



XML-Technologien

Heutige Vorlesung



- Organisatorisches
- Einführung und Überblick
- Klausurrelevante Literatur



Organisatorisches

Webseite der Lehrveranstaltung



- http://www.inf.fu-berlin.de/inst/ag-nbi/lehre/04/V_XML/
- hier finden sich:
 - Termine
 - Folien
 - Übungsblätter
 - Hinweise zur Lösung der Übungen
 - Musterlösungen
 - Hinweise auf Literatur

Anmeldung



- heute in Anmeldeleiste eintragen
- zusätzlich in Mailing-Liste eintragen:
http://lists.spline.inf.fu-berlin.de/mailman/listinfo/nbi_v_xml
- Msc-Studierende:
 - *verbindliche* Anmeldung zur Veranstaltung notwendig
 - Ohne diese Anmeldung dürfen *keine* Leistungen erbracht werden.
 - verbindliche Anmeldung mit Unterschrift in der nächsten Woche

Übungsbetrieb



- jeweils Mi 16:15-17:45
- abwechselnd Tutorium und betreute Rechnerübung

Tutorium



- ab 21.4. alle *zwei Wochen* Mi 16:15-17:45 im SR 055
- dort:
 - Fragen zur Vorlesung
 - Erläuterungen zum aktuellen Übungsblatt
 - Rückgabe des letzten Übungsblattes
- 6 Übungsblätter
 - Aufgaben mit projektcharakter
 - werden bewertet
- Es müssen Gruppen von 3 Studierenden gebildet werden.

Betreute Rechnerübung



- ab 28.4. alle *zwei Wochen* Mi 16:15-17:45 in einem der Rechnerräume im Keller
- dort:
 - *kein* Frontalunterricht
 - stehe bei Problemen bei der Bearbeitung des Übungsblattes zur Verfügung

XML-Editor



- 20 Lizenzen von XMLSpy stehen in den PC-Pools zur Verfügung
- XMLSpy unterstützt XML, DTDs, XML-Schema, XSLT, SOAP und WSDL
- XMLSpy-Einführung:
<http://www.xmlspy.de/documents/xmlspy2004proTutorial.pdf>

Scheinkriterien



- alle Übungsaufgaben erfolgreich gelöst
- die Klausur erfolgreich bestanden
- Gesamtnote:
 - Mittel aus Übungsnote und Klausurnote
 - bei überdurchschnittlicher Abweichung behalte ich mir Anpassung der Gesamtnote vor

Termine



- erstes Tutorium am 21.4.
- verpflichtende Anmeldung für MSc-Studierende am 21.4.
- Klausur am 28.7.

Kommunikation



- Veranstalter: schild@inf.fu-berlin.de
- Sprechstunde: Termin außerhalb der betreuten Rechnerübung per E-Mail vereinbaren
- Mailing-Liste: nbi_v_xml@lists.spline.inf.fu-berlin.de

Inhalt der Veranstaltung

Vorlesungsinhalt

XML-Basistechnologien (6 Termine)

- maschinenlesbare Dokumente (Maschine-Maschine-Kommunikation)
- *nicht* behandelt werden: XML-Technologien zur Präsentation von Dokumenten: XHTML oder WML (Mensch-Maschine-Kommunikation)

Web-Dienste (Web Services) (4 Termine)

- plattformunabhängige XML-Protokolle für verteilte Systeme

Semantic Web (1 Termin)

- Linkstruktur des WWW wird durch maschinenverarbeitbare Beziehungen ergänzt

Was ist XML?

HTML



hat sich für die Präsentation von Inhalten bewährt

Layoutunabhängige Repräsentation

- Trennung von Inhalt und Präsentation
 - Vielfalt von Endgeräten und Bandbreiten
- Austausch von Daten/Dokumenten zwischen Computern
 - z.B. Übermittlung eines Bestellformulars
 - z.B. Google-Suchanfrage ohne Browser

HTML: *keine* layoutunabhängige Repräsentation von Inhalten

XML

- Sprache des Webs zur layoutunabhängigen Repräsentation von Inhalten
 - XML *ersetzt* HTML nicht, es ergänzt HTML
- XML ist wie SGML eine verallgemeinerte Auszeichnungssprache

Binärdateien



- enthalten reichhaltige Informationen, wie Daten zu interpretieren sind (Metadaten)
- Z.B. kann Word-Dokument Informationen enthalten, dass ein bestimmtes Textfragment eine Kapitelüberschrift ist.
- **Nachteil:** werden nur von bestimmten Anwendungsprogrammen verstanden

Textdateien



- enthalten wie Binärdateien nur Bits
- Bits so angeordnet, dass sie Zahlen repräsentieren, die wiederum bestimmte Zeichen darstellen
- **Vorteil:** werden von jedem Anwendungsprogramm verstanden, das die entsprechende Zeichenrepräsentation kennt (z.B. Texteditoren)
- werden deshalb als anwendungsunabhängig bezeichnet
- **Nachteil:** Alle Zeichenketten werden gleich behandelt: keine Informationen, wie Daten zu interpretieren sind (Metadaten).

Auszeichnungssprachen



- kombinieren Vorteile von Binärdateien mit denjenigen von Textdateien:
- anwendungsunabhängige Dateiformate, die reichhaltige Metadaten enthalten können
- **Auszeichnungssprache (markup language):** textbasierte Sprache, die Dokumente/Daten mit so genannten *Tags* („Markierungen“) und dadurch mit zusätzlicher Information (Metadaten) versieht :

`<tag-name>ausgezeichneter Text</tag-name>`



- **Beispiel:** *Hypertext Markup Language* (HTML)

Verallgemeinerte Auszeichnungssprachen



- HTML: *vorgegebene* Auswahl von Tags, keine anderen dürfen verwendet werden.
- **verallgemeinerte Auszeichnungssprache (generalized markup language):** keine Tags vorgegeben, beliebige Tags möglich
- **Vorteil:** beliebige Metainformationen darstellbar
- **Nachteil:** Bedeutung der Metainformationen (Tags) offen und muss durch die Anwendung festgelegt werden
- **Beispiele:** SGML und XML

SGML



- *SGML = Standard Generalized Markup Language*
- 1969 von Charles Goldfarb und zwei seiner Kollegen bei IBM für das Dokumentenmanagement entwickelt.
- seit 1986 ein internationaler Standard
- *keine* vorgegebenen Tags, auch keine für das Layout von Dokumenten

Beispiel



- SGML erlaubt das Strukturieren von Dokumenten:

```
<book>
  <title>Beginning XML</title>
  <edition>2nd</edition>
  <authors>
    <author>David Hunter</author>
    <author>Curt Cagle</author>
    <author>Chris Dix</author>
  </authors>
  <date>2001</date>
  <publisher>Wrox Press</publisher>
  <abstract>...</abstract>
  <chapters>...</chapters>
</book>
```

SGML

- gibt *keine* konkreten Tags vor
- erlaubt, spezielle Auszeichnungssprachen mit konkreten Tags zu definieren (= Untermengen von SGML)
- Solche Untermengen von SGML werden auch als *Anwendungen* bezeichnet.
- Bekannteste Anwendung von SGML ist HTML.
- SGML wird häufig auch als *Meta-Sprache* bezeichnet.
- weitere Informationen:
<http://userpage.fu-berlin.de/~corff/SGML/>

Vor- und Nachteile von SGML

- + kombiniert SGML die Vorteile von Binärdateien mit denjenigen von Textdateien
- + beliebig erweiterbar
- + erlaubt die Definition von konkreten Auszeichnungssprachen wie HTML
- sehr komplex: Spezifikation über 500 Seiten lang
- SGML-Parser schwierig zu implementieren

Extensible Markup Language (XML)

- HTML hat sich für die Präsentation von Inhalten bewährt.
- HTML erlaubt aber keine layoutunabhängige Repräsentation von Inhalten.
- Hierfür sind verallgemeinerte Auszeichnungssprachen (wie SGML) besser geeignet.
- Für das Web ist allerdings SGML viel zu komplex.

XML ist eine konsequente Vereinfachung von SGML, die für Web-Anwendungen hinreichend allgemein ist.

Eine kurze Geschichte von XML

- 1969** Charles Goldfarb entwickelt bei IBM die Generalized Markup Language (GML).
- 1980** ANSI veröffentlicht ersten Entwurf von SGML.
- 1986** ISO verabschiedet SGML.
- 1989** Berners-Lee schlägt SGML-basiertes Hypertext-System vor.
- 1990** Berners-Lee entwickelt HTML, HTTP und URL. World Wide Web seinen Betrieb mit zwei Maschinen am CERN auf.
- 1994** Gründung des World Wide Web Consortiums (W3C)
- 1995** HTML 2.0
- 1998** XML 1.0
- 2000** XHTML 1.0 (Reformulierung von HTML in XML)
XML 1.0 (2nd Edition)

Extensible Markup Language (XML)

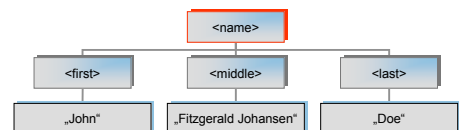
- wie SGML erlaubt XML das Strukturieren von Dokumenten

```
<name>  
<first>John</first>  
<last>Doe</last>  
</name>
```

- XML-Dokumente werden von den meisten modernen Browsern angezeigt.



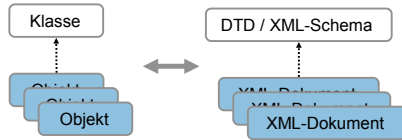
Baumstruktur von XML



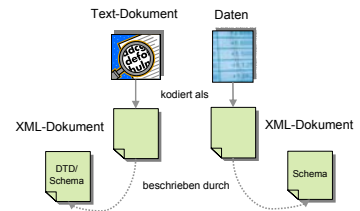
- Jedes XML-Dokument hat *genau ein* Wurzelement.
- Kind-Elemente geordnet

Dokument-Typen

- beschreiben generelle *Struktur* von Dokumenten/Daten, keine konkreten Inhalte
- beschreiben also eine Klasse von XML-Dokumenten
- können entweder mit DTDs (*Document Typ Definitions*) oder *XML-Schemata* spezifiziert werden.



DTDs vs. XML-Schemata



DTDs für Spezifikation von Text-Dokumenten ausreichend, XML-Schemata zur Spezifikation von Daten besser geeignet

Erweiterbarkeit

- X in XML steht für erweiterbar (engl. *extensible*).
- Was bedeutet Erweiterbarkeit?
- Vergleich mit HTML hilfreich:

HTML

- vorgegebene Auswahl an Sprachelementen
- Neues Sprachelement kann nur eingeführt werden, wenn sich das W3C auf eine neue HTML-Version einigt!

XML

- beliebige Elemente können benutzt werden
- Nur die Anwender des entsprechenden Elementes müssen sich auf eine gemeinsame Interpretation einigen.

Die XML-Familie: Der Kern

XML 1.0

- Syntax wohlgeformter XML-Dokumente
- Definition einfacher Dokument-Typen (DTD)

Namensräume

- gleichzeitige Verwendung unterschiedlicher Vokabularien in einem XML-Dokument
- z.B. Unterscheidung zwischen Titel einer Person vom Titel eines Buches

XML-Schema

- Definition komplexen Datentypen wie sie von Programmiersprachen bekannt sind

Der Rest der XML-Familie

Extensible Stylesheet Language (XSLT)

- Transformation von XML-Dokumenten in beliebige Text-Formate:

XML → HTML / WML / ASCII / XML / ...

