XML and Content Management — Intro

Heinz Wittenbrink

4. April 2017

Was ist XML?

XML-Basics

- XML: Extensible Markup Language
- Weiterentwicklung von SGML
- Lange als Alternative zu HTML verstanden
- Syntax für XML-Vokabulare
- W3C-Standard
- Basis für eine Familie von weiteren Standards und Technologien

XML-Syntax

- Spitzklammern unterscheiden Markup und Textinhalt
- Wichtigste syntaktische Features: Elemente, Attribute, Kommentare
- Wohlgeformtes XML folgt syntaktischen Regeln: Elemente sind geschlossen und - ineinander eingebettet, Attribute haben Werte in Anführungsstrichen, Groß und Kleinschreibung sind bedeutungsunterscheidend
- Valides XML entspricht einem definierten Dokumenttyp

Beispiel für ein XML-Dokument

```
<?xml version="1.0"?>
<!-- dictionary.xml
 - Copyright (c) 2014, HerongYang.com, All Rights Reserved.
-->
<dictionary>
 <word acronym="true">
  <name>XML</name>
  <definition reference="Herong&amp;apos;s Notes">eXtensible Ma
Language.</definition>
  <update date="2002-12-23"/>
 </word>
 <word symbol="true">
  <name>&amp;lt;</name>
  <definition>Mathematical symbol representing the "less th
operation, like: 1&amp;lt;2.</definition>
  <definition>Reserved symbol in XML to representing the be
tags, like: <![CDATA[<p>Hello world!</p>]]>
  </definition>
```

XML: Wichtige Eigenschaften und Begriffe

- Dokumente bestehen aus ineinander eingebetteten Elementen. Sie sind “Bäume”
- Es gibt ein und nur ein hierarchisch oberstes Element, das Dokument- oder Wurzelelement.
- Elemente dürfen sich nicht überlappen.
- Die XML-Syntax ist die Serialisierung einer Hierarchie.
- Attribute müssen einen Wert haben. Sie stehen gleichgeordnet bei einem Element.

Valides XML

- Valides XML entspricht den Regeln eines Dokumenttyps oder XML-Vokabulars
- Praktisch verwendete XML-Dokumente gehören fast immer zu einem Vokabular
- Validierung ist die formale Überprüfung der Regelkonformität eines Dokuments
- Dokumenttypen können formal u.a. in Dokumenttyp-Definitionen (DTDs) oder durch XML- bzw. Relax NG-Schemas definiert werden

XML-Verarbeitung

- Bei der XML-Verarbeitung entnimmt ein Parser dem Dokument die relevanten Informationen
- Dieser Prozess ist standardisiert, wobei es unterschiedliche Parser gibt XML-Parser stellen die Verarbeitung ein, sobald ein Dokument nicht wohlgeformt ist

Unterschiede XML und HTML

- XML ist case-sensitive, HTML nicht
- XML-Elemente müssen geschlossen sein, HTML-Elemente nicht
- XML-Attribute müssen einen Namen und einen Wert haben
- Die Dokumenttyp-Deklaration unterscheidet sich, siehe https://www.w3.org/QA/2002/04/valid-dtd-list.html
- HTML-Parser sind toleranter als XML-Parser

XML-Vokabulare im Web

- XHTML und (https://www.w3.org/TR/xhtml1/ und https://www.w3.org/TR/xhtml11/) Scalable Vector Graphics SVG (https://www.w3.org/Graphics/SVG/)
- MathML (https://www.w3.org/Math/)
- RSS (https://validator.w3.org/feed/docs/rss2.html) und Atom (https://validator.w3.org/feed/docs/atom.html)
- Epub (http://idpf.org/epub)

XML-Vokabulare für strukturierte Dokumentation

- DocBook (http://www.docbook.org/)und DITA (http://dita.xml.org/) sind wichtige Formate für das Single Source Publishing
- DocBook wurde als buchorientierter Dokumenttyp entwickelt
- DITA ist eine Topic-orientierte Alternative, die sich am Prinzip des didaktischen Minimalismus orientiert

Office-Formate

- Microsoft Office und Libre Office/Open Office verwenden XML als natives Format
- Office Open XML ist ein (umstrittener) ECMA-Standard (http://www.ecma-international.org/publications/standards/Ecma-376.htm)
- Das Open Document Format (https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office) ist XML basiert und ein OASIS-Standard
- Libre Office unterstützt auch DocBook

Entwicklung der Markup-Sprachen

- Basis: SGML, bereits seit den 1960er Jahren entwickelt
- HTML wurde als SGML-Anwendung entwickelt
- XML entstand in den 1990er Jahren als Alternative zu HTML und vereinfachte - Version von SGML
- HTML5 setzte sich gegen XHTML als lebender HTML-Standard durch.
- Parallel etablierte sich JSON als einfaches Format für den Datenaustausch im Web

Transformation und Präsentation von XML

- XML erlaubt eine strikte Trennung von Inhalt und Präsentation
- XML-Dokumente werden meist mit XSLT in andere XML-Vokabulare oder HTML überführt
- XSLT ist eine funktionale Programmiersprache, die die XML-Syntax verwendet Für die seitenorientierte Ausgabe wurde XSL-FO entwickelt

Microdata

- Als Microdata können maschinenlesbare Informationen in HTML-Dokmente eingebettet werden (https://www.w3.org/TR/microdata/)
- RFFa und Jason sind Alternativen
- Microformats haben eine ähnliche Zielsetzung, basieren aber auf älteren HTML-Mechanismem
- Empfehlenswert ist die Orientierung an Schema.org

RDF/RDFa

- RDF ist ein Standard für semantische Informationen im Web
- RDF kann als XML, aber auch in anderen Formaten serialisiert werden
- RDF wurde als Basistechnologie für das Semantic Web entwickelt
- Mit RDFa können RDF-Daten in HTML-Dokumente eingebettet werden
- Das von Facebook entwickelte Open Graph Protocol (http://ogp.me/) basiert auf RDFa

XML und Content Management

Task: