Kodierung strukturierter Dokumente: XML und XML-Technologie

Anne Brüggemann-Klein

Technische Universität München

Fakultät für Informatik

Organisatorisches zum Praktikum

Präsenztermin: Mittwochs, 10.15-12.00 Uhr, 01.07.023

Teilnahmeliste, Anmeldung in TUMonline (nur für bestätigte Studierende nutzen)

Format

- Gruppenarbeit (2-4 Personen pro Gruppe)
- Erarbeitungsphase (ca. 8 Termine am Mittwoch bis Mitte Dezember)
  - Technologieeinführungen mit tutoriellem Charakter
  - kein Praktikum am 2. November 2016 wegen FVV
  - Miniprojekt zur Demo (GuessANumber)
  - Übungsaufgaben mit Teilaufgaben zum Projekt, Gruppenpräsentationen mit Bonusregelung
- Projektphase (Termine am Mittwoch bis Semesterende nach Bedarf, Coaching)
  - Projekt Mancala oder Spiel nach Wahl
  - Abschluss mit benotetem Testat (Termin individuell pro Gruppe vereinbar)

Kommunikation außerhalb der Präsenztermine: Moodle, E-Mail

In eigener Sache: Geburtstagskolloquium

Die Fakultät für Informatik der Technischen Universität München lädt ein zum
60. Geburtstag von Prof. Dr. Anne Brüggemann-Klein

31. Oktober 2016 um 17 Uhr im TUM Institute for Advanced Study

16:00 Uhr	**Sektempfang** im Faculty Club (IAS)
17:00 Uhr	**Begrüßung** durch **Hans-Joachim Bungartz,** Dekan Fakultät für Informatik
17:15 Uhr	**Laudatio** von **Dr. Eva Sandmann,** Frauenbeauftragte der Technischen Universität München
17:30 Uhr	**Festvortrag** von **C. M. Sperberg-McQueen**, PhD, Black Mesa Technologies LLC *„30 Jahre SGML – 30 Jahre deskriptive Textauszeichnung“*
18:30 Uhr	**gemütlicher Ausklang**

Info unter www.in.tum.de/brueggemann-klein60
Zusage bitte an zusage@in.tum.de

Gesucht: 2 "Mundschenke" – HIlfe beim Ausschenken von Getränken
50 € pro Person, Meldung an brueggem@in.tum.de

Der rote Faden

Vogelperspektive: Elektronisches Publizieren, das Web, Document Engineering und XML-Technologie

Modell der strukturierten Dokumente

Technischer Einstieg XML, DTD, CSS

Projekte im Praktikum

Literatur

Elevator Speech XML-Technologie …

XML: Sprache zur Kodierung strukturierter Information (Text, Daten)

- niederschwelliges Klartext-Format
- flexible Strukturierung
  - mächtiger als reiner Text
  - flexibler als Tabellenstruktur (Datenbank)
  - zu Hause in jeder Domäne
- standardisiert und austauschbar
  - W3C Recommendation
  - basierend auf Unicode

… *Elevator Speech* XML-Technologie

XML-Technologie: standardisierte Sprachen und Werkzeuge
zur flexiblen Verwendung, Bearbeitung und Definition dieser Information

- transformieren (mehrfach verwenden), abfragen
- erfassen
- modellieren

➢ Full-stack Toolbox für Web-Anwendungen (*Look Ma, no Frameworks*),
Technologie der Wahl für Domänen-Expert/innen mit XML-Expertise

Was leistet XML-Technologie für Web-Anwendungen?

Kodierung von Daten und Dokumenten (XML, XML Namespaces, DocBook, SVG, ePUB, XLink)

Beschreibung von Dokumentenklassen (XML Schema)

Extraktion von Information (XQuery)

Kombination und Weiterverarbeitung (XSLT)

Formatierte Präsentation (XSL)

Datenerhebung mit Formularen, GUI (XForms, XHTML, CSS)

Aufbau von Web-Anwendung mit XRX-Architektur (XForms, REST, XQuery / XSLT)

Spezifikation von Verhalten (SCXML)

- Full-stack Toolbox für Web-Anwendungen (*Look Ma, no Frameworks*),
  Technologie der Wahl für Domänen-Expert/innen mit XML-Expertise

Abgrenzung Vorlesung / Praktikum

XML-Technologie ist Implementierungs-Technologie

Modelle, Konzepte, Grundsätze etc. weitgehend
in Vorlesung "Elektronisches Publizieren"