Document Object Model (DOM)

- Parsen, Modifizieren und Erstellen von XML-Dokumenten

XPath

- Zugriff auf beliebige Teile eines XML-Dokumentes, wie z.B. die Nachnamen aller Autoren

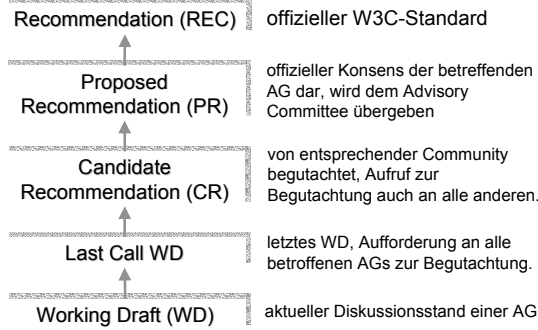
gesamte XML-Familie besteht aus lizenzfreien W3C-Standards

Exkurs: Das W3C



- 1994 als Projekt am MIT gegründet
- keine Normierungsorganisation im klassischen Sinn
- kann Einhaltung von Normen *nicht* auf rechtlichem Wege einklagen
- definiert deshalb lediglich Empfehlungen (*recommendations*)
- W3C-Recommendations sind lizenzfrei.

Standardisierungsprozess des W3C



© Klaus Schild, 2004

37

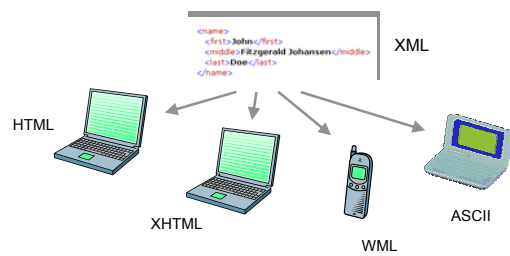
Anwendungen von XML

- Trennung von Inhalt und Präsentation
- anwendungsspezifische Standards
- Web-Dienste (Web Services)
- Semantic Web

© Klaus Schild, 2004

38

Trennung von Inhalt und Präsentation



Inhalt auf verschiedenen Endgeräten mit unterschiedlichen Bandbreiten unterschiedlich darstellen (*Multi Delivery*)

© Klaus Schild, 2004

39

Anwendungsspezifische Standards

- **XHTML**: Reformulierung von HTML in XML
- **WML**: Wireless Markup Language
Präsentation von Inhalten auf mobilen Endgeräten
- **DocBook**: strukturierte Darstellung von Büchern/Artikel
- **MathML**: Mathematical Markup Language
Standard für mathematische Ausdrücke
- **SVG**: Scalable Vector Graphics
Standard für Vektorgraphiken
- **SMIL**: Synchronized Multimedia Integration Language
Standard für Multi-Media-Anwendungen
- **VoiceXML**: Voice Extensible Markup Language
Standard für interaktive Sprachanwendungen

unterschiedliche Anwendungen, einheitliche Syntax!

© Klaus Schild, 2004

40

Was sind Web-Dienste?

© Klaus Schild, 2004

41

Was sind Web-Dienste (Web Services)?

traditionelle Web-Anwendung



Web-Dienst (Web Service)



© Klaus Schild, 2004

42

Beispiel: Google ohne Browser



With Google Web APIs, your computer can do the searching for you.

- Google gibt es auch als Web-Dienst: Suche und Rechtschreibkorrektur.
- Anwendungsprogramm sendet Google eine SOAP-Nachricht.
- Google antwortet wiederum mit SOAP-Nachricht.

Beispiel: Google ohne Browser



Google kann also aus einem Anwendungsprogramm heraus aufgerufen werden, um z.B.:

- in periodischen Abständen zu einem bestimmten Thema nach neuen Informationen zu suchen
- automatisch neue Trends im WWW zu identifizieren
- die Rechtschreibkorrektur von Google zu nutzen



Eigenschaften von Web-Diensten



- implementieren keine *neuen* Systeme
- Fassade für bestehende Systeme, um diese einfach zuzugreifen
- nutzen gängige Internet-Protokolle wie HTTP(S), SMTP und FTP
- verwenden XML-Standards SOAP und WSDL
- unabhängig von Programmiersprachen und Betriebssystemen

Was ist das Semantic Web?



Semantic Web



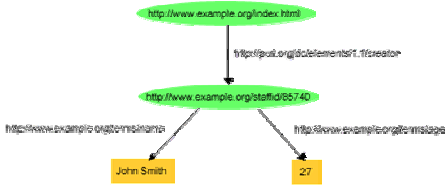
- heutige Webinhalte (Hypertexte) für Mensch-Maschine-Kommunikation ausgelegt
- Berner-Lees` Vision vom Semantic Web:
 - Webinhalte werden auch für Maschinen verständlich
 - dadurch können Computer komplexe Anfragen beantworten, die heute nur durch manuelles Surfen zu beantworten sind
 - Beispiel: Wie alt ist der Autor eines bestimmten Dokumentes?

Semantic Web



- XML: erster Schritt zum Semantic Web: Dokumente → maschinenverarbeitbare Daten
- Link-Struktur des heutigen Webs muss noch für Maschinen verständlich gemacht werden
- hierfür wurde RDF entwickelt

Ressource Description Framework



- maschinenverständliche Link-Typen (durch URIs identifiziert)
- verknüpfen zwei beliebige Web-Ressourcen

Literatur



Klausurrelevante Literatur

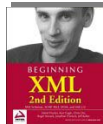


XML, Namensräume, XSLT

- Hunter et al., Beginning XML (2nd Edition), Wrox Press, 2001: S. 29-147.
- Semesterapparat der Fachbereichsbibliothek

XML-Schema

- <http://www.w3.org/TR/xmlschema-0/> oder
- S. 217-288 aus Hunter et al. (2001).



XML-Parser

- <http://java.sun.com/webservices/docs/1.2/tutorial/doc/>: Kap.1, "The SAX API" und "The DOM API".

Web-Dienste

- <http://www.w3.org/TR/soap12-part0/>
- <http://www.w3.org/TR/wsdl>

Wie geht es nächste Woche weiter?



- Organisatorisches
- Einführung und Überblick
- Klausurrelevante Literatur
 - XML-Syntax
 - Namensräume
 - Semantik von XML-Elementen

Beispiel



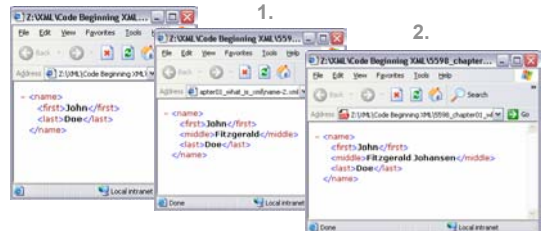
```
<location>
<latitude>32.904237</latitude>
<longitude>73.620290</longitude>
<uncertainty units="meters">2</uncertainty>
</location>
```

XML-Schema

DTD

- **Ortsangabe:** besteht aus Breitengrad, Längengrad und Unsicherheit der beiden Angaben.
- **Breitengrad:** Dezimalzahl zwischen -90 und +90.
- **Längengrad:** Dezimalzahl zwischen -180 und +180.
- **Unsicherheit:** eine nicht-negative Zahl.
- **Einheit für Unsicherheit:** entweder Meter oder Fuß

Beispiel für Erweiterbarkeit



1. Neues Element kann jederzeit eingeführt werden (verallgemeinerte Auszeichnungssprache).
2. Ein Element kann *unabhängig* von anderen Elementen erweitert werden (strikte Element-Struktur).