

Coding the Sources

Digitales Edieren in den Geisteswissenschaften

Online Summerschool und Workshop

Lübeck/Online, 17.08. – 27.08.2020



Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums



Textkodierung mit XML

Gerlinde Schneider
gerlinde.schneider@uni-graz.at
Lübeck/Online, 16.08.2020



Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums



Roadmap

- ★ Was ist XML?
- ★ Warum Textkodierung?
- ★ XML Basics
- ★ XML Syntax und Regeln
- ★ Oxygen XML Editor
- ★ Übung

Ziel: Verstehen

- how XML works and the relationship between XML and TEI

Was ist XML?

XML steht für e**X**tensible **M**arkup **L**anguage

und ist ein generischer Standard für die **Beschreibung** und den **Austausch** von Textdokumenten / textuellen Daten

W3C Standard

- 1.1 (Second Edition) - currently in use (4th edition of version 1.0) - 2006; 12 years ago

<https://www.w3.org/TR/2006/REC-xml11-20060816/>

Warum XML

- Eignet sich gut für die Auszeichnung semistrukturierter Daten / von Fließtext
- Ermöglicht eine Auszeichnung mit zweckmäßiger Ausdruckskraft und Komplexität
- Trennt Inhalt und Form

XML ist erweiterbar:

Keine vordefinierte Struktur oder Namen von Elementen und Attributen,
Leicht an die Bedürfnisse spezifischer Domänen und Anwendungsfälle anpassbar

- ☐ System- und plattformunabhängig
- ☐ Für Menschen und Maschinen lesbar
- ☐ Niedrigschwellig
- ☐ Unterstützt durch eine breite Palette von Software
- ☐ Internationale Nutzer- und Entwicklergemeinschaft
- ☐ Umfasst eine ganze Reihe von begleitenden Standards

Modellierung strukturierter Daten

- Häufige Anwendung in Softwareentwicklung and Informationswissenschaft
- Austausch von Daten und deren Speicherung
Z.B. Konfigurationsdateien, Metadatensätze ...

```
<item>
  <fullTitle>The waste land: a facsimile and transcript of
    the original drafts, including the annotations of
    Ezra Pound / T. S. Eliot ; edited by Valerie Eliot
  </fullTitle>
  <published>
    <publicationPlace>London</publicationPlace>
    <publicationDate>1971</publicationDate>
  </published>
  <publisher>Faber & Faber</publisher>
  <created>before October 1922</created>
  <format>Facsimile / Manuscript / Typescript</format>
  <language>English</language>
  <creator>
    <name>T S Eliot</name>
    <name>Ezra Pound</name>
    <name>Vivienne Eliot</name>
  </creator>
</item>
```

Source: British Library

Modellierung von narrativem Text

Besonders bedeutend für textorientierte Geisteswissenschaften

Mixed content: Elemente können Zeichenketten (Strings) ohne Markup + weitere Elemente enthalten

```
<text>
  <body>
    <p>The American poet, critic and publisher <persName>T S Eliot</persName> was born into a
      comfortable and historically distinguished family in <placeName>St. Louis</placeName>,
      <placeName>Missouri</placeName> in <date>1888</date>. He studied at Smith Academy
      and then Harvard, where he undertook an eclectic range of courses before settling on a
      BA in what would now be called Comparative Literature and an MA in English
      Literature.</p>
    <p>He spent a year studying at the Sorbonne in Paris, and returned to Harvard to work on
      the philosophy of consciousness. This can be seen as influential in his earlier poetry,
      much of which is concerned with fractured perceptions and mental illness.</p>
  </body>
</text>
```

Textkodierung

The Waste Land
By T. S. Eliot

FOR EZRA POUND
IL MIGLIOR FABBRO

I. The Burial of the Dead

April is the cruellest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
Winter kept us warm, covering
Earth in forgetful snow, feeding
A little life with dried tubers.
Summer surprised us, coming over the Starnbergersee
With a shower of rain; we stopped in the colonnade,
And went on in sunlight, into the Hofgarten,
And drank coffee, and talked for an hour.
Bin gar keine Russin, stamm' aus Litauen, echt deutsch.
And when we were children, staying at the arch-duke's,
My cousin's, he took me out on a sled,
And I was frightened. He said, Marie,
Marie, hold on tight. And down we went.
In the mountains, there you feel free.
I read, much of the night, and go south in the winter.

