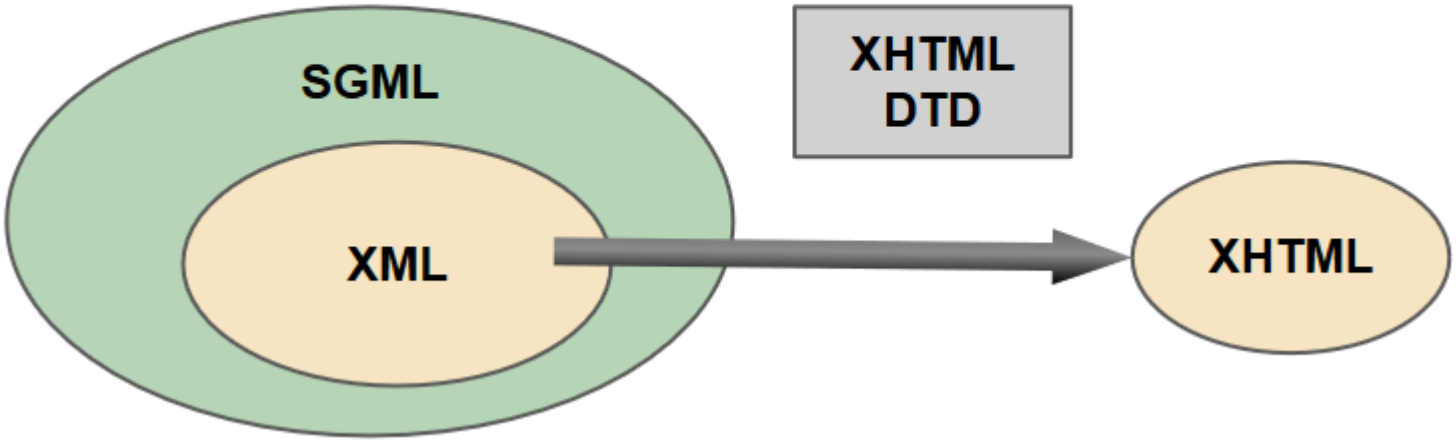
Von SGML über XML zu HTML

SGML, XML und HTML

Auszeichnungssprache **Anwendung**



Was ist SGML?

- Die Standard Generalized Markup Language (SGML) ist als Meta-Sprache dazu gedacht, neue Auszeichnungssprachen für den Dokumenten-austausch zu definieren.
- Insofern ist SGML also gar keine Sprache im engeren Sinne, sondern ein System um Sprachen zu erstellen.
  - Es wurde 1986 von der ISO standardisiert (ISO 8879:1986).
- Die grundlegende Idee einer Meta-Sprache ist es, sich für einen bestimmten Anwendungszweck eine Sprache zusammenzustellen.
- Hierbei liegt der Fokus von SGML auf Sprachen für das Erstellen von großen und umfangreichen Dokumenten, z.B. der Dokumentation für eine komplexe technische Anlage.
- Der große Vorteil von SGML liegt darin, dass die Dokumente als reine Textdokumente vorliegen und somit leicht mit beliebigen Werkzeugen wie Texteditoren bearbeitet werden können.

Was ist SGML?

- Auszeichnungen im Text werden bei SGML-basierten Sprachen durch spezielle Markierungen (Tags) vorgenommen, die z.B. Zitate kennzeichnen oder Querverweise zu anderen Dokumenten markieren.
- SGML trifft als Meta-Sprache allerdings keine Aussagen darüber, welche Tags in einem Dokument vorkommen dürfen sondern legt nur die Art fest, wie man die Tags definieren kann.
  - HTML ist eine Sprache, die mit Hilfe von SGML definiert wurde und die einen genau festgelegten Satz von Tags besitzt. In der Terminologie von SGML sagt man auch, dass HTML eine SGML-Anwendung ist.

Was ist SGML?

- Leider hat SGML einige Eigenschaften, die eine maschinelle Verarbeitung von SGML-Dokumenten erheblich erschweren:
  - Es gibt z.B. eine ganze Reihe von syntaktischen Freiheiten und Abkürzungen, die für menschliche Autoren durchaus angenehm sein können, bei der Verarbeitung durch ein Programm aber eine unnötige Komplexität erzeugen.
  - Die Freiheiten machen auch den SGML-Standard unnötig umfangreich (ca. 500 Seiten) und schwer zu verstehen.
  - Aus diesem Grund hat sich SGML außerhalb des HTML-Umfelds nie in breiter Front durchsetzen können.

Was ist XML?

- Inspiriert vom Erfolg von HTML und den Schwächen von SGML begann das World Wide Web Consortium (W3C) Ende der 1990er Jahre damit, eine Untermenge von SGML zu definieren, die viele der unnötigen Freiheiten aus SGML entfernt aber trotzdem die Möglichkeit bietet, beliebige Sprachen zu definieren und damit eine vollwertige Meta-Sprache ist.
- Diese Anstrengungen mündeten 1998 in die erste W3C-Recommendation zur Extensible Markup Language (XML).
- Dadurch, dass XML eine Untermenge von SGML ist, konnten die Werkzeuge und Technologien von SGML für XML weiterverwendet werden.
- Inzwischen hat sich die Situation allerdings insofern geändert, dass SGML keine Bedeutung mehr hat und ausschließlich XML zum Einsatz kommt.

Was ist XML?

- XML also ist eine Untermenge von SGML.
- XML lässt dabei alles weg, was die maschinelle Verarbeitung erschwert; es ist also eine strengere Syntax als SGML.
- Auch XML ist keine Sprache, sondern eine Technologie, um Sprachen zu definieren. Es ist also eine
  - Metasprache oder eine
  - Inhaltsbeschreibungssprache.
- XML-Dokumente können selbstbeschreibend sein.
- Ein XML-Dokument besteht aus Textzeichen und ist damit menschenlesbar.

Was ist wohlgeformtes XML?

- Das Dokument besitzt genau ein Wurzelelement.
- Als Wurzelelement wird dabei das jeweils äußerste Element bezeichnet, z.B. <html> in XHTML.
- Alle Elemente mit Inhalt besitzen einen Beginn- und einen End-Auszeichner, z. B. <eintrag>Eintrag 1</eintrag>.
- Elemente ohne Inhalt können auch in sich geschlossen sein, wenn sie aus nur einem Auszeichner bestehen, die mit /> abschließt, z. B. <eintrag />.
- Die Beginn- und End-Auszeichner sind ebenentreu-paarig verschachtelt.
  - Das bedeutet, dass alle Elemente geschlossen werden müssen, bevor die End-Auszeichner des entsprechenden Elternelements oder die Beginn-Auszeichner eines Geschwisterelements erscheint.
- Ein Element darf nicht mehrere Attribute mit demselben Namen besitzen.

Was ist gültiges XML?

- Soll XML für einen genormten Datenaustausch verwendet werden, so sollte das Format der Tags mittels einer Grammatik definiert sein.
- Eine mögliche Definition für eine solche Grammatik ist
  - eine Dokumenttypdefinition (DTD) oder
  - ein XML Schema.
- Der Standard definiert ein XML-Dokument als gültig, wenn es wohlgeformt ist, den Verweis auf eine Grammatik enthält und das durch die Grammatik beschriebene Format strikt einhält.

Beispiel für ein XML-Dokument: standalone: also keine Grammatik

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<verzeichnis>
     <titel>Wikipedia Städteverzeichnis</titel>
     <eintrag>
          <stichwort>Genf</stichwort>
          <eintragstext>Genf ist der Sitz von ...</eintragstext>
     </eintrag>
     <eintrag>
          <stichwort>Köln</stichwort>
          <eintragstext>Köln ist eine Stadt, die ...</eintragstext>
     </eintrag>
</verzeichnis>
```

Was ist HTML?

- Bei der Hypertext Markup Language (HTML) handelt es sich um eine Anwendung von SGML.
- Man spricht von einer Auszeichnungssprache (markup language), weil bestimmte Bestandteile eines Dokumentes durch spezielle Tags ausgezeichnet werden, wobei das gesamte Dokument in reiner Textform vorliegt.
- Der Web-Browser, als Anzeigekomponente für die Dokumente, setzt die Auszeichnungen dann in eine visuelle Darstellung um.
- Die Möglichkeit verschiedene Dokumente miteinander zu verknüpfen (hypertext) oder Verweise zu anderen Stellen desselben Dokuments aufzunehmen ist eine andere Schlüsseleigenschaft von HTML.
  - Die Verknüpfungen werden hierbei durch Hyperlinks hergestellt.

Was ist HTML?

- Ursprünglich stand der Inhalt und nicht das Layout im Vordergrund.
  - In den ersten Versionen von HTML war das Layout und die Darstellung einzig und allein Sache des Browsers und nicht des Autors einer HTML-Seite.
  - Der Benutzer des Browsers konnte z.B. die Schriftart und Größe für die Anzeige frei festlegen; der Autor der Web-Seite hatte hierauf keinen Einfluss.
  - Zusätzlich konnte man noch Bilder zur Veranschaulichung des Textes einbinden, an Audio oder Video hat man damals schon wegen der begrenzten Bandbreiten nicht gedacht.

Was ist HTML?

- Für wissenschaftliche Veröffentlichungen war der ursprüngliche Ansatz von HTML gut geeignet.
- Für Unternehmenswebseiten oder E-Commerce-Anwendungen aber vollkommen unpassend.
  - Aus diesem Grund setzten sich sehr schnell Erweiterungen durch, die erlauben, das Layout und die Darstellung von HTML-Seiten bis in das kleinste Detail zu bestimmen.
    - Dies geschieht heute über Cascading Style Sheets (CSS).
  - Zusätzlich hielten interaktive Elemente mit Formularen und JavaScript Einzug in die Web-Seiten.

Was ist XHTML?

- Bei HTML handelt es sich um eine über SGML definierte Auszeichnungs-sprache.
- HTML ist also eine SGML-Anwendung.
- XML ist eine neue, schlanke Variante von SGML.
- Daher liegt die Idee nahe, HTML noch einmal neu
  - auf Basis von XML
  - anstatt von SGML

  zu definieren und somit eine neue und saubere HTML-Variante als XML-Anwendung zu definieren.

Was ist XHTML?

- Genau dies hat das W3C im Jahr 2000 mit der Standardisierung von XHTML, das HTML völlig neu mit Hilfe von XML definiert hat.
- XHTML ist syntaktisch hundertprozentig kompatibel zu wichtigen XML-Standardsprachen.
- Durch die gemeinsame syntaktische Grundlage auf der Basis von XML ist es auch möglich, das Auslesen und Verarbeiten über Programmiersprachen zu vereinheitlichen.
- XHTML ist also XML-gerechtes HTML.
- Der "Nachbau" von HTML 4.0 in XHTML 1.0 ist so weit gelungen, dass es in XHTML die gleichen Elemente, Attribute und Verschachtelungsregeln gibt wie in HTML.

Was ist HTML5?

- Das W3C hat 2014 die fertige HTML5-Spezifikation vorgelegt.
- HTML5 wird damit als Nachfolger von HTML4 die Kernsprache („core language“) des Webs.
- Sie ersetzt damit die Standards
  - HTML 4.01,
  - XHTML 1.0 und
  - DOM HTML Level 2.
- Sie bietet neue Funktionen wie Video, Audio, lokalen Speicher und dynamische 2D- und 3D-Grafiken, die von HTML4 nicht direkt unterstützt wurden und sich ohne HTML5 nur mit zusätzlichen Plugins (z. B. Adobe Flash) umsetzen ließen.
- Zukunftsweisend sind weiterhin neue Elemente, die eine verbesserte semantische Struktur ermöglichen.

Was ist HTML5?

- Die verschiedenen, aus Gründen der Abwärtskompatibilität vorhandenen unübersichtlichen DOCTYPEs wurden stark vereinfacht zu:

```
<!DOCTYPE html>
```

- Die Dokumenttyp-Deklaration ist ein aus SGML und XML übernommenes Konzept.
- Andere Dokumenttyp-Deklarationen sollten Sie nicht mehr verwenden.
- HTML5 hat sich bis heute vollständig durchgesetzt.