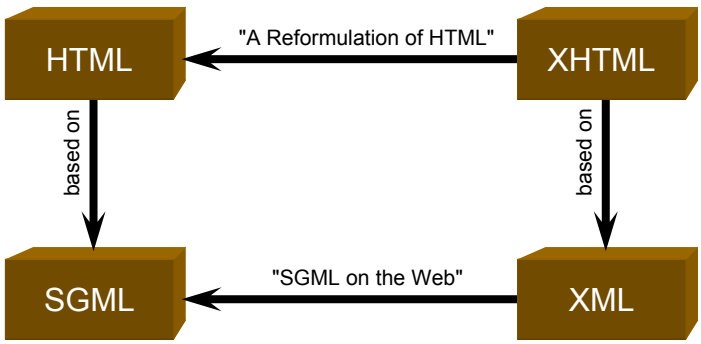
XML Parser



Vergleich SGML/XML/HTML

	SGML	XML	HTML
SGML Declaration	frei	fix	fix
DTD	frei	frei	fix
Document	frei	frei	frei

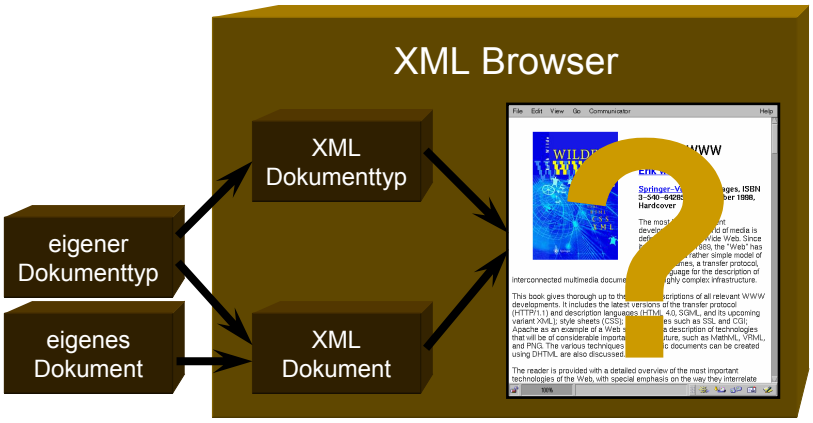
SGML/HTML und XML/XHTML



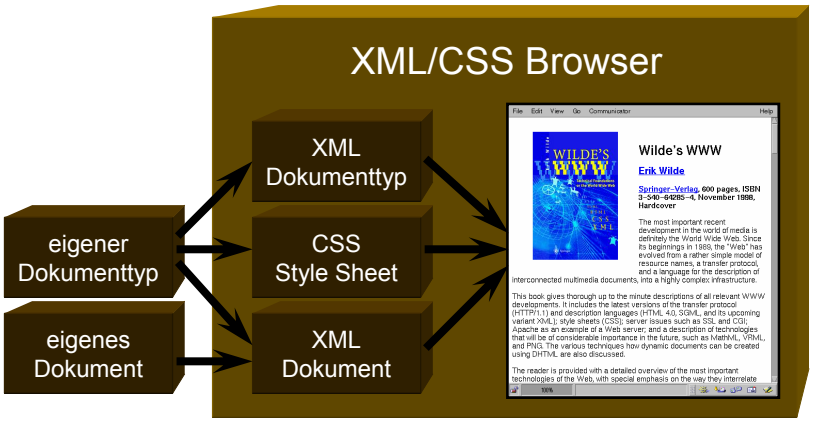
Unterschiede HTML/XHTML

- XML ist immer Case Sensitive
  - HTML benutzt SGML's Default ("ignore case")
- alle Tags in XML müssen angegeben werden
  - in HTML ist "Tag Omission" erlaubt
- XML Spezialschreibweise für leere Elemente
  - in HTML nicht erlaubt
- alle XML Attributwerte müssen Quotes benutzen
  - in SGML nur nötig, falls Spaces im Wert
- XML Attributnamen müssen angegeben werden
  - in HTML sind "Short Tags" erlaubt

Publishing mit XML

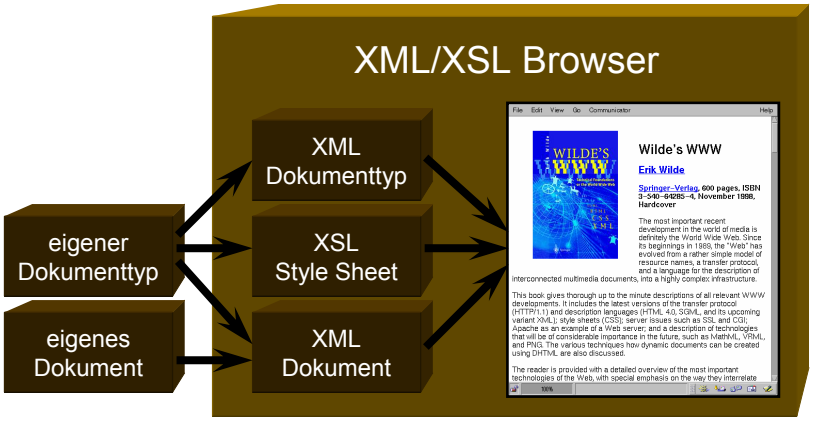


Publishing mit XML/CSS



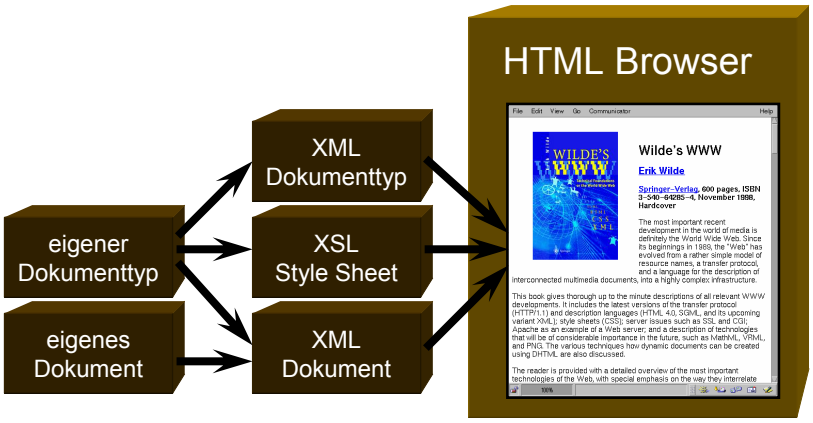
13.3.2006 XML Vorlesung FHA SS 2006 27

XML/XSL Publishing (Client-Side)

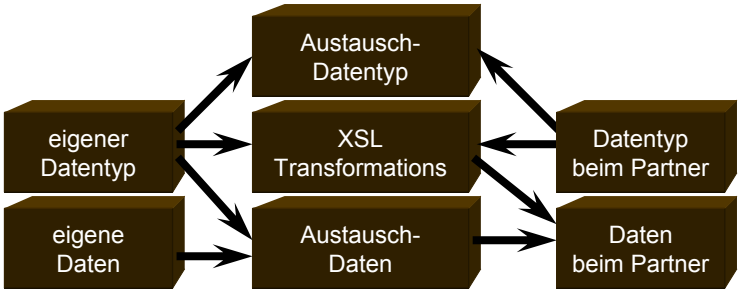


13.3.2006 XML Vorlesung FHA SS 2006 28

XML/XSL Publishing (Server-Side)



B2B und XML/XSLT (Server-Side)



XML im B2B-Bereich

- mit Abstand grösster Erfolg von XML
  - XML als politisch erfolgreiche Lösung
    - keine Bindung an Hardware oder Betriebssystem
    - keine Bindung an Hersteller oder Konsortien
  - XML kam zur richtigen Zeit
    - Anfang 1998: Internet-Boom in voller Fahrt
    - Investitionswille praktisch unbeschränkt
- Anwendungen auf dem Web noch am Anfang
  - bisher nur möglich in Form von XHTML
  - fehlende Bausteine und fehlende Software

XML und verwandte Standards

- XML selber ist ein einfacher Standard
  - Syntax für baumstrukturierte Dokumente
  - Syntax für eine einfache Schema-Sprache
- XML wird von vielen Standards begleitet
  - Hauptziel der Vorlesung: Orientierung bieten
  - Standards kennenlernen und einschätzen können
- W3C entwickelt viele XML-Standards
  - z.T. von Konsortiumsmitgliedern eingebracht
- viele Entwicklungen in anderen Organisationen

Zusammenfassung

- Grundlagenvorlesung zum Thema XML
  - Basis für Anwendungen in vielen Bereichen
  - praktische Erfahrung mit XML-Technologien
- XML als Web-Standard
  - entwickelt als "SGML on the Web"
  - Anwendungen in verschiedensten Bereichen
- XML lebt vom Umfeld
  - orientieren und einschätzen lernen
- XML wird es noch eine ganze Weile geben...

Entwickeln Sie Ihr XML-Szenario

- XML-Dokumente sind strukturierte Daten
  - hierarchisch organisierte Daten
  - u.U. typisierte Daten
  - aufeinander verweisende Daten
- ähnlich einer Datenbank
  - allerdings ohne das DBMS
  - mit einem anderen Datenmodell
  - basierend auf anderen Technologien
- nehmen Sie ein Beispiel für Sie wichtiger Daten
  - überlegen Sie sich Strukturen und Zusammenhänge
  - Darstellung in ER-Form oder anderer Notation

Übungsgrundlage

- XML-Beschreibung von XML-Technologien
  - Einträge für alle relevanten Technologien
  - Verweise zwischen den Einträgen
    - basiert auf, definiert von, Konkurrenz zu, ...
  - Beschreibungen der Einträge
  - Links zu den Einträgen
  - Klassifizierung der Einträge
    - Protokolle, Formate, APIs, ...
  - Bücher zum Thema
    - ISBN, Titel, Verlag, Author, ...
- Beispiel unter `http://dret.net/glossary/`