Vorgehensweisen, Architekturen, Praxis, Implementierungen etc. weitgehend
im Praktikum "XML-Technologie"

Wegen grundlegender Bedeutung für XML-Technologie heute auch im Praktikum:
Grundsätzliches zum Modell der strukturierten Dokumente

Modell der strukturierten Dokumente

Modell der strukturierten Dokumente …

Etablierte Methode der Informatik: Suche nach einem dem Problembereich angemessenem Modell [Aufbau, Konstruktionsprinzip, Bildungsmuster]

Im Document Engineering etabliertes Modell: Modell der strukturierten Dokumente

- entwickelt in den 80er Jahren
- entstanden in den Bereichen Verlagswesen und technische Dokumentation
- auch als Grundlage für Informationsverarbeitung auf Basis von Dokumenten geeignet (XML-Datenbanken)

Grundlage von XML-Technologie

- zur Kodierung und
- zur Bearbeitung strukturierter Dokumente

… Modell der strukturierten Dokumente …

Beispiel: Formatierung

- Ursprünglich: Direkte Formatierung (hier Kursivschrift) im Dokument

  Bitte hinterlassen Sie eine *message*! [Grund: andere Sprache]
  Bitte nur *einmal* anklicken! [Grund: Betonung]
  Wir nennen das eine *Pipette*. [Grund: Fachbegriff]

- Modell der strukturierten Dokumente:

  (1) Markierung mit dem Grund direkt im Text (logische Struktur)

  Bitte hinterlassen Sie eine <foreign lang="en">message</foreign>
  Bitte nur einmal anklicken!
  Wir nennen das eine <term>Pipette</term>

  (2) Tabelle mit Vorschriften zur Umsetzung (Stylesheet)

  foreign → setze kursiv
  em → setze kursiv
  term → setze kursiv

  (3) Kombination von logischer Struktur und Stylesheet: Software

… Modell der strukturierten Dokumente …

Beispiel: Nummerieren und Referenzieren

- Ursprünglich: Direkte Nummerierung/Referenzierung im Dokument

  3 Ernährung …

  7 Sport

  In diesem Kapitel nehmen wir einiges aus Kapitel 3 wieder auf…

- Modell der strukturierten Dokumente:

  (1) Markierung mit dem Grund direkt im Text (logische Struktur)

  <heading id="h.ern">Ernährung</heading> …

  <heading id="h.sp">Sport</heading>

  In diesem Kapitel nehmen wir einiges aus Kapitel <ref id="h.ern"/>

  wieder auf …

  (2) Tabelle mit Vorschriften zur Umsetzung (Stylesheet)

  heading → setze fett, nummeriere durchgehend

  ref → füge Nummer des referenzierten Elements ein

  (3) Kombination von logischer Struktur und Stylesheet: Software

… Modell der strukturierten Dokumente …

Idee: Ersetze im ursprünglichen Dokument alle Information, die mit Verwendungszweck zu tun hat,

- durch den inhaltlichen (semantischen) Kern dieser Information und
- durch (separierbare, wiederverwendbare) Vorschriften zu ihrer Bearbeitung

Ergebnis entspricht Modell der strukturierten Dokumente

… Modell der strukturierten Dokumente …

Zwei Bestandteile eines strukturierten Dokuments

- Text-Inhalt
- Annotation von Textbereichen mit semantisch bedeutsamer Metainformation: (logische) Struktur

Kombinierbar mit separater, auswechselbarer Formatvorlage (Stylesheet) für formatierte Präsentation

Kombinierbar mit weiteren separaten, auswechselbaren und ausführbaren Bearbeitungsvorschriften für Verarbeitung bis hin zu semantischer Interpretation

➢ Unterstützung semantischer Verarbeitung im Grenzbereich Dokumente ◁▷ Daten

… Modell der strukturierten Dokumente

Vorteile des Modells

- Flexibilität [setze Fachbegriffe fett statt kursiv]
- Stabilität bei Veränderungen [füge neues Kapitel ein]
- Konsistenz [ein Stylesheet für verschiedene Dokumente]
  - n : 1 Verhältnis Dokumente : Stylesheets
- Single Source [alternative Stylesheets für ein Dokument]
  - 1 : n Verhältnis Dokumente : Stylesheets
- Verarbeitbarkeit [liste alle Fachausdrücke für Glossar auf]

