Semistrukturierte Daten

XML

Stefan Woltran
Emanuel Sallinger

Institut für Informationssysteme
Technische Universität Wien

Sommersemester 2014

XML

Was ist XML?

- XML steht für *Extensible Markup Language*
- Industriestandard des W3C (World Wide Web Consortiums)
- Syntax zur Beschreibung semistrukturierter Daten

Eigenschaften von XML

- Trennung von Struktur und Präsentation
- Erlaubt Spezifizierung anwendungsspezifischer Dokumenttypen
- als Datensaustauschformat sehr gut geeignet

XML will be the ASCII of the web – basic, essential, unexciting (Tim Bray)

Geschichte

- 1945: Hypertext
- 1969: GML
- 1986: SGML (ISO Standard)
- 1989: HTML (Tim Berners-Lee, CERN)
- 1994: W3C gegründet
- 1996: SGML Subset Arbeitsgruppe gegründet
- 1998: XML 1.0
- 1999: XSLT
- 2001: XML Schema
- ... laufend neue Industriestandards

HTML vs. XML

- HTML
  - Fix definierte Elementnamen
  - vor allem zur Präsentation bzw. Layout
  - Browser verarbeiten HTML fehlertolerant
  - verschiedenste Erweiterungen ("lebender" Standard HTML5)
- XML
  - Elementnamen haben keine vordefinierte Bedeutung
  - Metasprache für Markup Sprachen
  - Syntax muss strikt eingehalten werden
  - viele ergänzende Standards (Schema-, Abfragesprachen)

Was XML nicht ist...

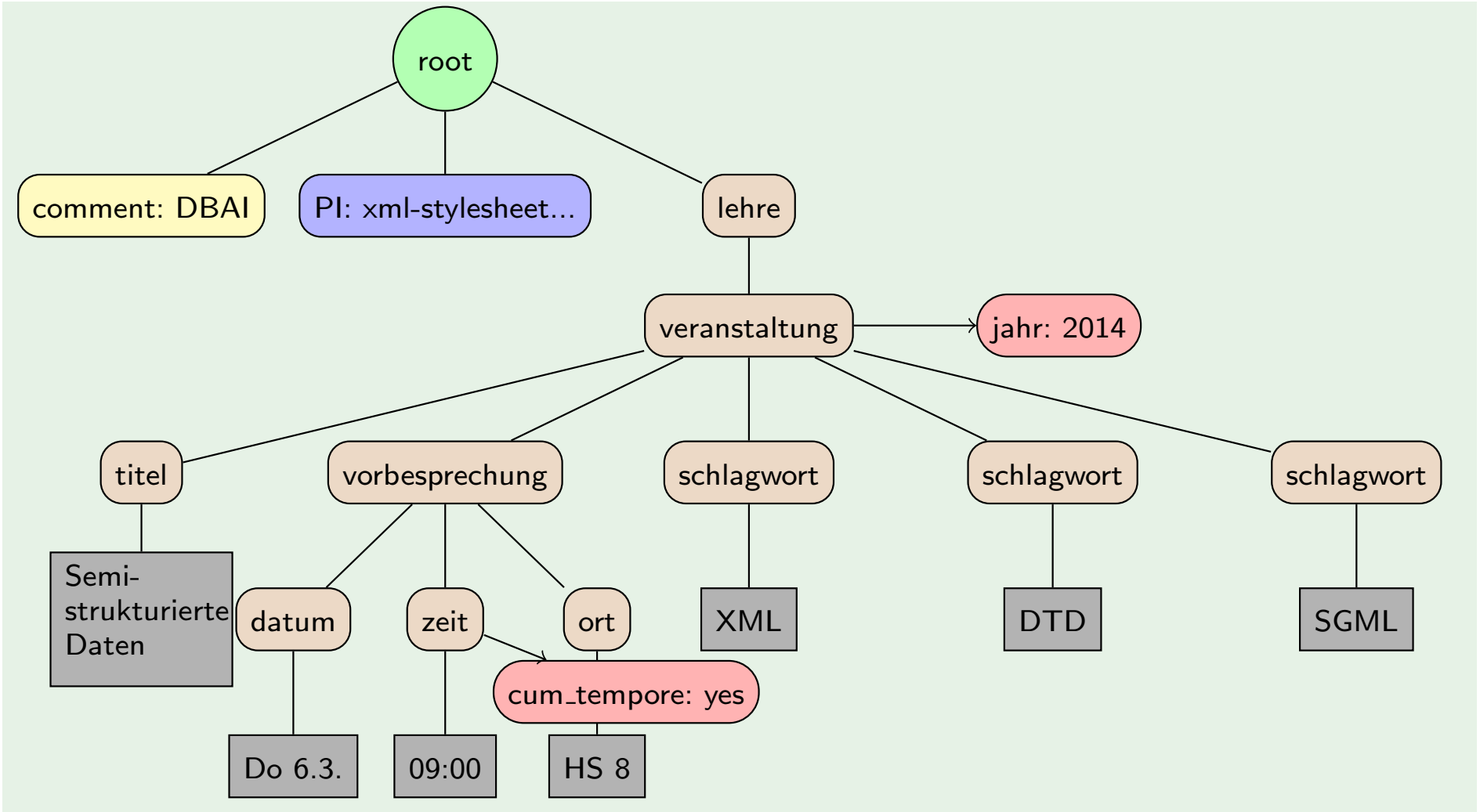
- XML ist keine Programmiersprache
- XML ist kein Netzwerkprotokoll
- XML ist keine Datenbank

