XML-Basics

Seminar "(Digitale) Editorik historischer Quellen"

Max Grüntgens (max.gruentgens@adwmainz.de)
Dominik Kasper (dominik.kasper@adwmainz.de)

Sommersemester 2018

Ziel von XML

- Trennt Inhalt und Struktur.
- Bringt Meta-Daten auf unterschiedlichen Ebenen an.
- Ermöglicht maschinelle Weiterverarbeitung.
- **Implizite Information explizit machen!**

XML-Grundbegriffe

Notationstypen

- Elemente `<element>`

```
<letter>
  <sender>Max Mustermann</sender>
  <recipient>Mina Musterfrau</recipient>
  <date/>
</letter>
```

- Attributes `<element attribut="attributwert">`

```
<letter identifiere="xyz">
  <sender id="m0001">Max Mustermann</sender>
  <recipient id="f0001">Mina Musterfrau</recipient>
  <date when="2016-11-28"/>
</letter when>
```

- Namespace-Präfixe `<präfix:element>`

```
<root xmlns:prj="http://url/to/namespace/prj">
```

```
  <letter identifiere="xyz">
    <sender id="m0001">Max Mustermann</sender>
    prj:letter>Blackletter script</prj:letter>
  </letter when>
</root>
```

- Leere Elemente werden meist als selbstschließende Elemente notiert.
- Wohlgeformtheit –> entspricht den allgemeinen Notationsregeln.
- Validität –> entspricht den spezifischen Regeln eines Schemas.
- Eingebaute Schema-Parser geben Feedback innerhalb des Editors.
- **Grundlage für weitere maschinelle Verarbeitung!**

XML-Notation und Syntax

- Ein geöffnetes Element muss wieder geschlossen werden.
- XML stellt eine Baumstruktur dar Keine Überlappung von Elementbereichen erlaubt!
- Die genaue "Grammatik", also Abfolge von Elementen, Inhalt von Attributen etc., wird im zugrundeliegenden Schema festgelegt.
- **Best Practice: Abschnitte eines Quellentextes oder konstituierende Attribute einer Person als Elemente, Meta-Daten wie Identifikationsnummern von Normdateien als Attribut annotieren.**

Text Encoding Initiative (TEI) – Guideline zur Annotation verschiedener Textgattungen

- TEI Root-Element `<TEI>` klammert das gesamte Dokument.
- TEI Namespace `<TEI xmlns="http://www.tei-c.org/ns/1.0">`
- Metadaten/Kopfdaten-Sektion `<teiHeader>` mit:
  - Bibliographische Sektion `<fileDesc>`
  - Encoding-Sektion `<encodingDesc>`
  - Textprofil-Sektion `<profileDesc>`
  - Revisions-Sektion `<revisionDesc>`
- Transkriptions-Sektion (`<text>`)
  - Enthält einen oder mehrere Texte
  - Die Transkription-Sektion kann eine Vielzahl an Tags zur Kodierung unterschiedlichster Strukturen und Bezüge enthalten.
  - Beispiel: Annotieren von grammatischen Worteinheiten (tokens) [1][2]
- **Während des Kodierens sind regelmäßig die TEI-P5-Guidelines zu konsultieren!**
- **Siehe auch das Übersichtspapier zu TEI-Lite.**

Abfragen und Aggregation mit XPath

- XPath fragt Pfade und Achsen innerhalb eines XML-Baumes ab und gibt gefundene Knoten, Attributwerte und Inhalte zurück.
  - Beliebiger Startpunkt der Suche wird mit `//` notiert.
  - Nodes als `//node1/node2`, Attribute in der Navigation `//node\[@Attr="Wert"\]` als Endpunkt @Attribute
- **Hinweis: Namespaces sind gegebenenfalls mit anzugeben!**
- **Siehe auch das separate Übungsblatt XPath.**

Ressourcen

M. Grüntgens, D. Kasper: Markup in geisteswissenschaftlichen Forschungs- und Publikations-Kontexten am Beispiel der Extensible Markup Language (XML). Mainz 2016 https://digicademy.github.io/mainzed_lunch_lectures_markup/#/step-1

M. Grüntgens, D. Kasper: Semantische Annotation & Kodierung. Verstehen – Auszeichnen – Abfragen. Mainz 2017 https://digitale-methodik.adwmainz.net/mod5/5c/slides/annotationen/XML_2017/#/step-1

Ron Van den Branden, Melissa Terras & Edward Vanhoutte. TEI by Example. http://www.teibyexample.org

TEI: P5 Guidelines. http://www.tei-c.org/Guidelines/P5/ http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

Übungsblatt XPath

Zum Download: http://www.benedictus.mgh.de/openmgh/bsb00000824.zip

Ausdruck	Selektion
`//Q{http://www.tei-c.org/ns/1.0}lb`	Angabe des TEI-Namespaces für Linebreak (`lb`)
`//text/body/p`	Alle Paragraphen (p) im Pfades `/TEI/text/body`
`//text/body/p\[1\]`	Der erste Paragraph (p) im Pfades `/TEI/text/body`
`//text/body/p\[@type='first'\]`	Der erste Paragraph (p) mit dem Attribut (type='first') innerhalb des Pfades `/TEI/text/body`
`//body//p/child::\*`	Alle Kind-Knoten des Knotens `p`
`//body//p/child::text()`	Alle Textknoten des Knotens `p`

- `/TEI/text`
- `/TEI/text/body`

- /TEI/text/body//lb\[5\]
- //div\[@type\]
- //w\[@type="*"\]
- //lb\[@n\<6\]
- //biblFull/*/*/child::text()

Datenbanken

eXist