What are the roots that clutch, what branches grow
Out of this stony rubbish? Son of man,

In einem Text sind viel mehr Informationen enthalten, als durch Zeichenkodierung ausgedrückt werden können.

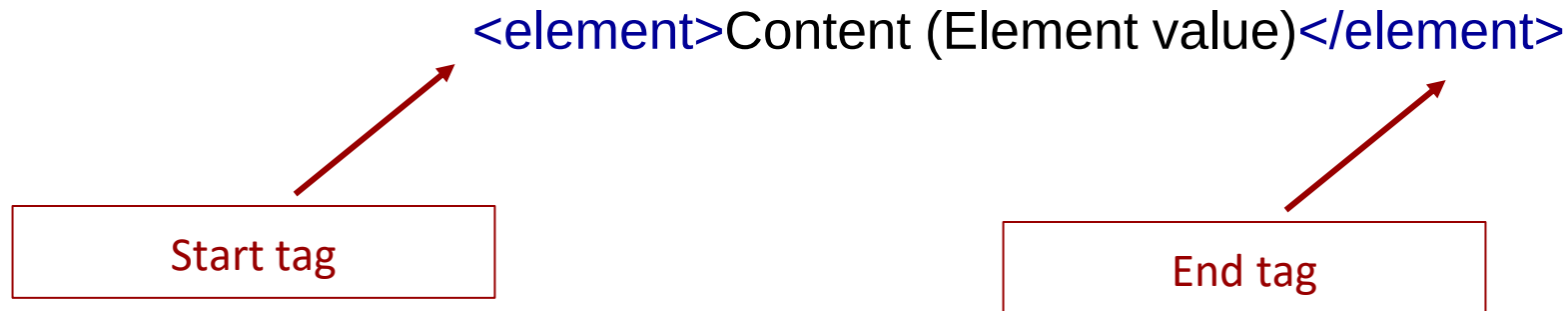
Implizite **Inhaltsinformationen** oder die **Textstruktur** werden durch Markup explizit (maschinenlesbar) gemacht.

Auch **unterschiedliche Interpretationen** eines Textes und **unterschiedliche Lesarten** können durch Markup expliziert werden.

Basic Syntax: Elemente

Die zentrale Einheit bildet das **XML-Element**

- Elemente sind Daten, die von einem Tag umgeben sind.
- Elemente müssen ein Start- und ein End-Tag haben
- Elemente können andere Elemente, Text oder beides enthalten



Elemente

- Leere Elemente haben keinen Inhalt
und werden durch ein spezielles Tag dargestellt

`<element />`



Attribute

Attribute enthalten zusätzliche Informationen zu einem Element

- Dem Start-Tag eines Elements zugewiesen
- Name/Wert-Paar
- Elemente können eine unbegrenzte Anzahl von Attributen haben,
- ein Name darf aber nur ein Mal vorkommen
- Werte müssen in Anführungszeichen stehen

```
<element attribute="value">Content</element>
```

Attribut oder Element

```
<word>  
  <text>specific</text>  
  <type>adjective</type>  
  <lemma>specific</lemma>  
</word>
```

ODER

```
<w type="adjective" lemma="specific">specific<w>
```

Regeln für XML-Namen

Gelten für Element- und Attribut-Namen

- Namen können beliebige alphanumerische Zeichen, Bindestriche, Punkte oder Unterstriche enthalten
- Namen müssen mit einem alphabetischen Zeichen, Unterstrich oder Doppelpunkt beginnen
- Namen dürfen nicht mit einer Zahl beginnen
- Bei den Namen wird zwischen Groß- und Kleinbuchstaben unterschieden:

`<title> ≠ <Title>`

- Die Namen können beliebig lang sein
- Die Verwendung von `<` `>` `&` `'` und `"` ist nicht erlaubt

Entity references

Für reservierte Zeichen

<	<	less than
>	>	greater than
&	&	ampersand
'	'	apostrophe
"	"	quotation mark

Diese müssen im Content immer escaped werden und dürfen nicht in XML Namen verwendet werden!

Root Element und Nesting

- Ein XML-Dokument hat ein einziges **Wurzelement** (root)!