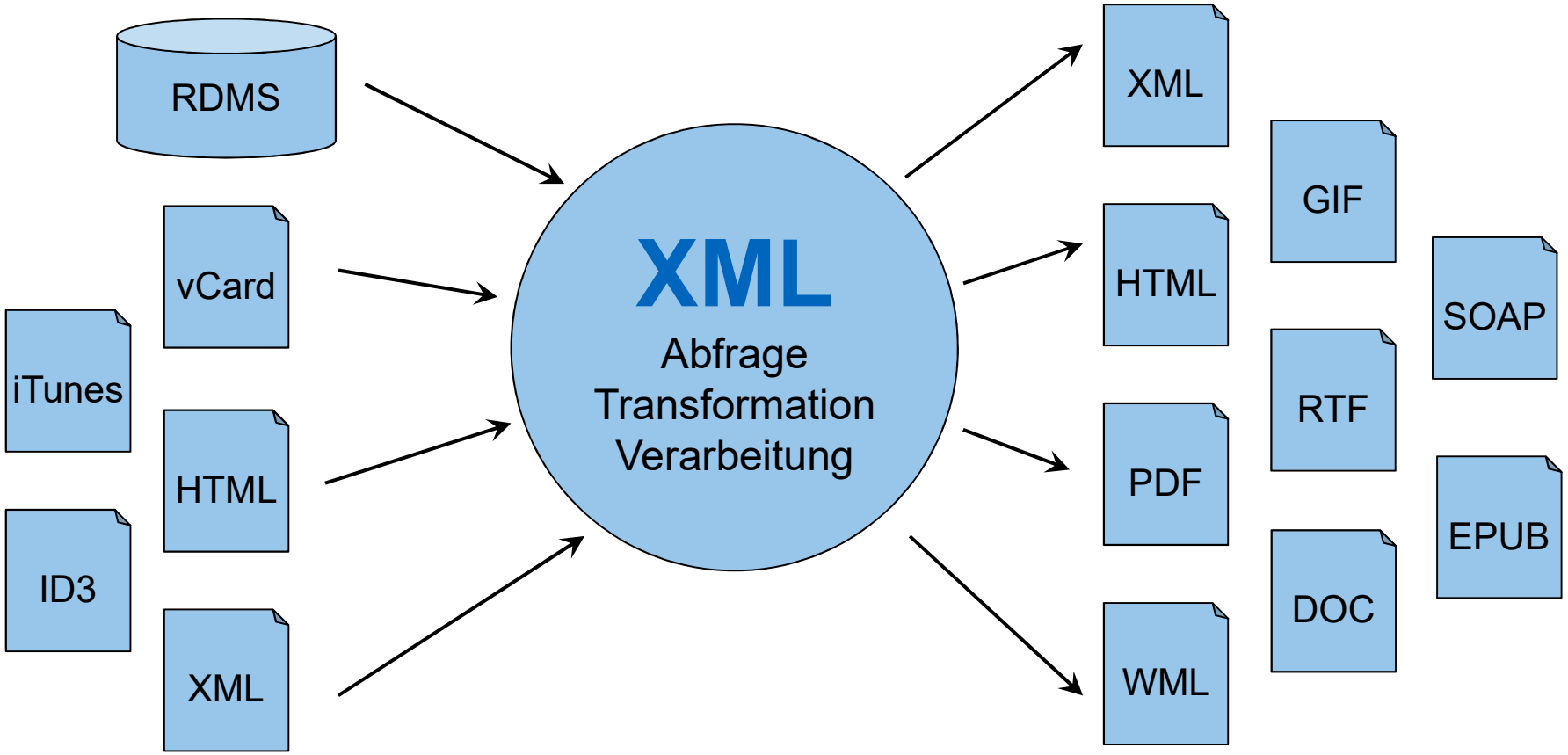
Schlüsseltechnologie zur Umsetzung: XML …

Sprache zur **Kodierung** von strukturierten Dokumenten

Standardisiertes **Austauschformat** für Dokumente
(Plattformunabhängigkeit von XML)

Zentrales Format (**Drehscheiben-Charakter**)

Drehscheibencharakter von XML



… Schlüsseltechnologie zur Umsetzung: XML

Zentrales Format (**Drehscheiben-Charakter**)

- **Import** aus heterogenen Datenquellen, auch Content Syndication
- **Export** in verschiedene Kanäle, Medien, Präsentationsformate (Cross-Media, Multi-Channel)
- automatisierte und computergestützte **Verarbeitung**
  bis hin zu semantischer Interpretation

Unabhängigkeit von Quell- und Zielformaten durch Drehscheiben-Format XML:
n+m statt n·m viele Beziehungen

Satelliten-Technologien

Anwendung

Aus dem früheren Praktikumsprojekt XTunes

- Datenbestand (kodiert mit XML, XLink)
  - Information zu Musiktiteln, KomponistInnen, Werken, ausführenden KünstlerInnen
- Anwendungen (Daten aufbereitet mit XQuery, XSLT; Daten präsentiert mit HTML, SVG)
  - Timeline aller Violinkonzerte mit Links zu Aufnahmen
  - Jubiläumskalender

Unterstützung semantischer Verarbeitung im Grenzbereich Daten-Dokumente

Technischer Einstieg: XML, DTD, CSS

XML (Extensible Markup Language) am Beispiel

Dokumentenformat des W3C für Webdokumente

Markupsprache: Markiert Textbereiche an Anfang und Ende mit Tags (attributierte Elemente):
..\..\compProp\KonzeptUTF8.xml, ../../compProp

Tag-Syntax wie bei HTML: `<eName aName="…">…</eName>`
mit Variante `<eName aName="…"/>` ohne Inhalt

Tags müssen immer geschlossen werden (anders als bei HTML): Wohlgeformtheit

Zusammengehörige Tags markieren attributierte Elemente

Referenzen `&refName;` inkludieren Sonderzeichen (wie in HTML)
oder sogar ganze Teildokumente

Freies Vokabular für Elementnamen, Attributnamen, Referenzen,
formal definierbar in Document Type Definition (DTD): ../.. /compProp/Konzept.dtd

Formatierung von XML-Dokumenten: ../../compProp/KonzeptXML.css

Erster Steckbrief für XML

HTML mit freien Elementen, Attributen, Referenzen
▷ XML (ohne DTDs)

XML als Syntax für strukturierte Dokumente [Metasprache für Markupsprachen]:
Standardisiertes allgemeines Format für moderne Dokumente zur Kodierung von Text und semantischen Rollen

XML-Anwendung, Markupsprache: spezifisches Vokabular/Format für spezifische Anwendungen
Beispiele: XHTML, TEI XML, MathML, CML, Persistente Objektrepräsentation (Java), B2B (EDI), ECommerce (OTP), Topic Map Markup Language, XSL-FO, EPUB, SVG, Konfigurationen, Metadaten, DocBook

XML-Syntax: Dokument (Instanz) …

XML-Deklaration

```
<?xml version="1.0" encoding=""
    standalone=""?>
```

Element mit Attributen, kodiert durch Tags

```
<eName aName1="w1" ... aNamex="wx">
    Element-Inhalt
</eName>
<eName aName1='w1' ... aNamex='wx'/>
```

Einschränkungen zu Alphabet (alphanumerisch plus '-', '_', '.', ':') und Syntax für Namen (Elemente, Attribute etc.)

Ausschluß bestimmter Funktionszeichen für Attributwerte und textuellen Inhalt

… XML-Syntax: Dokument (Instanz)

Mixed Content (erzählerische Dokumente vs. Daten)

Zeichen: Attributwerte und textueller Inhalt

- Direkte Code-Position
- *Character Reference*
  `&#Dezimalzahl;` oder `&#xHexzahl;`
- Vordefinierte *Entity References* für Funktionszeichen: `<`, `>`, `&`, `&apos;`, `&quot;`
- Deklarierte *Entity References*

CDATA *Section*
`<![CDATA[InhaltOhne']]>']]>`
Markup im Innern einer CDATA Section wird nicht erkannt; nicht schachtelungsfähig

White Space

Konzepte und Werkzeuge

XML, DTD, CSS

Encoding

Wohlgeformtheit und Validität

Parser

Editor, Entwicklungsumgebung

Browser

Wohlgeformtheit und Validität

XML-Dokumente müssen wohlgeformt sein
(Überprüfung durch Parser, XML-Prozessor)

- DTD optional
- korrekte Klammerung
- keine mehrfachen Attribute in Elementen

Optional: Validität
(Überprüfung durch validierende Parser)

- Überprüfung der Instanz gegenüber DTD

Parser …

Programme zur XML-Syntaxüberprüfung heißen Parser

Aufgaben eines Parsers

- Syntaxüberprüfung
- Abstraktion von "syntaktischem Zucker"
- Überprüfen der Wohlgeformtheit
- Auswertung von DTD (optional)
  - Expansion von Referenzen
  - Einsetzen von Attributwerten
  - Validierung
- Weitergabe der erforderlichen Information an Anwendung (SAX-API: Strom von Events)

… Parser

Parser-Software Xerces von Apache XML unter http://xerces.apache.org/,
zum Beispiel als Java-Implementierung Xerces-J

- xercesImpl.jar, xml-apis.jar und xercesSamples
  im Classpath
- Aufruf `java sax.Counter [-v] XMLDocument`
- implementiert SAX-API

Online-Parser

Integrierter Parser in Browsern (keine Validierung)

ntegrierter Parser in allgemeinen oder XML-spezifischen Entwicklungsumgebungen,
z.B. oXygen, Eclipse

Bestandteil von APIs für Programmiersprachen,
z.B. Java, Python

➢ Parser sind allgegenwärtig und werden eher im Hintergrund benutzt

Web-Browser

Moderne Web-Browser können XML-Dokumente darstellen

- quelltext-nah, aber mit Einrückungen und klappbaren Strukturen
- formatiert, nach Vorgaben eines CSS-Stylesheets formatiert

Anbindung eines Stylesheets in XML-Dokument mit Processing Instruction
```
<?xml-stylesheet type="text/css" href="KonzeptXML.css"?>
```
[Associating Style Sheets with XML documents, W3C Recommendation 29 June 1999]
[http://www.w3.org/TR/xml-stylesheet]

Beispiel (..\..\compProp)

Editieren von XML-Dokumenten

Editieren von XML mit Texteditoren möglich (WordPad)

Spezialisierte XML-Editoren und Entwicklungsumgebungen: XML Spy, oXygen, XMetaL, Eclipse mit folgender Funktionalität:

- Syntax-Highlighting, Syntaxvervollständigung
- Navigation in Element-Hierarchie
- nach Wahl formatierte Darstellung
- Überprüfen der Wohlgeformtheit, Validierung
- evtl. Integration von Werkzeugen (XSLT, XQuery) und Debuggern
- spezielle Visualisierungen
  - tabellenähnliche Dokumente
  - DTDs und Schema-Dokumente

Projekte im Praktikum

Projekte im Praktikum

Projekte

- Durchgängiges Beispiel im Praktikum: GuessANumber

Umgesetzt als Web-Anwendungen rein mit XML-Technologie

Anwendung von bewährten Architekturen und Prinzipien des Software Engineering zur Sicherstellung von Qualitätskriterien

- Modell-getriebene Ansätze (Domain-Driven Design)
- Objekt-orientierter Programmierstil (Kapselung von Daten und Methoden)
- Architekturstil Model-View-Controller
- Herausforderung: Observer-Pattern im Request-Response-Zyklus von Web-Anwendungen

Literatur

Literatur zum Einstieg …

Skript (PDF in Moodle, Ausdrucke über Fachschaft)

White papers

- J. Bosak: *XML, Java and the Future of the Web*.
  http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm.
- J.P. Morgenthal: *Portable Data / Portable Code: XML & Java Technologies*.
  http://java.sun.com/xml/ncfocus.html.
- T. Berners-Lee u.a.: *The Semantic Web*.
  Scientific American 2001.

… Literatur zum Einstieg …

Textbücher

- A. Moller, M. Schwartzbach: *An Introduction to XML and Web Technologies*. Addison-Wesley 2006.
- J. Fawcett, L. Quin, D. Ayers: Beginning XML. Wrox 2012.
- E.R. Harold u.a.: *XML in a Nutshell*. O'Reilly 2001.
- G. Kappel u.a.: *Web Engineering*. DPunkt 2003.

Grenzgängerische Lektüre

- D. Levy: Scrolling Forward. Making Sense of Documents in the Digital Age. Arcade Publishing 2001.
- S. Abiteboul, P. Buneman, D. Suciu: *Data on the Web*. Morgan Kaufmann Publishers 2000.

… Literatur zum Einstieg

Online-Quellen

- XML-Seite des W3C (http://www.w3.org/XML/).
- Online-Tutorial von W3Schools. (http://www.w3schools.com/xml/).
- Kurs XML der TEIA-AG.
  http://www.teialehrbuch.de/XML/
- XML-Portal von O‘Reilly (http://www.xml.com/).