Ein XML-Dokument *existiert* einfach. Es *tut* nichts.

Beispiel (XML Dokument)

```
<?xml version="1.0"?>
<!-- DBAI -->
<?xml-stylesheet type="text/css"href="lehre.css"?>
<lehre>
  <veranstaltung jahr="2014">
    <titel>Semistrukturierte Daten</titel>
    <vorbesprechung>
      <datum>Do 6.3.</datum>
      <zeit cum_tempore="yes">09:00</zeit>
      <ort>HS 8</ort>
    </vorbesprechung>
    <schlagwort>XML</schlagwort>
    <schlagwort>DTD</schlagwort>
    <schlagwort>SGML</schlagwort>
  </veranstaltung>
</lehre>
```

Beispiel (Dokumentbaum)



Struktur eines XML Dokuments

- Baumstruktur
  - keine Einschränkung der Baumstruktur durch den XML Standard
  - ist selbstbeschreibend
  - die Ordnung der Knoten ist signifikant
- Zeichendaten vs. Markup
  - Markup repräsentiert die Struktur
  - Zeichendaten repräsentiert die restliche Information
  - beide sind einfach als Text abgelegt
  - Markup steht innerhalb spitzer Klammern `<...>` (oder `&...;`)

Elemente

- repräsentieren die strukturelle Information
- der Name des Elements steht in spitzen Klammern

```
<datum>Do 6.3.</datum>
```

- Der Inhalt des Elements wird begrenzt durch den Start Tag

```
<datum>
```

- und den End Tag

```
</datum>
```

Elemente

- Der Inhalt eines Elements sind
  - Elemente
  - Text
  - oder beides beliebig gemischt

```
<datum><tag>Do</tag>7.3.</datum>
```

- Elemente mit leerem Inhalt können abgekürzt werden:

```
<datum/>
```

Elemente

- Verschachtelung von Elementen
  - verschachtelte Elemente ermöglichen Baumstruktur
  - beliebig tiefe Verschachtelung möglich
- Verschränkte Tags sind ein Syntaxfehler

erlaubt:

```
<b>bold<i>bold-italic</i>bold</b>
```

nicht erlaubt:

```
<b>bold<i>bold-italic</b>italic</i>
```

Namen

- Der XML Standard selbst definiert (fast) keine Namen
- Namen sind case sensitive
- Genaue Regeln sind komplex (basierend auf Unicode)

Erlaubt sind:

- Buchstaben, Ziffern, Unterstrich, Bindestrich, Punkt
- Doppelpunkt erlaubt (aber hat später Spezialbedeutung)
- Fast der gesamte Unicode Zeichensatz erlaubt

Verboten sind:

- Beginn mit Ziffern
- Beginn mit den Zeichen `xml` (in beliebiger Groß-/Kleinschreibung)

Dokumente

- Bestehen aus genau einem Element
  - je nach Standard Wurzelelement oder Dokumentelement genannt
- Optional vor dem Wurzelelement kann eine XML Deklaration stehen

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

- Kann zusätzliche Information wie das Encoding enthalten
  - wenn nicht deklariert: `UTF-8`

```
<?xml version="1.0" encoding="ISO-8859-1"?>
```

Attribute

- Elemente werden durch Attribute genauer beschrieben
- Attribute werden im Start Tag des Elements definiert
  - bestehen aus dem Namen des Attributs gefolgt von
  - dessen Wert in Anführungszeichen

```
<zeit cum_tempore="yes">
```

- Ein Element kann
  - beliebig viele Attribute enthalten
  - jeder Attributname darf allerdings nur einmal vorkommen
  - wobei die Reihenfolge nicht signifikant ist

Attribute

- Der Wert eines Attributs ist reiner Text
  - das Zeichen < ist nicht erlaubt
  - es kann kein Anführungszeichen vorkommen

```
<zeit cum_tempore="yes">
```

- Alternative Notation mit Apostrophen
  - das Zeichen < ist ebenfalls nicht erlaubt
  - es kann kein Apostroph vorkommen

```
<zeit cum_tempore='yes'>
```

Kommentare

- Für Menschen bestimmte Kommentare
  - werden vom Paser nicht unbedingt an die Applikation weitergereicht
  - dürfen die Zeichenfolge `--` nicht enthalten
  - sind überall erlaubt, wo ein Element stehen darf

```
<!-- DBAI -->
```

Processing Instructions

- Für Applikationen bestimmte Processing Instructions
  - werden vom Paser an die aufrufende Applikation weitergereicht
  - bestehen aus Target (alles vor dem ersten Leerzeichen)
  - und dem Inhalt (alles nach dem ersten Leerzeichen, wenn vorhanden)
  - sind überall erlaubt, wo ein Element stehen darf

```
<?xml-stylesheet type="text/css"href="lehre.css"?>
```

- der Inhalt einer Processing Instruction muss nicht in XML Syntax sein!

Character References

- Referenzen auf Zeichen mit Spezialbedeutung
  - `<` für <
  - `>` für >
  - `&quot;` für "
  - `&apos;` für ’
  - `&` für &
- das Zeichen `&` hat daher natürlich auch eine Spezialbedeutung

- mehr Referenzierungsmöglichkeiten mit DTD
  - im Standard werden `lt`, `gt`, `quot`, `apos`, `amp` vordefinierte Entitäten genannt
  - und es können selbst zusätzliche Entitäten definiert werden

Whitespace

- Whitespace bezieht sich auf
  - Leerzeichen
  - Zeilenumbrüche
  - Tabulator
- Tritt in zwei Rollen auf:
  - innerhalb von Elementinhalt, Attributwert:
    wird an die Applikation weitergereicht
  - zwischen Attributwerten, vor und nach dem Dokumentelement:
    nicht signifikant
- Genaue Behandlung komplex und teilweise parserabhängig
  - insbesondere findet eine Normalisierung von Whitespace statt
  - zum Teil steuerbar durch Spezialattribute

Rund um XML

Durch das W3C definierte Industriestandards:

- Schemasprachen
  - DTD (Teil des XML Standards)
  - XML Schema
- Abfragesprachen
  - XPath
  - XQuery
  - XSLT

Rund um XML

- In vielen Spezifikationen verwendet:
  - Namespaces
  - Datatypes (Teil des XML Schema Standards)
- Sehr viele Standards verwenden XML Syntax
  - Office-Dokument Formate
  - Präsentationsformate (z.B. XHTML)
  - Web Services
  - ...