Elemente innerhalb eines anderen Elements zu haben, heißt **Verschachtelung** (nesting).

Root



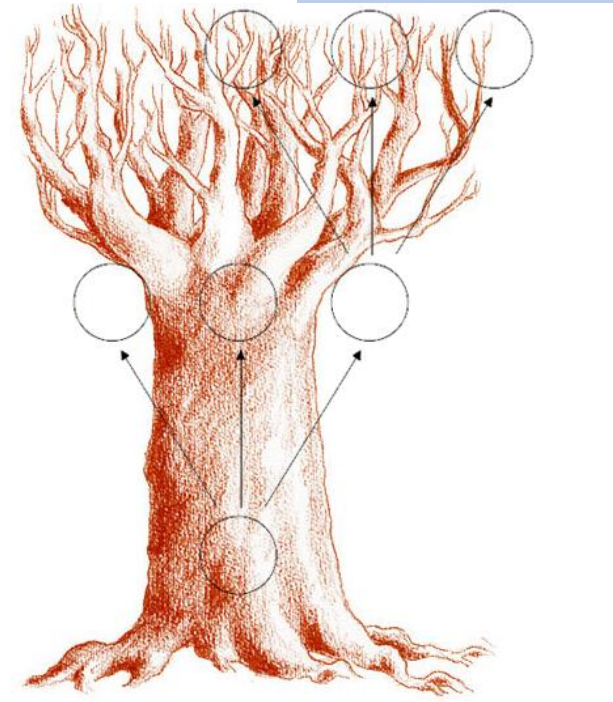
```
<p>A paragraph includes a specific  
<w>word</w> and other words.  
<s>And some sentences.</s>  
<s>And some sentences.</s>  
</p>
```

Kein "cross-nesting"!

Tree structure

- Die Verschachtelung der Elemente, beginnend mit einem Wurzel-Element ergibt eine Baumstruktur
-
- Ordered Hierarchy of Content Objects (OHCO)
 - Content objects (Knoten/Elemente)
 - Hierarchisch (Relation zwischen den Knoten)
 - Geordnet (Sequenz von Knoten)

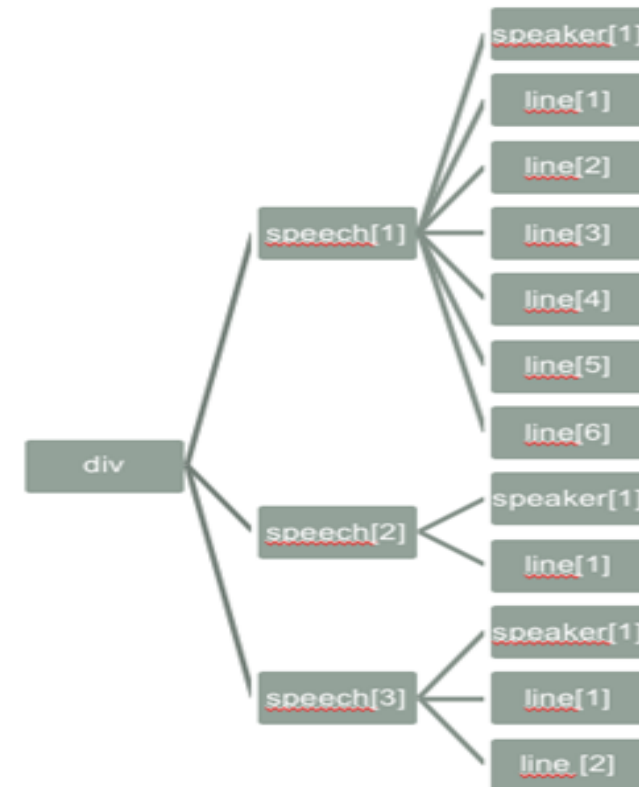
Baum: © 2008 Ilka Janke



Beispiel

Hamlet

```
<?xml version="1.0" encoding="UTF-8"?>
<div n="1">
  <speech>
    <speaker>HAMLET</speaker>
    <line>I would not hear your enemy say so,</line>
    <line>Nor shall you do mine ear that violence,</line>
    <line>To make it truster of your own report</line>
    <line>Against yourself: I know you are no truant.</line>
    <line>But what is your affair in Elsinore?</line>
    <line>We'll teach you to drink deep ere you depart.</line>
  </speech>
  <speech>
    <speaker>HORATIO</speaker>
    <line>My lord, I came to see your father's funeral.</line>
  </speech>
  <speech>
    <speaker>HAMLET</speaker>
    <line>I pray thee, do not mock me, fellow-student;</line>
    <line>I think it was to see my mother's wedding.</line>
  </speech>
</div>
```



XML Document structure

- ❑ XML Declaration

`<?xml version="1.0" encoding="UTF-8"?>`

- ❑ Processing Instructions

`<?xml-stylesheet type="text/xsl" href="transformation.xsl"?>`

- ❑ Wurzel-Element + verschachtelte Elemente

`<root> ... </root>`

- ❑ Kommentare

`<!-- This is a comment -->`

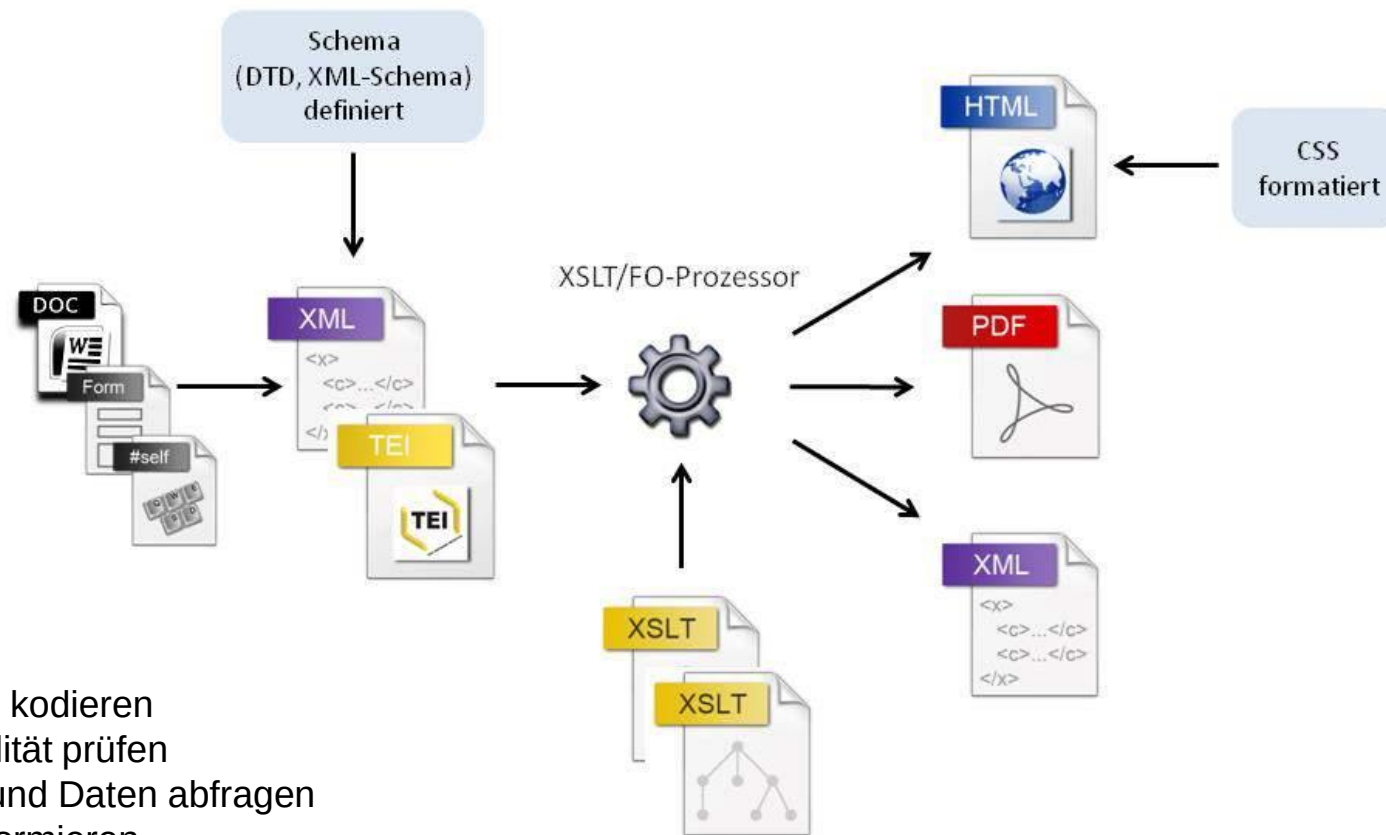
XML Document structure

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="transformation.xsl"?>
```

```
<!-- Metadata starts here -->
```

```
<item>
  <fullTitle>The waste land: a facsimile and transcript of the original drafts, including the
    annotations of Ezra Pound / T. S. Eliot ; edited by Valerie Eliot
  </fullTitle>
  <published>
    <publicationPlace>London</publicationPlace>
    <publicationDate>1971</publicationDate>
  </published>
  <publisher>Faber & Faber</publisher>
  <created>before October 1922</created>
  <format>Facsimile / Manuscript / Typescript</format>
  <language>English</language>
  <creator>
    <name>T S Eliot</name>
    <name>Ezra Pound</name>
    <name>Vivienne Eliot</name>
    <!-- Another comment -->
  </creator>
</item>
```

XML Processing

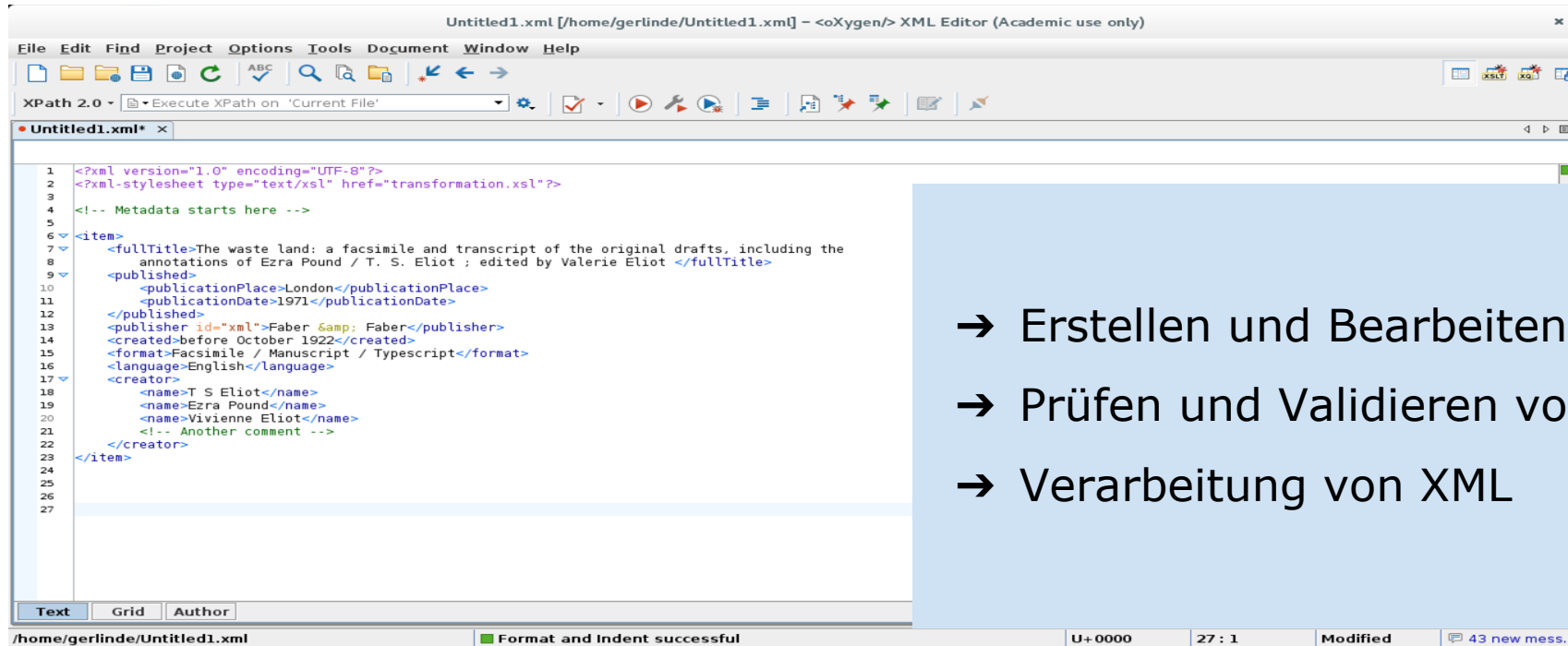


XML: Informationen kodieren
Schema: Datenqualität prüfen
XPath: Navigieren und Daten abfragen
XSLT: Daten transformieren

Source: Fritze, 2017

Oxygen editor

Oxygen ist ein Texteditor/ Entwicklungsumgebung speziell für die Bearbeitung von XML-Dokumenten.



- Erstellen und Bearbeiten von XML
- Prüfen und Validieren von XML
- Verarbeitung von XML

Oxygen editor

- Plattformunabhängig
- Subversion-Client, Add-on für Git-Integration
- Unterstützt TEI
- Unterstützt alle gängigen Schemasprachen
- Syntax-Vervollständigung
- Integrierte Dokumentation
- Eingebaute XSLT- und FOP-Prozessoren

Ist das Dokument wohlgeformt?

Ein wohlgeformtes XML-Dokument erfüllt die Regeln des Standards:

1. Es gibt genau ein Wurzelement
2. Jedes Element hat einen Start- und einen entsprechende End-Tag
3. Elemente müssen richtig verschachtelt sein - keine überlappenden Strukturen
4. Attributwerte müssen in Anführungszeichen stehen
5. Ein Element kann nicht zwei Attribute mit dem gleichen Namen haben
6. Reservierte Zeichen müssen escaped werden

Harold/Means, 2004

Ist das Dokument wohlgeformt?

- `<name>Franz Kafka</name>` 
- `<name><forename>Franz</forename><surname>Kafka</surname></name>` 
- `<name><forename>Franz<surname></forename>Kafka</surname></name>` 
- `<name type="person">Franz Kafka</name>` 
- `<name type=person>Franz Kafka</name>` 
- `<name type="person">Franz Kafka<name/>` 
- `<name type="person">Franz Kafka</Name>` 
- `<name>Franz Kafka<person/></name>` 
- `<name type="person" type="schriftsteller">Franz Kafka</name>` 
- `<name type="person author">Franz Kafka</name>` 

Validierung

Ein XML-Dokument kann nicht nur wohlgeformt sein, sondern auch auf seine **Gültigkeit** überprüft werden

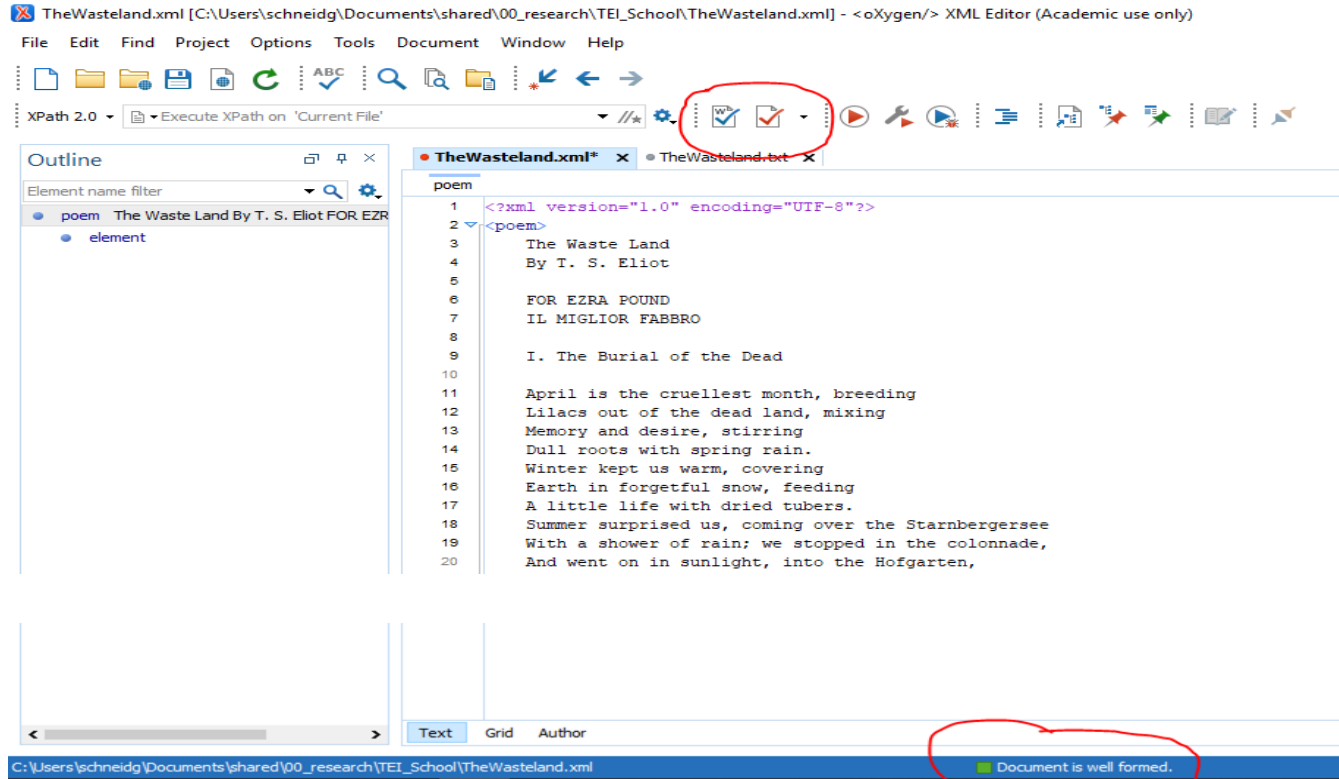
Ein valides XML-Dokument erfüllt ein Set an Regeln, die in einem spezifischen Schema definiert und an das Dokument angefügt werden

z.B. ...

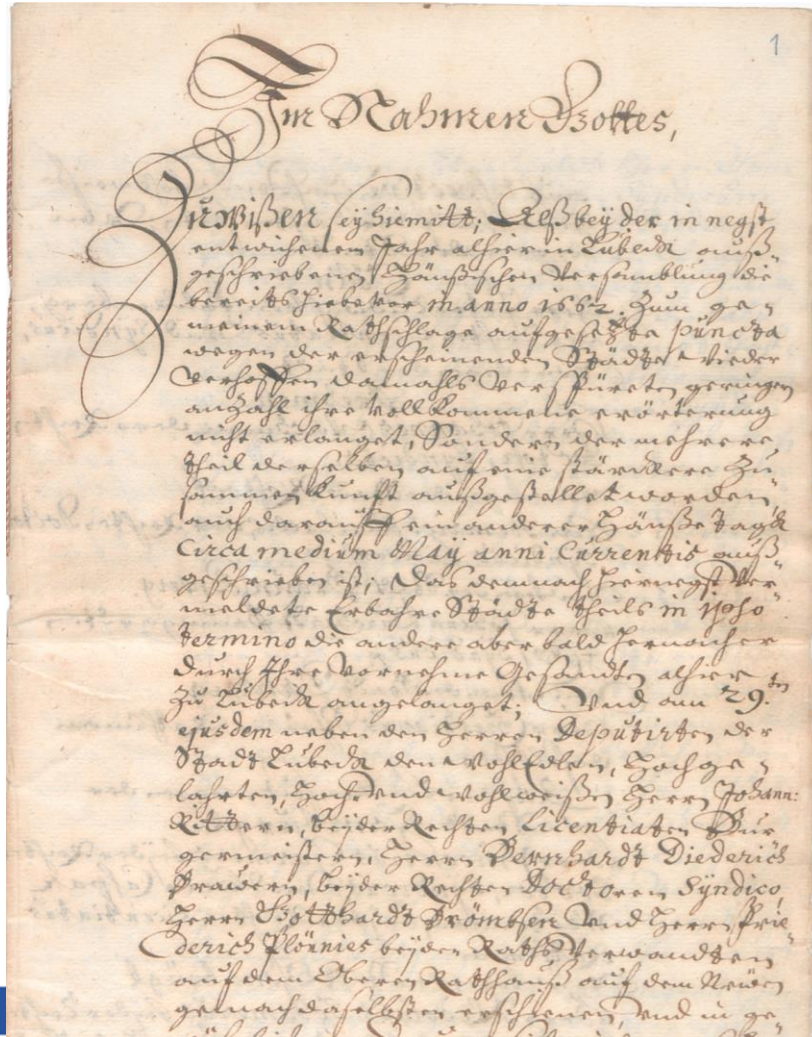
- das verwendete Vokabular (Element- und Attributnamen)
- die Struktur eines Dokuments und die Reihenfolge der Elemente

Verschiedene Schemasprachen, z.B. Document Type Definition (DTD) oder XMLSchema, ermöglichen unterschiedliche Arten der Validierung

Document check



Übung: Hanserezess



Der Lübecker Rezess des Hansetags von 1669

Signatur Archivalie: AHL - Hanseatica 247

In dieser Quelle gibt es viele interessante
Phänomene auszuzeichnen!

Transkription Rezess zur hansischen Tagfahrt 1669, Lübecker Version

[Seite 1]

Im Nahmen Gottes,

Zu wißen seÿ hiemitt; Alß beÿ der in negst
entwichenem Jahr alhier in Lubeck auß=
geschriebenen Hænßischen versamblung die
bereits hiebevör in anno 1662. zum ge=
meinem Rathschlage aufgesetzte puncta
wegen der erscheinenden Städte wieder
verhoffen damahls verspüreten geringer
anzahl ihre vollkommene erörterung
nicht erlanget, Sondern der mehrere [...]

Übung: Brainstorming

1. Welche strukturellen und inhaltlichen Elemente und Informationen kannst du in der Quelle identifizieren?
1. Welche dieser Informationen sind für dein Vorhaben relevant?
Welche Eigenschaften möchtest du auszeichnen?
1. Definiere aufbauend darauf ein Vokabular, das du konsistent zur Auszeichnung verwenden möchtest.
Wie werden die Tags heißen?
Welche Attribute wirst du verwenden?

Übung: Hands-on

- Starte den Oxygen XML Editor
- Öffne ein neues XML Dokument

Strg+N | File -> New File | -> XML Document 

- Erstelle ein Wurzelement (<text></text>)
- Kopiere die erste Seite des Transkripts in dein Wurzelement
- Beschreibe die Quelle mit den aussagekräftigen XML-Elementen und Attributen, die im ersten Schritt definiert wurden

XPath 2.0 ▾

Execute XPath on 'Current File' ▾



Hanserezess1669.xml X

text

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <text>Im Nahmen Gottes,
3 Zu wißen sey" hiemitt; Alß bey" der in negst
4 entwichenem Jahr alhier in Lubeck auß=
5 geschriebenen Hänßischen versamblung die
6 bereits hiebevor in anno 1662. zum ge=
7 meinem Rathschlage aufgesetzte puncta
8 wegen der erscheinenden Städte wieder
9 verhoffen damahls verspüreten geringer
10 anzahl ihre vollkommene erörterung
11 nicht erlanget, Sondere der mehrere
12 theil derselben auf eine stärkere Zu=
13 sammenkunft außgestallet worden,
14 auch darauff ein anderer Hänße tagk
15 circa medium May" anni currentis aus=
16 geschrieben ist; das demnach hiernegst ver=
17 meldete Erbahre Städte theils in ipso
18 termino die andere aber bald hernacher
19 durch ihre vornehme Gesandten alhier
20 zu Lubeck angelanget; vnd am 29.ten
21 eiusdem neben den Herren Deputirten der
22 Stadt Lubeck den wohlEdlen, Hochge=
23 lahrten, Hoch= und wohlweißen Herrn Johann:
24 Rittern, bey"der Rechten Licentiaten Bur=
25 germeistern, Herrn Bernhardt Diederich
26 Brawern, bey"der Rechten Doctoren Sy"ndico,
27 Herrn Gotthardt Brömben vnd Herrn Frie=
28 derich Plönnies bey"den RathsVerwandten
29 auf dem Oberen Rathhauß auf dem Newen
30 gemach daselbstn erschienen, vnd in ge=
31 wöhnlicher ordnung sich niedergesetzt
32 haben, die auch wohlEdle, Hochgelahrte
33 auch </text>
```

Text Grid Author

Referenzen:

XML in a Nutshell by [Elliote Rusty Harold](#), W. Scott Means 2004 by O'Reilly Media, Third edition

Textkodierung mit XML, Summer School "Digitale Edition" 2017, Christiane Fritze

<http://www.w3schools.com/xml/>

Fragen?

Gerlinde Schneider, gerlinde.schneider@uni-graz.at, [@lindesc](#)



Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums

