

# Datenmodellierung 3: Text, Markup, XML

Vorlesung *Einführung in die Digital Humanities*

Prof. Dr. Christof Schöch  
Wintersemester 2018/19



# Semesterüberblick

23.10.: Digital Humanities im Überblick

30.10.: Digitalisierung: Text und Bild

06.11.: Grundbegriffe des Programmierens

13.11.: Datenmodellierung 1: Modellierung

20.11.: Datenmodellierung 2: Datenbanken

## **27.11.: Datenmodellierung 3: Text, Markup, XML**

30.11.: Digitale Edition (Vorlesung statt Übung)

11.12.: Geschichte der Digital Humanities

18.12.: Informationsvisualisierung

22.12.-6.1.: Weihnachtspause

08.01.: Natural Language Processing

15.01.: Quantitative Analyse 1: Stilometrie, Topic Modeling

22.01.: Quantitative Analyse 2: Superv. Machine Learning

29.01.: Open Humanities

05.02.: Klausurtermin

# Sitzung im Überblick

1. Was ist Text? (konzeptuelles Datenmodell)
2. Grundlagen der Textkodierung
3. Textkodierung mit XML
4. Suche in XML-Dokumenten mit XPath
5. Praxis der Textkodierung

# **1. Was ist Text? (= konzeptuelles Datenmodell)**

# Beispiel Email

[Centernet] Job: HathiTrust Research Center Digital Humanities Specialist at University of Illinois at UrbanaChampaign

 **Senseney, Megan Finn** <mfsense2@illinois.edu>  
to centernet 

Nov 14 (3 days ago)   

Dear Centernet members,

Please consider applying to and/or sharing the job opening for the HathiTrust Research Center Digital Humanities Specialist. Details below.

All best,

Megan Senseney

--

HathiTrust Research Center Digital Humanities Specialist  
(Visiting Academic Professional)  
University of Illinois at Urbana Champaign Library

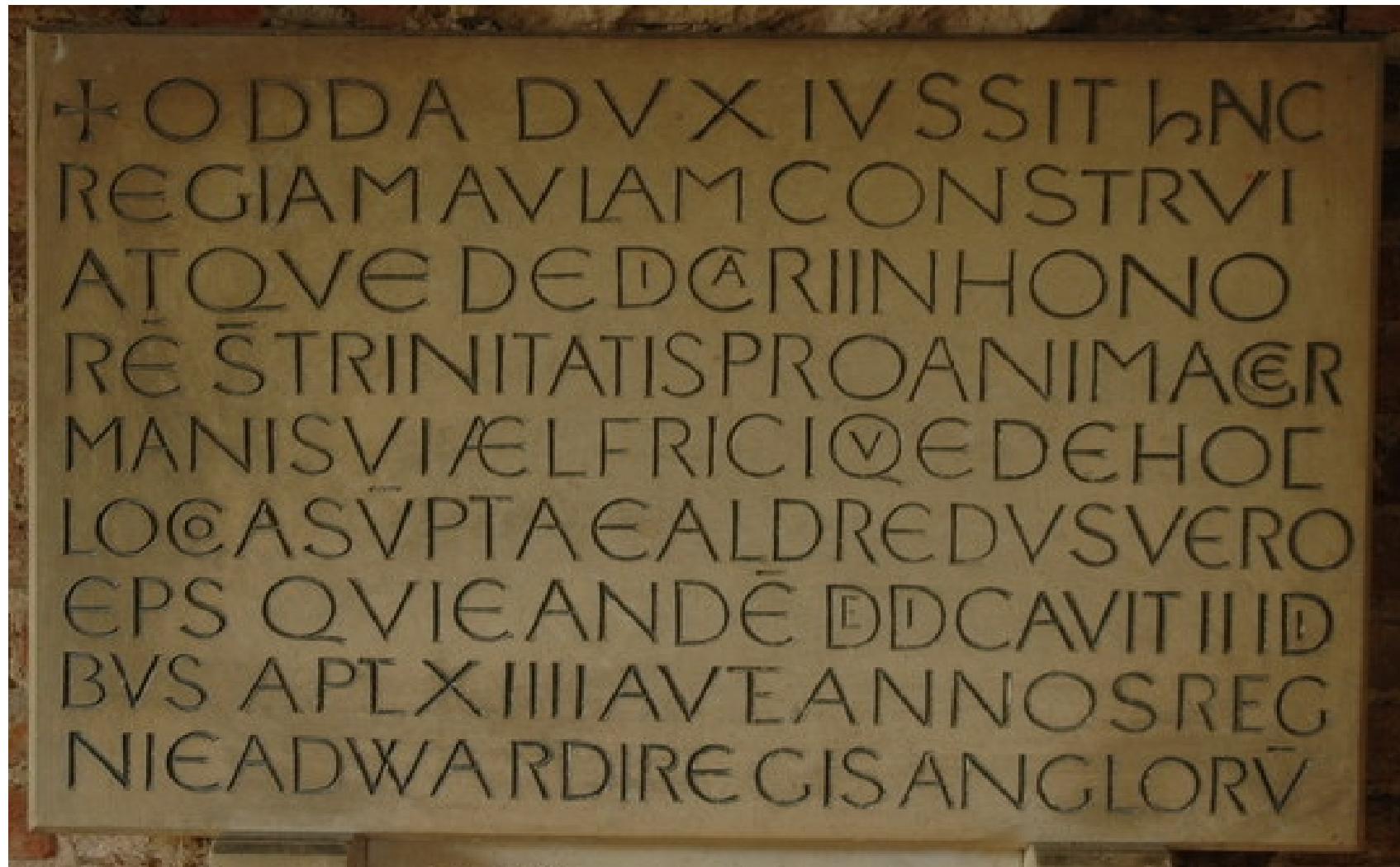
Position available immediately. This is a full-time visiting academic professional appointment in the University of Illinois at Urbana-Champaign Library, funded for two years with the possibility of renewal or being made permanent.

# Beispiel: Postkarte von 1920



Quelle: „Dortmund, Nordrhein-Westfalen“: Bahnhof, in: 10.000 Ansichtskarten von Deutschland um 1900, Dortmund: Hermann Lorch Kunstanstalt, o.J. Zeno.org. <http://images.zeno.org/Ansichtskarten/l/big/AK02828b.jpg> gemeinfre.

# Beispiel: Lateinische Inschrift

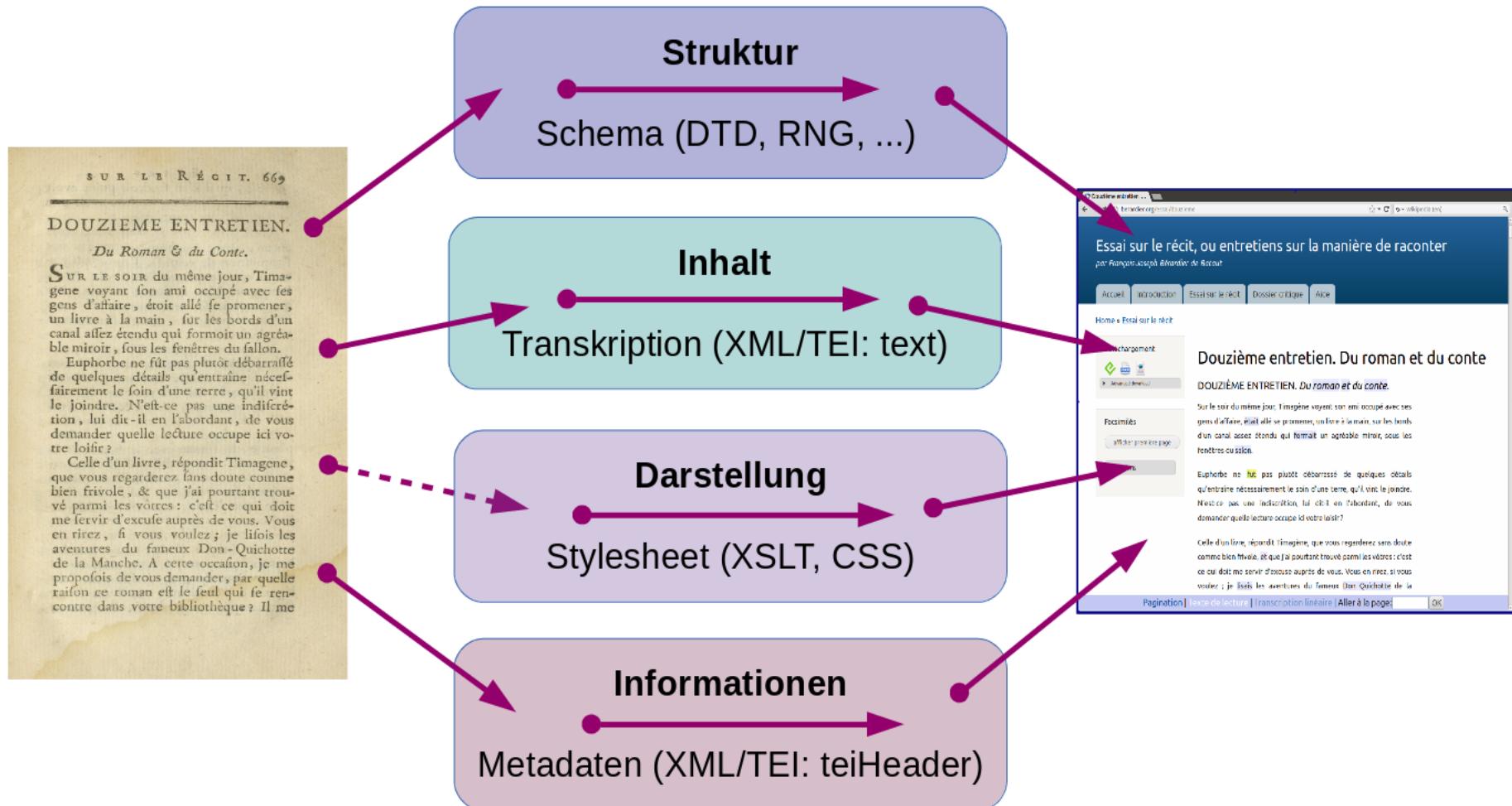


Quelle: Wikimedia Commons, Odda's Chapel Deerhurst,  
[https://commons.wikimedia.org/wiki/File:Latin\\_inscription,\\_Odda%27s\\_Chapel.jpg](https://commons.wikimedia.org/wiki/File:Latin_inscription,_Odda%27s_Chapel.jpg). Creative Commons CC-BY-SA.

# Was also ist Text?

- Buchstaben
- Leerzeichen
- Satzzeichen
- Absatzmarkierungen
- Kapitelanfänge
- Titelblätter ...
  
- inhaltlicher und sprachlicher Zusammenhang?
- Anfang, Mitte und Ende

# Aspekte des digitalen Texts



# Struktur

- Explizite Kodierung von Strukturinformationen
  - z.B. Kapitelgrenze
  - Absatzgrenze
  - Seitengrenze
- Standards zur Kodierung von Strukturen
  - XML
  - TEI

# Textauszeichnung der Struktur

```
<div type="Werk">
  <head>Die Leiden des jungen Werthers</head>
  <div type="Teil"><head>Erster Theil</head>
    <p> [...] <p>
    <div type="Kapitel"><head>am 4. May 1771.</head>
      <p>Wie froh bin ich, daß ich weg bin! Bester
        Freund, was ist das Herz des Menschen!
      [...]</p>
      [...]
    </div>
  </div>
</div>
```

# Layout: zwei Verfahren

- Direkte Eingabe im Text

**Das ist die fettgedruckte Überschrift**

<bold>Dies ist die fettgedruckte Überschrift</bold>

- Zuordnung zur Struktur

Im Text:

<h1 class="Kapitel">Kapitel 1. Einleitung</h1>

Im Stylesheet

.Kapitel { font-family:sans-serif; font-size:24pt; color:blue; }

# Metadaten

- Informationen über den Text werden im Text kodiert
  - z.B. Autor oder Editor
  - Erstellungsdatum des Textes oder der digitalen Form
  - uvm.
- Standards zur Kodierung von Metadaten:
  - Dublin Core
  - TEI Header
  - HTML „META“ Bereich

# Metadaten (im HTML HEAD)

The screenshot shows a web browser window with the title bar 'E06\_Mann-HTML\_pg12108.html'. The main content area displays the source code of a HTML document. The code includes a head section with various meta tags for DCTERMS and MARCREL, and a body section containing paragraphs about Gustav Aschenbach.

```
1 <html>
2   <head>
3     <title>Der Tod in Venedig</title>
4     <link href="http://purl.org/dc/terms/" rel="schema.DCTERMS" />
5     <link href="http://id.loc.gov/vocabulary/relators" rel="schema.MARCREL" />
6     <meta content="Der Tod in Venedig" name="DCTERMS.title" />
7     <meta content="http://www.gutenberg.org/files/12108/12108-8.txt" name="DCTERMS.source" />
8     <meta content="de" scheme="DCTERMS.RFC4646" name="DCTERMS.language" />
9     <meta content="Public domain in the USA." name="DCTERMS.rights" />
10    <meta content="Mann, Thomas, 1875-1955" name="DCTERMS.creator" />
11    <meta content="Authors -- Fiction" scheme="DCTERMS.LCSH" name="DCTERMS.subject" />
12    <meta content="Munich (Germany) -- Fiction" scheme="DCTERMS.LCSH" name="DCTERMS.subject" />
13    <meta content="Venice (Italy) -- Fiction" scheme="DCTERMS.LCSH" name="DCTERMS.subject" />
14    <meta content="2004-04-01" scheme="DCTERMS.W3CDTF" name="DCTERMS.created" />
15    <style type="text/css">
16      .pageno { position: absolute; right: 95%; font: medium sans-serif; text-indent: 0 }
17      .pageno:after { color: gray; content: '[' attr(title) ']' }
18      .lineno { position: absolute; left: 95%; font: medium sans-serif; text-indent: 0 }
19      .lineno:after { color: gray; content: '[' attr(title) ']' }
20    </style>
21  </head>
22  <body>
23    <p id="id00008" style="margin-top: 4em">Thomas Mann</p>
24    <p id="id00009">Der Tod in Venedig</p>
25    <p id="id00013" style="margin-top: 6em">Erstes Kapitel</p>
26    <p id="id00014" style="margin-top: 2em">
27      Gustav Aschenbach oder von Aschenbach, wie seit seinem fünfzigsten
28      Geburtstag amtlich sein Name lautete, hatte an einem
29      Frühlingsnachmittag des Jahres 19..., das unserem Kontinent monatelang
30      eine so gefahrdrohende Miene zeigte, von seiner Wohnung in der
31      Prinz-Regentenstraße zu München aus, allein einen weiteren Spaziergang
32      unternommen. [...]
33    </p>
34  </body>
35 </html>
```

# Metadaten (im teiHeader)

Screenshot of EditPlus showing XML code for a TEI header.

```

EditPlus - [F:\xml\xtei\jgoethe\jgoethe0.xml *]
File Edit View Search Document Project Tools Window Help

<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Der junge Goethe in seiner Zeit. Texte und Kontexte. </title>
      <author>Johann Wolfgang Goethe</author>
      <respStmt>
        <resp>Herausgegeben von </resp>
        <name>Karl Eibl, Fotis Jannidis, Marianne Willems</name>
      </respStmt>
    </titleStmt>
    <editionStmt>
      <edition>Erste Ausgabe </edition>
      <date>1998</date>
    </editionStmt>
    <publicationStmt>
      <p>Der junge Goethe in seiner Zeit</p>
      <p>Herausgegeben von Karl Eibl, Fotis Jannidis und Marianne Willems</p>
      <p>Erste Auflage 1998</p>
      <p>ISBN gebunden 3-458-16914-8</p>
      <p>Taschenbuch 3-458-33800-4</p>
      <p>Insel Verlag Frankfurt/Main</p>
      <p>Weitere Informationen: http://www.jgoethe.uni-muenchen.de/</p>
    </publicationStmt>
    <sourceDesc><p>Die Texte des Jungen Goethe entstammen, so weit nicht anders angegeben, der Ausgabe: Der junge Goethe. <br/> Neu bearb. 3. Ausgabe. 5 Bde. und 1 Reg.-Bd. Hrsg. v. Hanna Fischer-Lamberg. <br/> Berlin bzw. Bd. 5 und Reg.-Bd. Berlin und New York 1963 - 1974. (Sigle FL). <br/> Das Nähere jeweils im Kommentar. Die Herkunft der anderen Vorlagen wird jeweils unter dem Titel genannt. <br/> Wo kein spezieller Herkunftsvermerk steht, war das Original die Vorlage.</p>
    </sourceDesc>
  </fileDesc>
  <encodingDesc>
    <projectDesc>
      <p>Die TEI-Dateien wurden durch Konvertierung der Folio Flat Files gewonnen. Zur besseren Handhabung mit XML-Browsern wurden die Daten in 24 Dateien gespeichert: jgoethe0.sgm - jgoethe23.sgm. Die zugehörigen Entities befinden sich in den gleichnamigen Dateien mit der Endung .ent </p>
    </projectDesc>
    <editorialDecl>
      <p>Die editorischen Prinzipien der Ausgabe sind im Nachwort (Teil des Textes) nachzulesen. <br/> Eine Beschreibung der Tags und zur Formatierung wesentlicher Attribute befindet sich in der beiliegenden readme.txt</p>
    </editorialDecl>
  </encodingDesc>
</teiHeader>

```

For Help, press F1

# Inhalt

- Informationen werden nach inhaltlichen Gesichtspunkten ausgezeichnet:
  - Morpho-syntaktische Informationen (ling. Annotation)
  - Normdaten / Named Entities (Personen, Orte, Organisationen, etc.)
  - Registerinformationen: Autoren, Werke, Konzepte (ggfs. auf Grundlage einer Taxonomie)
- Wie macht man das?
  - Standards für linguistische Annotation: diverse Tagsets
  - Standards zur Kodierung von Inhalten: Normdaten (GND, VIAF, Getty Thesaurus)
  - Standards für die Kodierung von Konzepten: Taxonomien, Ontologien, Schlagwortsysteme (siehe bspw. Satorbase)

## **2. Textauszeichnung (Markup)**

# Was ist Markup?

- Ursprung im Druckwesen: Anweisungen für den Setzer
- “markup”
  - (Subst.) = Auszeichnung, Markierung
  - (Verb) = auszeichnen, markieren, kodieren
- Macht Eigenschaften explizit
  - benennt und/oder charakterisiert Teile der Zeichenkette
  - auf formalisierte und kohärente Weise

# Textauszeichnung: Druck vs. Digital

- Textauszeichnung im Druck:
  - Einfügen von Setzeranweisungen ins Manuskript
- Textauszeichnung im digitalen Text:
  - Vom Text systematisch unterschiedene Information über den Text bzw. Teile des Texts
  - Explizitmachen von Eigenschaften

# Exkurs: Setzschild, Setzrahmen, etc.



## Quellen:

Foto links oben: Wilhei, *Handsatz Winkelhaken*, 2009.

<https://commons.wikimedia.org/wiki/File:Handsatz.jpg>.

Lizenz: Creative Commons Attribution 3.0 Unported license.

Foto links unten: Roletschek, Ralf, *Nördlingen, alte Druckerei. Schriftstock Bleisatz*, 2009.

<https://commons.wikimedia.org/wiki/File:2009-04-18-noerdingen-rr-05.jpg>

Lizenz: Gemeinfrei.

Foto rechts oben: Wilhei, *Handsatz im Setzschild*, 2009.

[https://commons.wikimedia.org/wiki/File:Handsatz\\_im\\_Setzschild.jpg](https://commons.wikimedia.org/wiki/File:Handsatz_im_Setzschild.jpg)

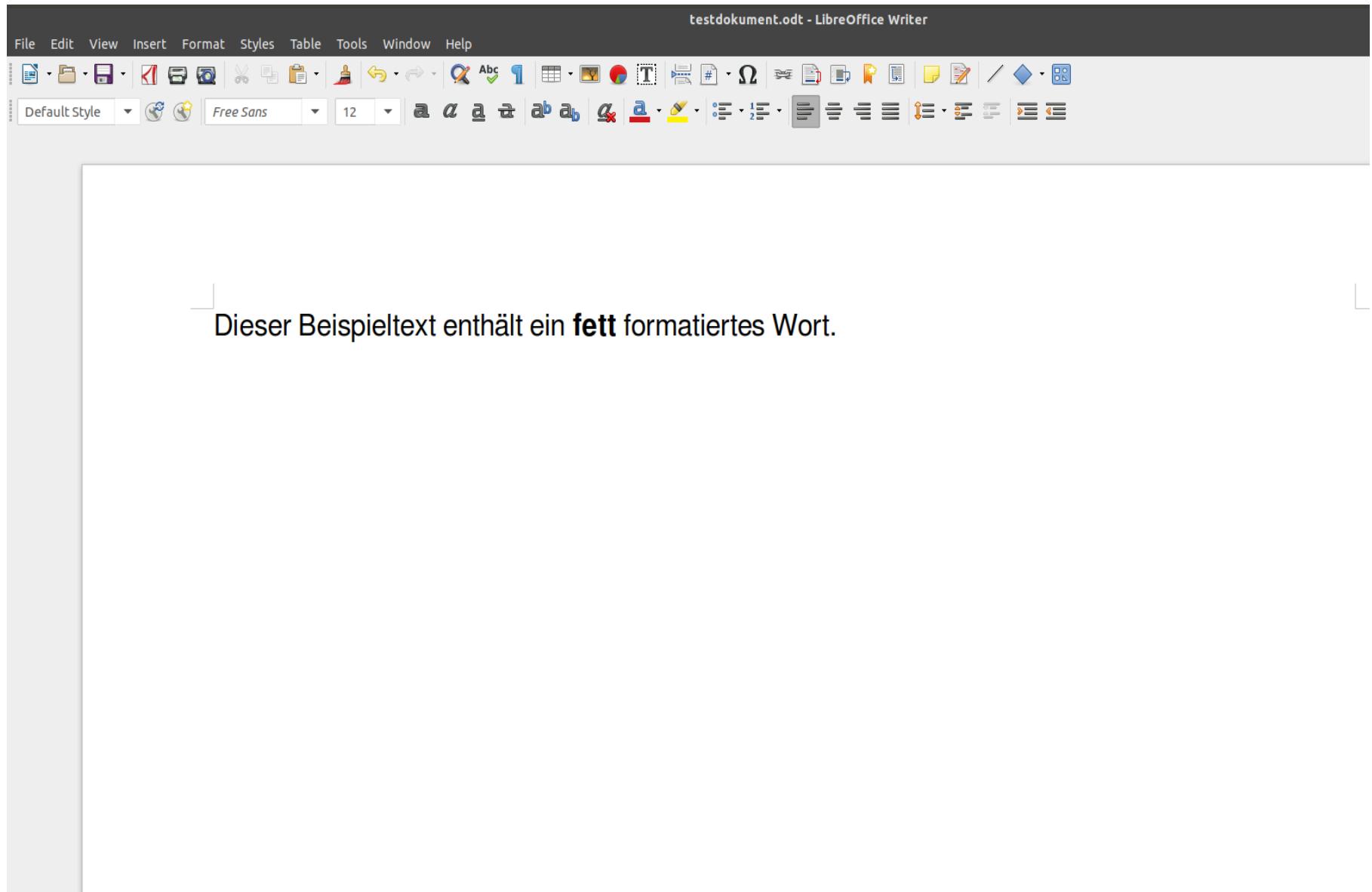
Lizenz: Creative Commons Attribution 3.0 Unported.

Foto rechts unten: Wilhei, *Handsatz-Werkzeug: Ahle, Kolumnenschnur, Winkelhaken und ausgebundene Kolumne auf einem Setzschild*, 2009.

[https://commons.wikimedia.org/wiki/File:Handsatz\\_Zubehoer.jpg](https://commons.wikimedia.org/wiki/File:Handsatz_Zubehoer.jpg)

Lizenz: Creative Commons Attribution 3.0 Unported.

# Textauszeichnung: im Word Processor



# .rtf im plain-text Editor

testdokument.rtf

```
File Edit Search View Document Project Build Tools Help
testdokument.rtf x
1 {\rtf1\ansi\deff4\adeflang1025
2 {\fonttbl{\f0\froman\fprq2\fcharset0 Times New Roman;}{\f1\froman\fprq2\fcharset2 Symbol;}{\f2\fswiss\fprq2\fcharset0 Arial;}{\f3\froman\fprq2\fcharset0 Liberation
Serif\{\*\falt Times New Roman;}{\f4\fswiss\fprq0\fcharset128 Free Sans;}{\f5\fnil\fprq0\fcharset2 StarSymbol\{\*\falt Arial Unicode MS;}{\f6\fnil\fprq2\fcharset0 Noto
Sans CJK SC Regular;}{\f7\fnil\fprq2\fcharset0 Lohit Devanagari;}{\f8\fnil\fprq0\fcharset128 Lohit Devanagari;}{\f9\fnil\fprq0\fcharset0 Lohit Devanagari1;}}
3 {\colortbl{\red0\green0\blue0;\red0\green0\blue255;\red0\green255\blue255;\red0\green255\blue0;\red255\green0\blue255;\red255\green0\blue0;\red255\green255\blue0;\red255\green255\blue255;\red0\green0\blue128;\red0\green128\blue128;\red0\green128\blue0;\red128\green0\blue128;\red128\green128\blue0;\red128\green128\blue128;\red192\green192\blue192;}}
4 {\stylesheet{\s0\snext0\nowidctlpar\hyphpar0\cf0\kerning1\dbch\af6\langfe2052\dbch\af7\afs24\alang1081\loch\f4\hich\af4\fs24\lang1023 Normal;}}
5 {\*\cs15\snext15 Footnote Characters;}
6 {\*\cs16\snext16 Numbering Symbols;}
7 {\*\cs17\snext17\dbch\af5\dbch\af5\afs18\loch\f5\fs18 Bullets;}
8 {\*\cs18\snext18 Endnote Characters;}
9 {\*\cs19\snext19\super Footnote Anchor;}
10 {\*\cs20\snext20\b\ab0 wT1;}}
11 {\*\cs21\snext21\b0\ab0 wHyperlink;}}
12 {\*\cs22\snext22\b0\ab0 wFollowedHyperlink;}}
13 {\*\cs23\snext23\b0\ab0 wCommentReference;}}
14 {\s24\sbasedon0\snext25\sb240\sa120\keepn\dbch\af6\dbch\af7\afs28\loch\f4\fs28 Heading;}}
15 {\s25\sbasedon0\snext25\sl288\slmult1\sb0\sa140 Text Body;}}
16 {\s26\sbasedon25\snext26\sl288\slmult1\sb0\sa140\dbch\af8\loch\f4 List;}}
17 {\s27\sbasedon0\snext27\sb120\sa120\oneline\i\dbch\af8\afs24\ai\loch\f4\fs24 Caption;}}
18 {\s28\sbasedon0\snext28\oneline\dbch\af8\loch\f4 Index;}}
19 {\s29\snext29\nowidctlpar\hyphpar0\dbch\af6\dbch\af7\cf0\kerning1\langfe2052\afs24\alang1081\loch\f4\fs24\lang1023 wdefault-paragraph-style;}}
20 {\s30\sbasedon29\snext30\dbch\af6\dbch\af7\loch\f4\fs24\lang1023 wStandard;}}
21 {\s31\sbasedon30\snext30\sb0\sa0\dbch\af6\dbch\af7\loch\f4\fs28\lang1023 wHeading;}}
22 {\s32\sbasedon30\snext32\sl288\slmult1\sb0\sa0\dbch\af6\dbch\af7\loch\f4\fs24\lang1023 wText_20_body;}}
23 {\s33\sbasedon32\snext33\sl288\slmult1\sb0\sa0\dbch\af6\dbch\af9\loch\f4\fs24\lang1023 wList;}}
24 {\s34\sbasedon30\snext34\sb0\sa0\dbch\af6\dbch\af9\loch\f4\fs24\lang1023 wCaption;}}
25 {\s35\sbasedon30\snext35\dbch\af6\dbch\af9\loch\f4\fs24\lang1023 wIndex;}}
26 {\s36\sbasedon30\snext36\dbch\af6\dbch\af7\loch\f4\fs24\lang1023 wP1;}}
27 {\s37\snext37\nowidctlpar\hyphpar0\afs20\cf0\kerning1\dbch\af6\langfe2052\dbch\af7\alang1081\fs20\loch\f4\hich\af4\lang1023 wCommentText;}}
28 {\s38\sbasedon37\snext37\afs20\fs20 wCommentSubject;}}
29 {\*\generator LibreOffice/5.4.1.2$Linux_X86_64
LibreOffice_project/40m0$Build-2}\{\info{\creatin\yr2017\mo12\dy6\hr15\min11}\{\revtim\yr2017\mo12\dy6\hr15\min11}\{\printim\yr0\mo0\dy0\hr0\min0\}}{\*\userprops{\propname
Category}\proptype30{\staticval }\{\propname Editor}\proptype30{\staticval LibreOffice/5.4.1.2$Linux_X86_64 LibreOffice_project/40m0$Build-2}\{\propname
HyperLinkBase}\proptype30{\staticval }\{\propname Language}\proptype30{\staticval }\{\propname Manager}\proptype30{\staticval }\{\propname Version}\proptype30{\staticval
}\deftab454\deftab454\deftab454
30 \viewscale200
31 {\*\pgdsctbl
32 {\pgdsc0\pgdcuse451\pgwsxn12240\pghsxn15840\marglsxn1134\margtsxn1134\margbsxn1134\pgdscnxt0 Default Style;}}
33 {\pgdsc1\pgdcuse451\pgndec\pgwsxn12240\pghsxn15840\marglsxn1134\margtsxn1134\margbsxn1134\pgdscnxt2 First Page;}}
34 {\pgdsc2\pgdcuse451\pgndec\pgwsxn12240\pghsxn15840\marglsxn1134\margtsxn1134\margbsxn1134\pgdscnxt2 Standard-1;}}
35 {\formshade{\*\pgdscnol}\paperh15840\paperw12240\margl1134\margr1134\margt1134\margb1134\sectd\sbknone\sectunlocked1\pgndec\pgwsxn12240\pghsxn15840\marglsxn1134\margtsxn1134\margbsxn1134\titlepg\ftnbj\ftnstart1\ftnrstcont\ftnnar\aeenddoc\aftnrstcont\aftnstart1\aftnrlc
36 {\*\ftnsep\chftnsep}\pgndec\pard\plain \s36\dbch\af6\dbch\af7\loch\f4\fs24\lang1023\{\rtlch \ltrch\loch
37 Dieser Beispieldatei enthu228'3fl ein }\{\cs20\b\ab0\rtlch \ltrch\loch
38 fett }\{\rtlch \ltrch\loch
39 formatiertes Wort. }
40 \par }
```

# Textauszeichnung: HTML

```
<HTML>
  <HEAD>
    <TITLE> 1. HTML Beispiel</TITLE>
  </HEAD>
  <BODY>
    <H1>Textauszeichnung</H1>
    <P>Dies ist <B>der</B> Text.</P>
  </BODY>
</HTML>
```

# Zwei Typen von Markup

- "**prozedural**": visuell, typographisch:
  - Anweisung, wie ein Stück Text dargestellt werden soll
  - Schwerpunkt auf Darstellung, Aussehen
  - häufig mehrdeutig, unflexibel
- "**deskriptiv**": semantisch, funktional, strukturell
  - Explizieren, welche Funktion ein Stück Text hat
  - Schwerpunkt auf Struktur und Bedeutung
  - separat davon die Darstellung definieren
  - mehr, eindeutigere Information
  - Darstellung leichter anpassbar

# Prozeduraler Markup

## Woody Allens Urteil

*Krieg und Frieden* ist ein fantastisches, aber *viel* zu langes Buch.

## "prozedural"

```
<markup>
  <fett>Woody Allens Urteil</fett><umbruch/><kursiv>Krieg und
Frieden</kursiv> ist ein fantastisches, aber
<kursiv>viel</kursiv> zu langes Buch.
</markup>
```

# Deskriptiver Markup

## Woody Allens Urteil

*Krieg und Frieden* ist ein fantastisches, aber *viel* zu langes Buch.

## "deskriptiv"

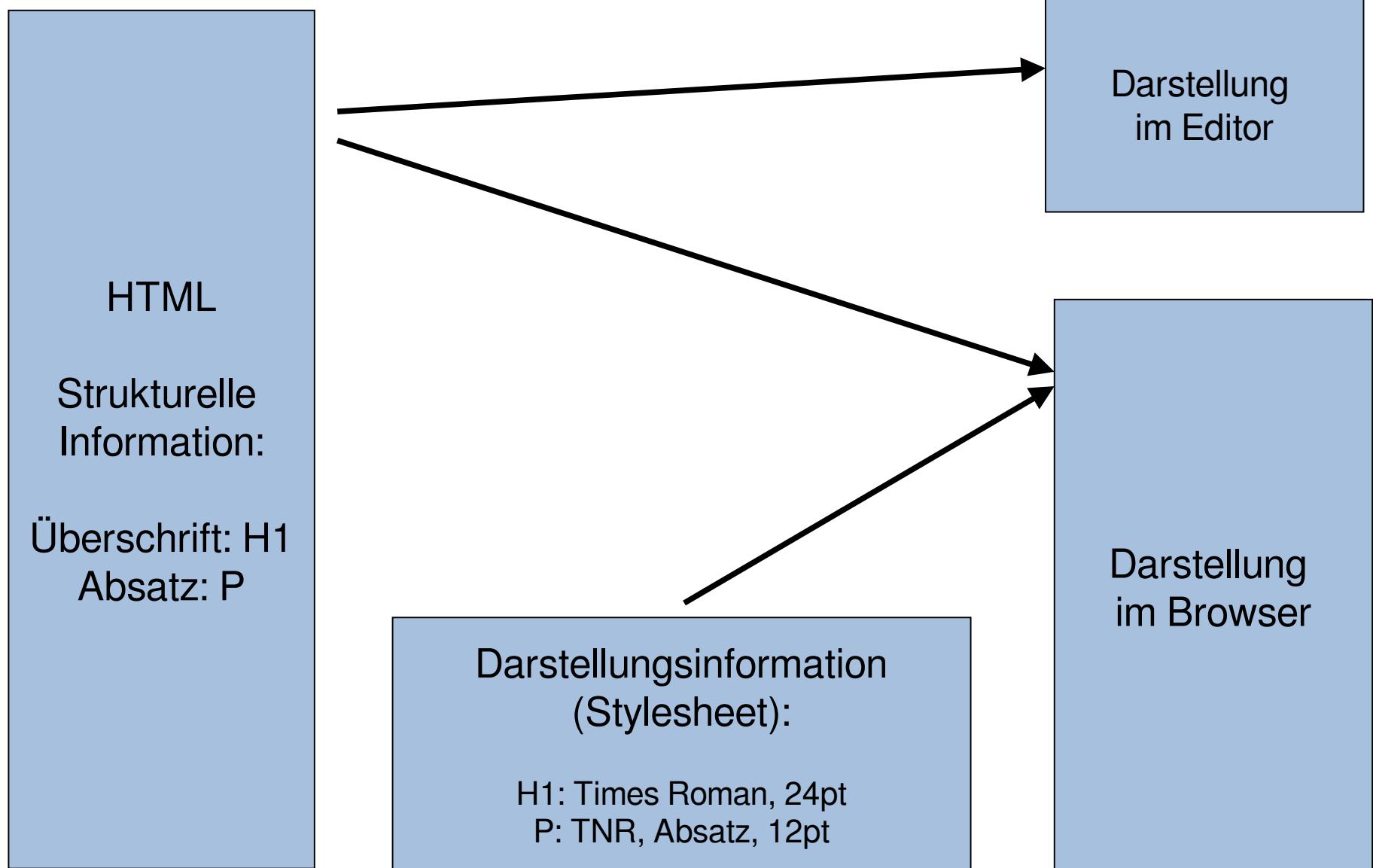
```
<markup>
  <titel>Woody Allens Urteil</titel><absatz> <buch>Krieg und
Frieden</buch> ist ein fantastisches, aber <emph>viel</emph> zu
langes Buch.</absatz>
</markup>
```

## Stylesheet



```
<stylesheet>
  titel: fett + Umbruch; buch: kursiv;
  absatz: normal, Umbruch; emph: kursiv;
</stylesheet>
```

# Rendering mit Stylesheet



# Anforderungen: System zum Textprozessieren

- Verwendbar für viele Zwecke und vom Benutzer an seine Interessen anpassbar
- Beschreibend (deskriptiv) und nicht verarbeitend (prozedural)
- Lesbar für Menschen und Maschinen
- Unterstützung bei Texteingabe sowie Kontrolle auf Konformität mit Textmodell durch formale Beschreibung
- Kein proprietäres System, sondern offener Standard

# **3. Textkodierung mit XML (= logisches Datenmodell)**

# Was ist XML?



- eXtensible Markup Language
- W3C-Standard
- Metasprache zur Definition von XML-Formaten
- Standard für digitale Repräsentation von Daten
- Prinzipien + Syntax
- einfach (wenige, mächtige Mechanismen)
- anwendungs- und plattformunabhängig

## **3.1 Bestandteile eines XML-Dokuments**

# XML: Elemente, Attribute, Werte, Strings



The screenshot shows a code editor window with the title bar "XML-Elemente-Attribute-Werte.xml". The XML document contains two crime entries (Krimi) with their titles, authors, and names. The code is color-coded: blue for tags, purple for attributes, and black for strings.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <Krimi-Sammlung>
3  <Krimi n="1">
4      <titel>Piège pour Cendrillon</titel>
5      <autor status="bekannt">
6          <name>Japrisot</name>
7          <vorname>Sébastien</vorname>
8      </autor>
9  </Krimi>
10 <Krimi n="2">
11     <titel>Meurtres pour mémoire</titel>
12     <autor status="berühmt">
13         <name>Daeninckxs</name>
14         <vorname>Didier</vorname>
15     </autor>
16 </Krimi>
17 </Krimi-Sammlung>
18
```

# Element / Tags

<überschrift>1. Kapitel</überschrift>

# Leere Elemente

<neueZeile/>

# Attribute und Werte

- <div type="chapter">
- <div type="chapter" nr="1">

# Entity

Entities sind Platzhalter für andere Zeichen. Sie funktionieren ungefähr so, wie Textbausteine in einer Textverarbeitung.

Zwei Formen von Entities:

- Parameter Entity (nur in der DTD)  
%isolat1;
- General Entity (nur in der Dokumentinstanz)  
Ich k&ouml;nnte glauben, da&szlig; Sie .....

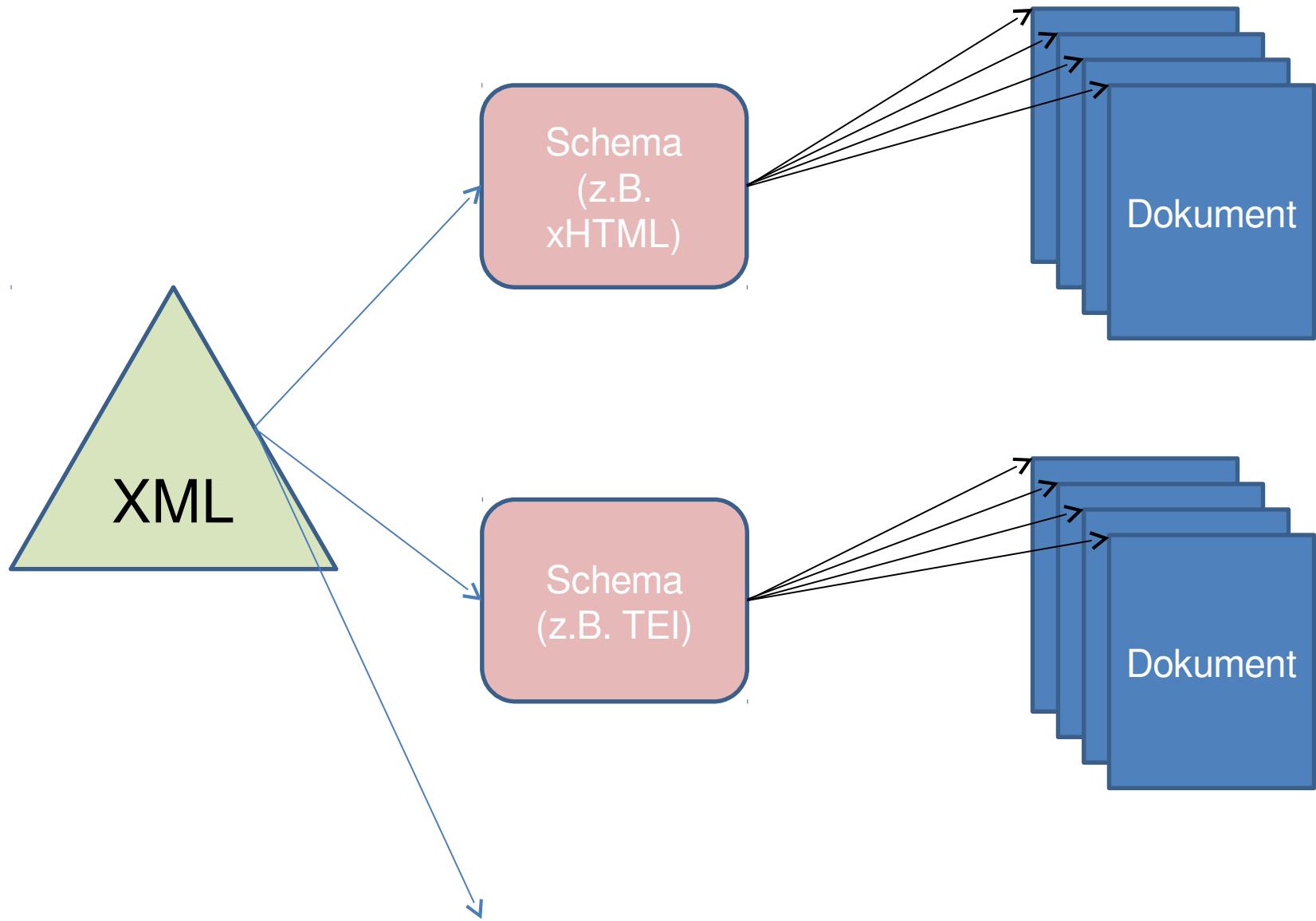
# XML-Entities

- < &lt;
- & &amp;
- Fontane & Sohn  
`<p>Fontane &amp; Sohn</p>`

# Kommentar

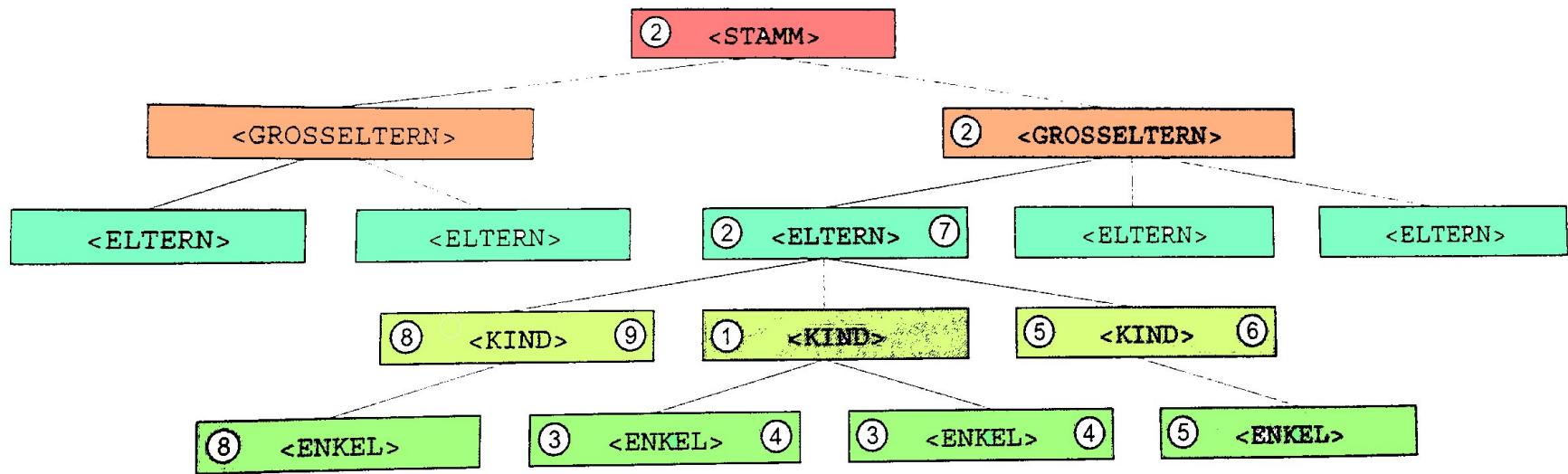
<!-- stimmt die Auszeichnung  
hier? -->

# XML-Welt



## **3.2 Die Baumstruktur von XML**

# Baumstruktur: hierarchische Relationen



(Quelle: RRZN, XML Grundlagen, 2008)

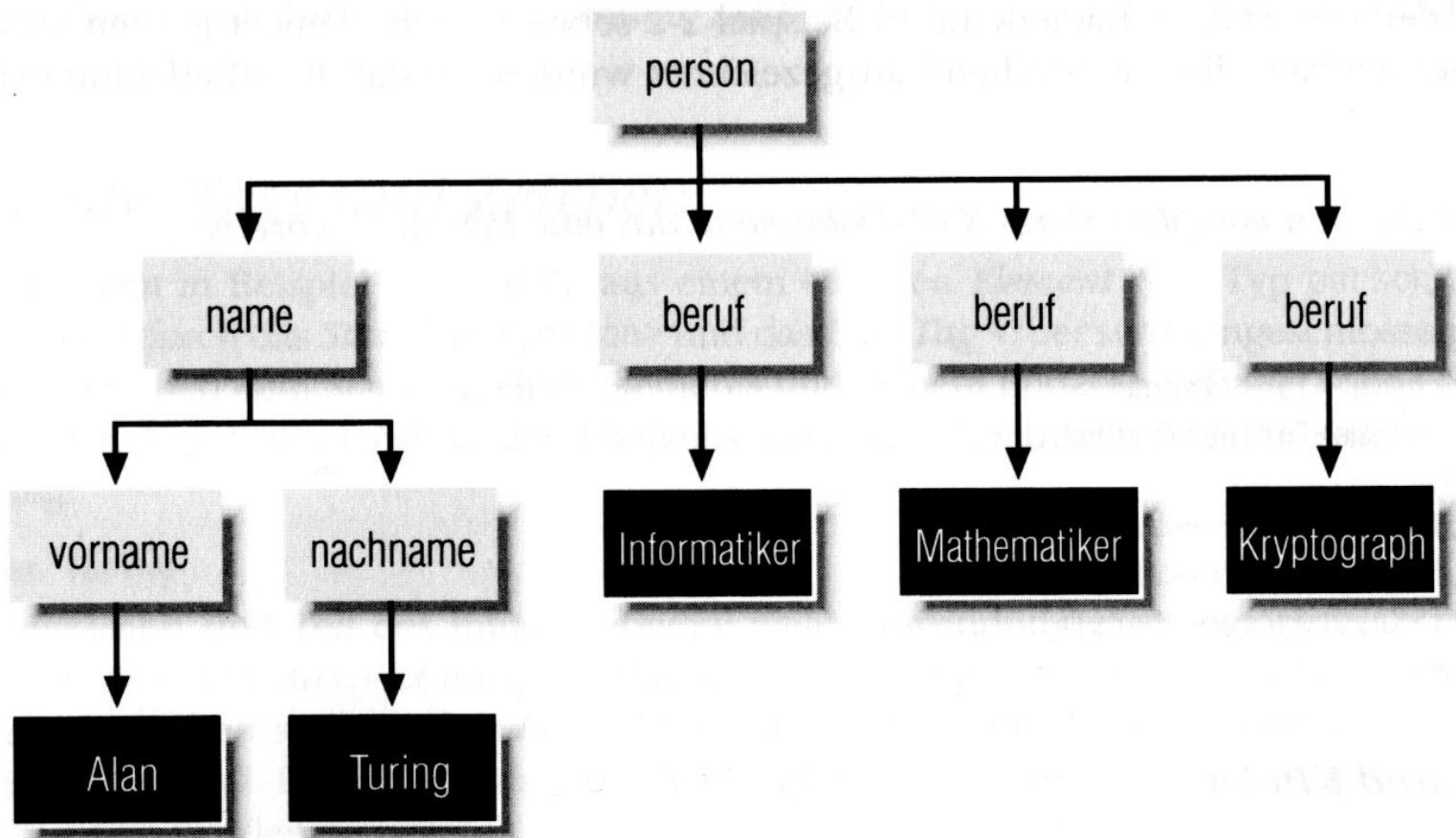
# Baumstruktur = Texttheorie

- Konzeptuelles Datenmodell (was ist Text?):  
“Ordered Hierarchy of Content Objects” (OHCO)
- Logisches Datenmodell:  
hierarchische Struktur von XML

# Kodierung in XML

```
<person>
  <name>
    <vorname>Alan</vorname>
    <nachname>Turing</nachname>
  </name>
  <beruf>Informatiker</beruf>
  <beruf>Mathematiker</beruf>
  <beruf>Kryptograph</beruf>
</person>
```

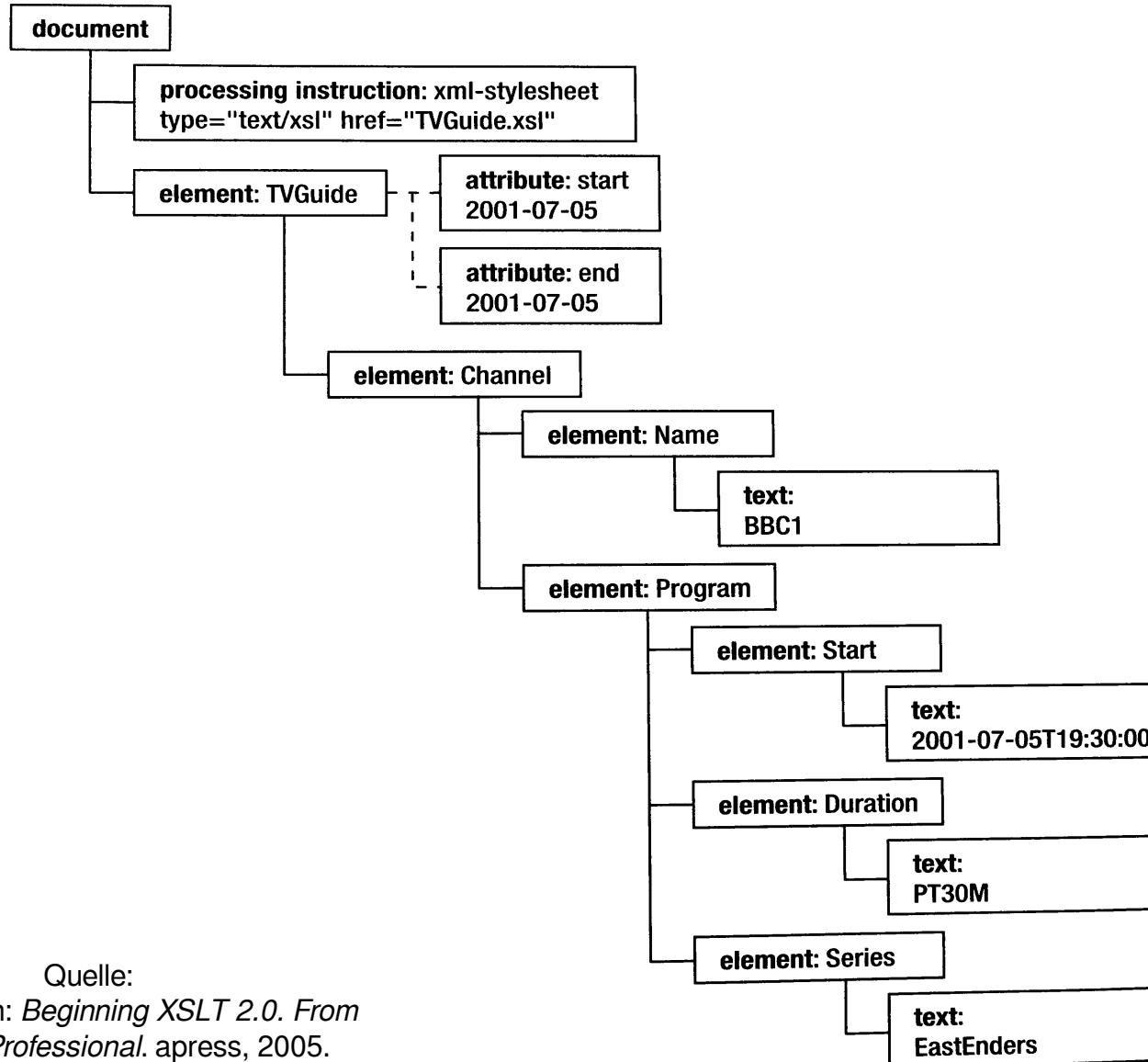
# Baumstruktur als Diagramm



# Baumstruktur: Kodierung

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet type="text/xsl" href="TVGuide.xsl"?>
<TVGuide start="2001-07-05" end="2001-07-05">
  <Channel>
    <Name>BBC1</Name>
    <Program>
      <Start>2001-07-05T19:00:00</Start>
      <Duration>PT30M</Duration>
      <Series>QuestionOfSport</Series>
      <Title></Title>
    </Program>
    <Program rating="5" flag="favorite">
      <Start>2001-07-05T19:30:00</Start>
      <Duration>PT30M</Duration>
      <Series>EastEnders</Series>
    </Program>
  </Channel>
</TVGuide>
```

# Baumstruktur: Diagramm



Quelle:

Jeni Tennison: *Beginning XSLT 2.0. From Novice to Professional.* apress, 2005.

## **3.2 DTDs und Schemas für XML-Dokumente**

# Dokumentstruktur

Festgelegt in einem Schema, nach verschiedenen Standards:

- Document Type Definition
  - <http://www.w3.org/TR/REC-xml/#dt-doctype>
- Relax NG
  - [http://standards.iso.org/ittf/PubliclyAvailableStandards/c052348\\_ISO\\_IEC\\_19757-2\\_2008\(E\).zip](http://standards.iso.org/ittf/PubliclyAvailableStandards/c052348_ISO_IEC_19757-2_2008(E).zip)
- W3C Schema
  - <http://www.w3.org/standards/techs/xmlschema>

# Beispiel Gedicht

Gaiku 123

early dew  
the water contains  
teaspoons of honey

(2009)

(Quelle: Yael Netzer et al., „Gaiku: Generating Haiku with word association norms, 2009)

# Konzeptionelles Datenmodell

- Titel
- Gedicht
- Zeilen
- Jahr

# Relax NG-Schema für „Haiku-Sammlung“

```
start = haikus
haikus = element haiku+
haiku = element haiku {head?, body}
head = element head {title?, year?}
title = element title {text}
year = element year {xsd:integer}
body = element body {stanza}
stanza = element stanza {line+}
line = element line {text}
```

# Nach Schema kodiertes Gaiku

```
haikus.xml x
1 <haikus>
2   <haiku>
3     <head>
4       <title>Gaiku 123</title>
5       <year>2009</year>
6     </head>
7     <body>
8       <stanza>
9         <line>early dew</line>
10        <line>the water contains</line>
11        <line>teaspoons of honey</line>
12      </stanza>
13    </body>
14  </haiku>
15 </haikus>
16
17
```

# well-formed vs. valid

- „well-formed“ (wohlgeformt)
  - Dokument entspricht den allgemeinen Prinzipien von XML
  - die Kriterien sind immer gleich
  - die Kriterien sind allgemein
- „valid“ (valide)
  - Dokument entspricht der Syntax und dem Lexikon eines spezifischen XML-Formats
  - Kriterien hängen von der jeweiligen Definition (DTD, Schema) ab
  - Kriterien sind meist sehr detailliert

# Kriterien für "Wohlgeformtheit"

- Prolog: XML-Version, Zeichensatz
- Nur ein Element auf oberster Ebene
- Jedes Element hat Anfangs- und Endtag
- Hierarchische Struktur: keine überlappenden Elemente
- Elemente können Unterelemente haben
- Elemente können Attribute haben
- Attribute können Werte haben
- Die Werte sind in Anführungszeichen gesetzt
- Alle Zeichen entsprechen dem ang. Zeichensatz

# Noch ein paar Regeln

- kann auch Kommentare enthalten  
`<!-- Kommentar -->`
- kann auch „processing instructions“ enthalten  
`<? Anweisung ?>;` Beispiel Prolog
- Element-Namen sind „case sensitive“:  
`<name> ≠ <Name>`
- Leere Elemente können abgekürzt werden:  
`<pb></pb> = <pb/>`

# Kriterien für "Validität"

- wohlgeformt
  - Definition (Schema/DTD) vorhanden: intern/extern
  - Dokument entspricht der Definition
- 
- Alle notwenigen, nur erlaubte Elemente
  - Alle notwendigen, nur erlaubte Attribute
  - Alle Werte haben eine gültige Form/Ausprägung
  - Elemente und Attribute kommen nur dort vor, wo sie auch erlaubt sind

## **3.3 Ein paar weitere Themen**

# Namespace

- Erlaubt es, mehrere verschiedene Schemata in einem Dokument zu verwenden und eindeutige Elementnamen zu verwenden.
- Beispiel der Bestimmung eines Namespace für ein ganzes Dokument:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
```

# Beispiel: Namespaces

```
<html xmlns="http://www.w3.org/1999/xhtml"
      xmlns:svg="http://www.w3.org/2000/svg">
<head>
  <title>Beispiel mit mehreren Namensräumen</title>
</head>
<body>
  <h1>Eine Mathe-Formel:</h1>
  <math xmlns="http://www.w3.org/1998/Math/MathML">
    <mi>x</mi><mo>=</mo><mn>2</mn>
  </math>
  <p>Und noch ein kleines Bild dazu:</p>
  <svg:svg>
    <svg:rect x="0" y="0" width="10" height="10" />
  </svg:svg>
</body>
</html>
```

# Überlappendende Hierarchien

```
<lg>
<l><s>Er streckt ins Dunkel seine
Fleischerfaust.</s></l>
<l><s>Er schüttelt sie.</s><s>Ein Meer von
Feuer jagt </l>
<l>Durch eine Straße.</s><s> Und der Glutqualm
braust</l>
<l>Und frißt sie auf, bis spät der Morgen
tagt.</s></l>
</lg>
```

# Lösungsansätze

## Zerlegen und Verbinden

<l><s>Er schüttelt sie.</s><s>Ein Meer von Feuer jagt</s> </l>

<l><s>Durch eine Straße.</s><s>Und der Glutqualm braust</s></l>

## Leeres Element mit Verweis

<l><s>Er schüttelt sie.</s><anchor subtype="sentStart"/>Ein Meer von Feuer jagt </l>

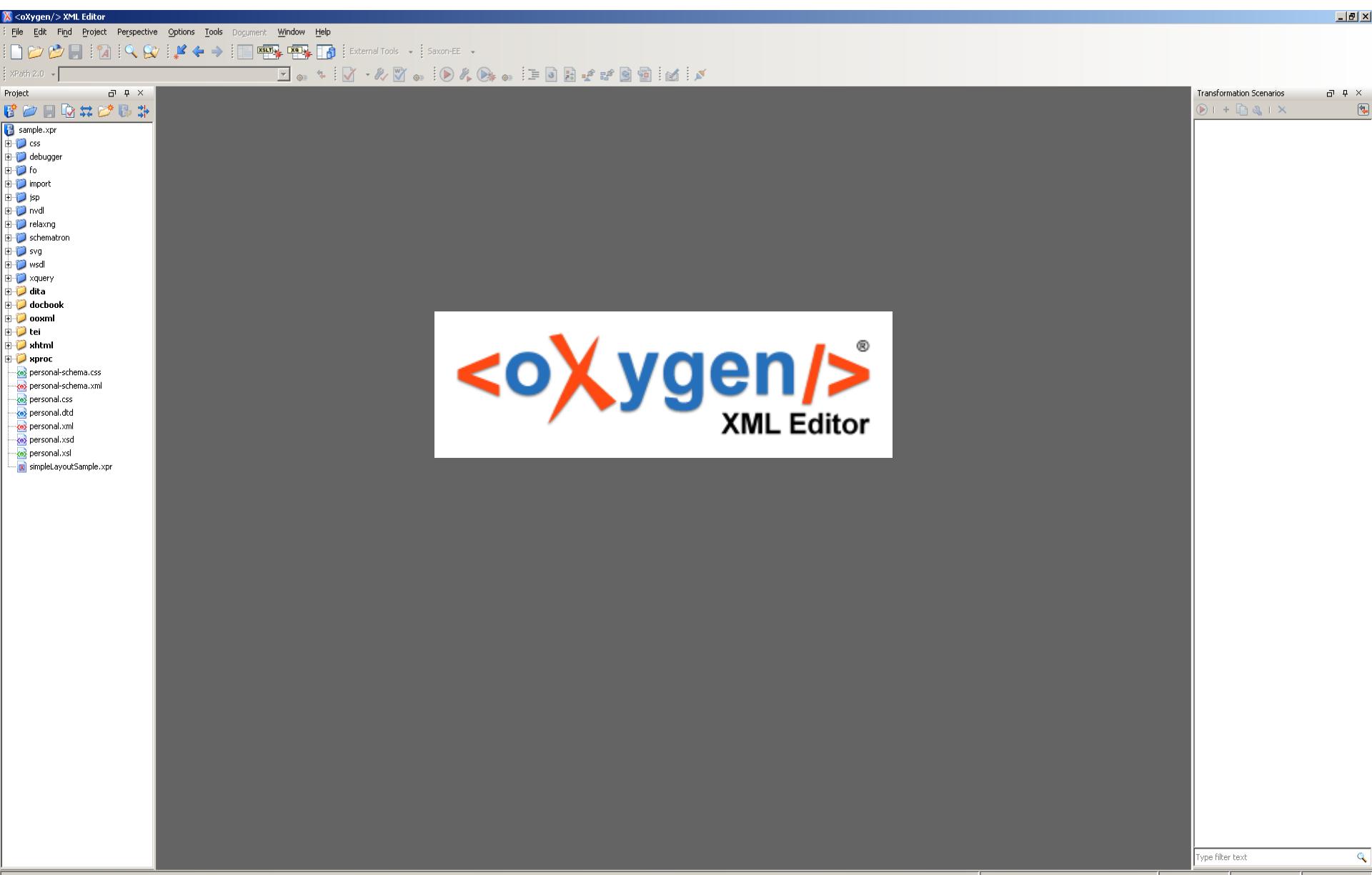
<l>Durch eine Straße.<anchor subtype="sentEnd"/><anchor subtype="sentStart"/>Und der Glutqualm braust</l> ... <anchor subtype="sentEnd"/>

# X-Technologien im Überblick

- Technologien zur Definition von XML-Formaten
  - DTD (Document Type Definition)
  - Schema (z.B. Relax NG)
- Verwandte Technologien für Transformation und Analyse
  - XSL
  - XSLT
  - XPath (=> wichtig für Suche mit lxml!)
- Mit XML definierte spezielle Markup Languages
  - xHTML
  - MathML, MusicXML
  - TEI, MEI , CEI, ...

# **4. Praxis der Textkodierung**

# oXygen: XML-Editor



# oxygen: „wohlgeformt?“

The screenshot shows the oXygen XML Editor interface with the following details:

- Title Bar:** Untitled2.xml [D:\vorlesung\Untitled2.xml] - <oXygen/> XML Editor
- Menu Bar:** File, Edit, Find, Project, Perspective, Options, Tools, Document, Window, Help
- Toolbar:** Includes icons for file operations, search, and various XML-related functions.
- Project Explorer:** Shows a project structure with files like sample.xpr, css, debugger, fo, import, jsp, nvd1, relaxing, schematron, svg, wsdl, xquery, dita, docbook, ooxml, tei, xhtml, xproc, and personal-schema.rnc.
- XML Editor:** The main workspace displays the XML code:

```
<?xml version="1.0" encoding="UTF-8"?>
<person>
    <vorname>Alan</vorname>
    <nachname>Turing</nachname>
    <beruf>Informatiker</bruf>
    <beruf>Mathematiker<beruf>
    <beruf>Kryptograph</beruf>
</person>
```

The line containing the invalid element `<beruf>Mathematiker<beruf>` is highlighted with a red underline, indicating a syntax error.
- Outline View:** Shows a tree structure of the XML document nodes.
- Attributes View:** A panel on the right showing attributes for the selected node, currently empty.
- Transformation Scenarios:** A panel showing available transformation scenarios: XML transformation with XSLT, XML transformation with XQUERY, DITA OT transformation, XSLT transformation, XProc transformation, XQUERY transformation (with sub-options Execute XQuery and Execute SQL), and SQL transformation (with sub-option Execute SQL).
- Status Bar:** Shows the file path D:\vorlesung\Untitled2.xml, character position U+0066, time 5:14, and status Modified.
- Message Bar:** F [Xerces] The element type "beruf" must be terminated by the matching end-tag "</beruf>".

# oXygen: „wohlgeformt?“

The screenshot shows the oXygen XML Editor interface with the following details:

- Title Bar:** Untitled2.xml [J:\vorlesung\Untitled2.xml] - <oXygen/> XML Editor
- Menu Bar:** File, Edit, Find, Project, Perspective, Options, Tools, Document, Window, Help
- Toolbar:** Includes icons for file operations, search, and various XML-related functions.
- Project Explorer:** Shows a project structure with files like sample.xpr, css, debugger, fo, import, jsp, nvd1, relaxing, schematron, svg, wsdl, xquery, dita, docbook, ooxml, tei, xhtml, and xproc.
- Outline View:** Displays the XML structure with nodes like person, vorname, nachname, and beruf.
- XML Editor:** The main pane contains the XML code:

```
<?xml version="1.0" encoding="UTF-8"?>
<person>
    <vorname>Alan</vorname>
    <nachname>Turing</nachname>
    <beruf>Informatiker</beruf>
    <beruf>Mathematiker</beruf>
    <beruf>Kryptograph</beruf>
</person>
```

A red circle highlights the word "Informatiker" in the fifth line. A blue arrow points from the right margin to the start of the word "Informatiker". Another blue arrow points from the right margin to the end of the word "beruf" in the same line.
- Attributes View:** Shows attributes for the beruf element, with "beruf" listed under the "Attribute" column.
- Transformation Scenarios:** A tree view showing transformation scenarios: XML transformation with XSLT, XML transformation with XQUERY, DITA OT transformation, XSLT transformation, XProc transformation, XQUERY transformation (with Execute XQuery and SQL transformation), and Execute SQL.
- Status Bar:** J:\vorlesung\Untitled2.xml, U+0066, 5:14, Modified.
- Message Bar:** F [Xerces] The element type "beruf" must be terminated by the matching end-tag "</beruf>".

F [Xerces] The element type "beruf" must be terminated by the matching end-tag "</beruf>".  
Text Grid Author

# oXygen: „wohlgeformt?“

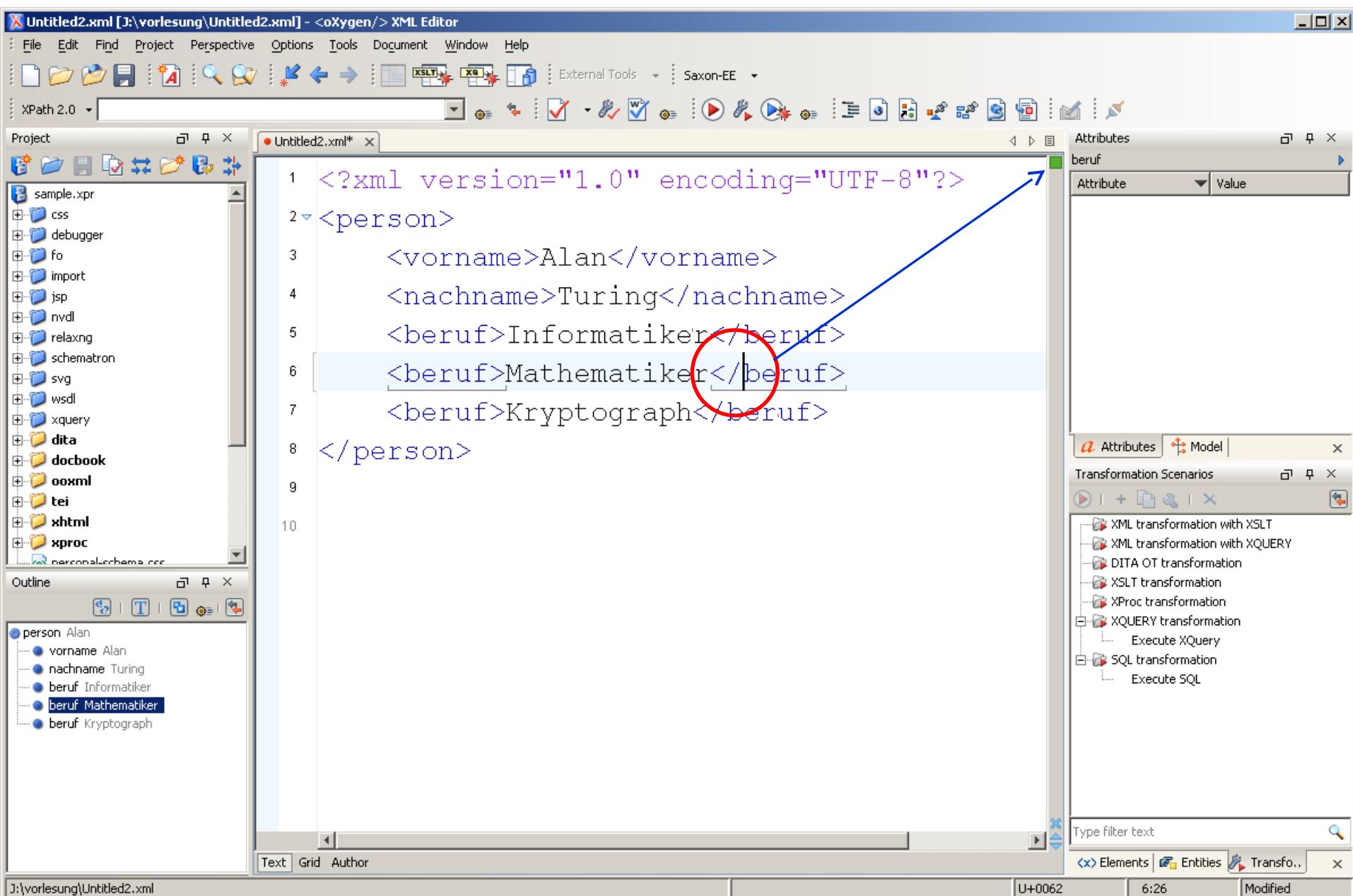
The screenshot shows the oXygen XML Editor interface with the following details:

- Title Bar:** Untitled2.xml [D:\vorlesung\Untitled2.xml] - <oXygen/> XML Editor
- Menu Bar:** File, Edit, Find, Project, Perspective, Options, Tools, Document, Window, Help
- Toolbar:** Includes icons for file operations, search, and various XML-related functions.
- Project Explorer:** Shows a sample.xpr project with various schema files like css, debugger, fo, import, jsp, nvd1, relaxing, schematron, svg, wsdl, xquery, dita, docbook, ooxml, tei, xhtml, and xproc.
- Outline View:** Displays the XML structure with nodes like person, vorname, nachname, beruf, and their values.
- Main Editor:** Contains the XML code:

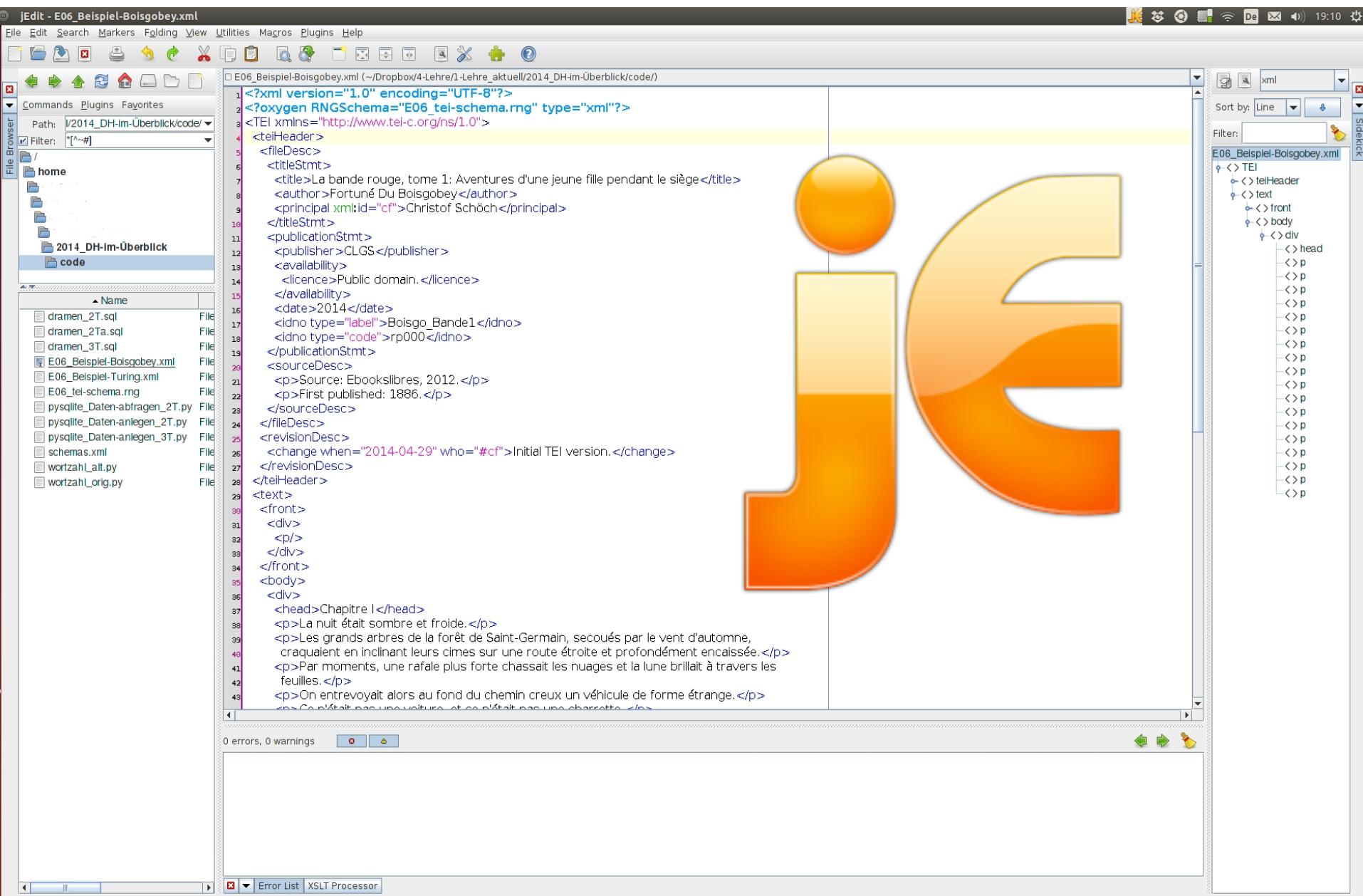
```
<?xml version="1.0" encoding="UTF-8"?>
<person>
    <vorname>Alan</vorname>
    <nachname>Turing</nachname>
    <beruf>Informatiker</beruf>
    <beruf>Mathematiker</beruf>
    <beruf>Kryptograph</beruf>
</person>
```

A blue arrow points from the end of the third 'beruf' element to the end of the document, while another blue arrow points from the start of the fourth 'beruf' element to its closing tag.
- Attributes View:** Shows attributes for the person element.
- Transformation Scenarios:** A tree view showing available transformations: XML transformation with XSLT, XML transformation with XQUERY, DITA OT transformation, XSLT transformation, XProc transformation, XQUERY transformation (with Execute XQuery), and SQL transformation (with Execute SQL).
- Status Bar:** Shows the file path D:\vorlesung\Untitled2.xml, a warning icon, and the status "Modified".
- Bottom Status Bar:** Shows the message "F [Xerces] The element type "beruf" must be terminated by the matching end-tag "</beruf>"."

# oxygen: „wohlgeformt?“



# Alternative: jEdit mit XML-Plugin



The image shows the jEdit text editor interface with an XML file open. The file is named E06\_Beispiel-Boisgobey.xml and contains XML code for a TEI document. The code includes details about the book "La bande rouge, tome 1: Aventures d'une jeune fille pendant le siège", its author Fortuné Du Boisgobey, and its publisher CLGS. It also includes publication information and a front page description.

```
<?xml version="1.0" encoding="UTF-8"?>
<oxygen RNGSchema="E06_tel-schema.rng" type="xml"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <telHeader>
    <fileDesc>
      <titleStmt>
        <title>La bande rouge, tome 1: Aventures d'une jeune fille pendant le siège</title>
        <author>Fortuné Du Boisgobey</author>
        <principal xml:id="#cf">Christof Schöch</principal>
      </titleStmt>
      <publicationStmt>
        <publisher>CLGS</publisher>
        <availability>
          <licence>Public domain.</licence>
        </availability>
        <date>2014</date>
        <idno type="label">Boisgo_Bande1</idno>
        <idno type="code">rp000</idno>
      </publicationStmt>
      <sourceDesc>
        <p>Source: Ebookslibres, 2012.</p>
        <p>First published: 1886.</p>
      </sourceDesc>
    </fileDesc>
    <revisionDesc>
      <change when="2014-04-29" who="#cf">Initial TEI version.</change>
    </revisionDesc>
  </telHeader>
  <text>
    <front>
      <div>
        <p>La nuit était sombre et froide.</p>
        <p>Les grands arbres de la forêt de Saint-Germain, secoués par le vent d'automne,<br/> craquaient en inclinant leurs cimes sur une route étroite et profondément encassée.</p>
        <p>Par moments, une rafale plus forte chassait les nuages et la lune brillait à travers les feuilles.</p>
        <p>On entrevoyait alors au fond du chemin creux un véhicule de forme étrange.</p>
        <p><!-- C'était pas une voiture, et ce n'était pas une charrette. --></p>
      </div>
    </front>
    <body>
      <div>
        <head>Chapitre I</head>
        <p>La nuit était sombre et froide.</p>
        <p>Les grands arbres de la forêt de Saint-Germain, secoués par le vent d'automne,<br/> craquaient en inclinant leurs cimes sur une route étroite et profondément encassée.</p>
        <p>Par moments, une rafale plus forte chassait les nuages et la lune brillait à travers les feuilles.</p>
        <p>On entrevoyait alors au fond du chemin creux un véhicule de forme étrange.</p>
        <p><!-- C'était pas une voiture, et ce n'était pas une charrette. --></p>
      </div>
    </body>
  </text>

```

The right side of the interface features a large, stylized watermark with the letters "je". The status bar at the bottom indicates "0 errors, 0 warnings".

# **4. Suchen in XML-Dokumenten mit XPath**

# Was ist XPath?

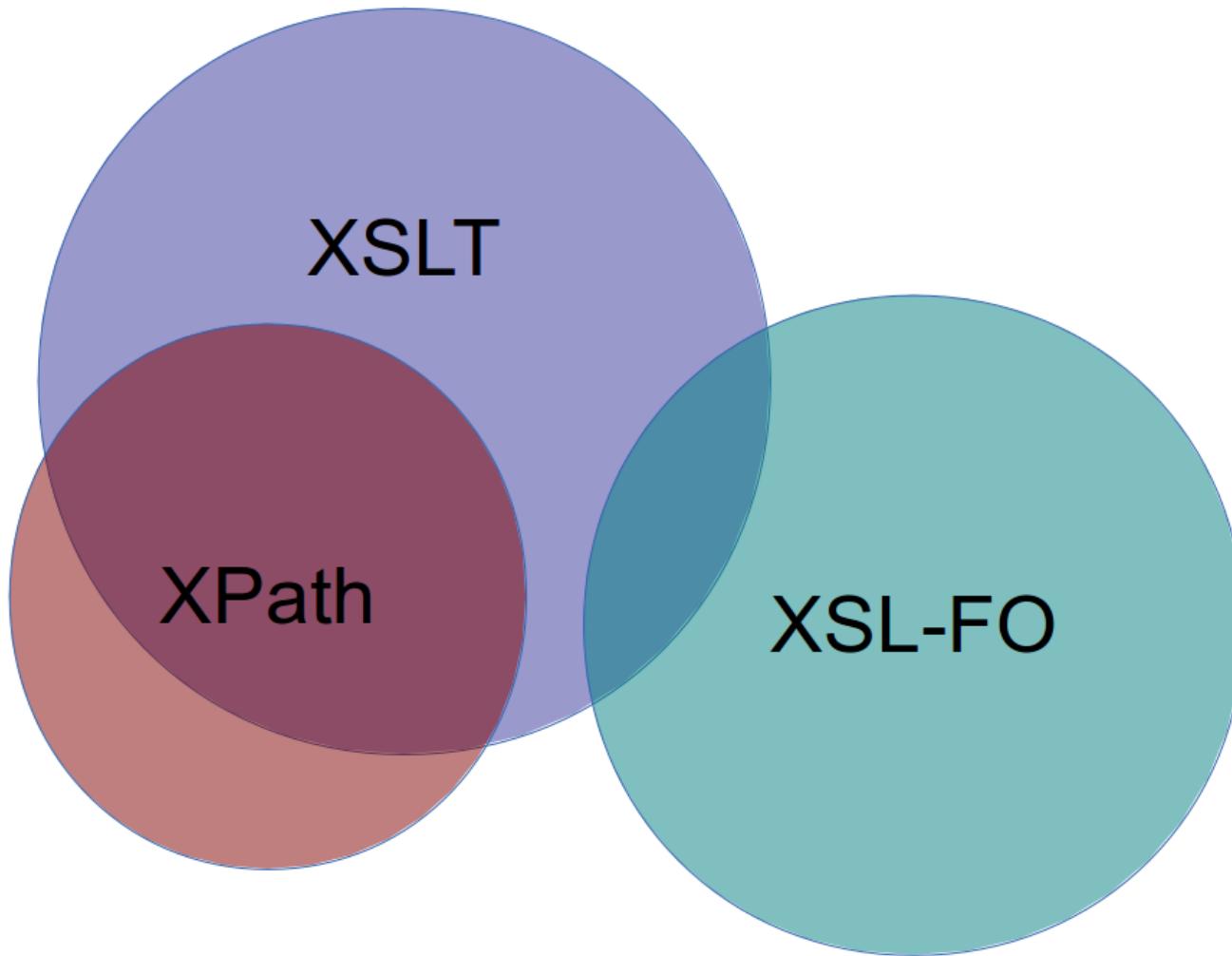
## Grundidee

- "XPath is a language for addressing parts of an XML document (XPath Specifications)
- XPath wird u.a. in XSLT eingesetzt

## Spezifikation

- Die XPath-Spezifikation wird vom W3C (World Wide Web Consortium) gepflegt
- <http://www.w3.org/TR/xpath/>

# Xpath im Kontext



# Beispiele für XPath

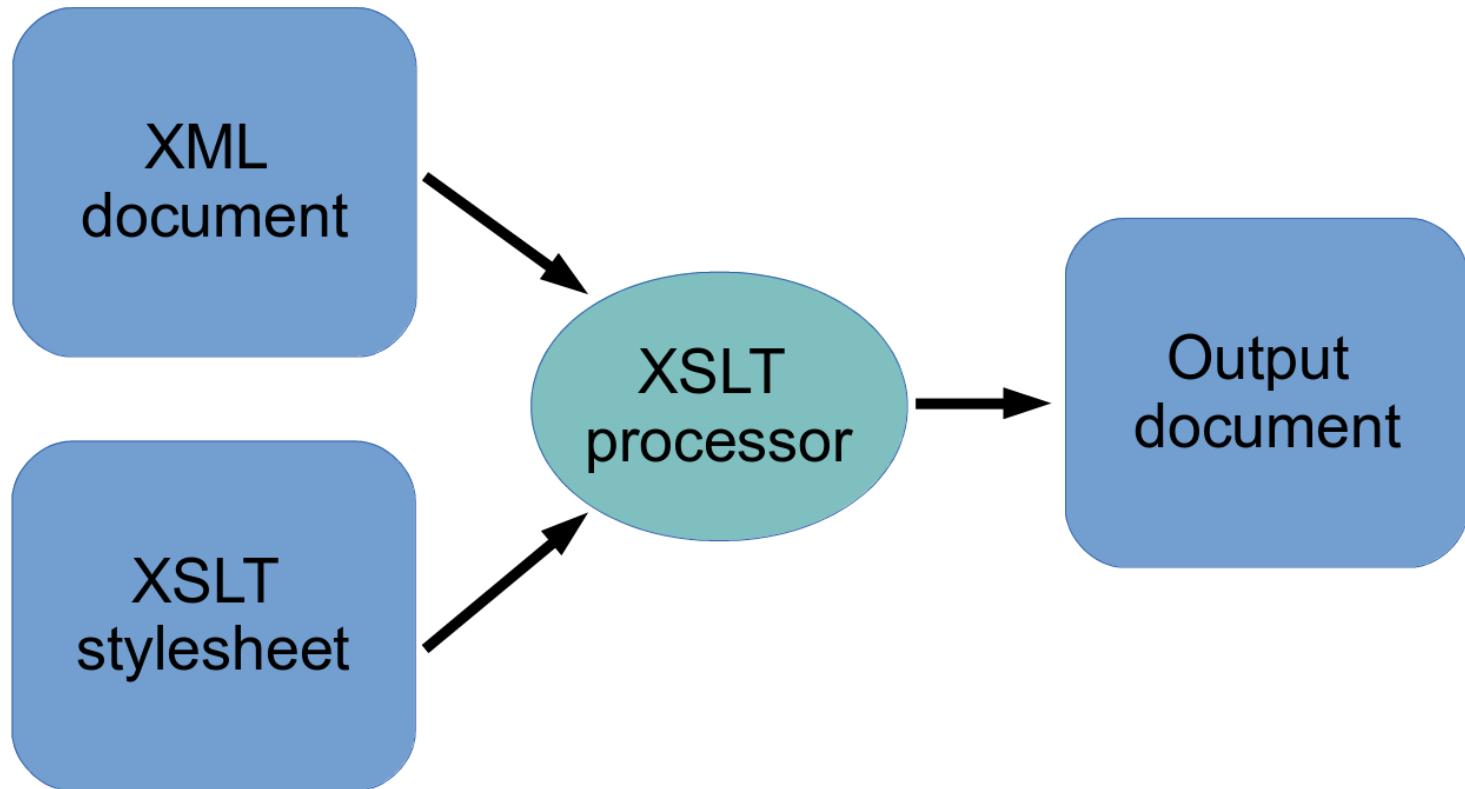
## Prinzip

- Lokalisierungsschritte: Achse + Knotentest (+ Prädikat)

## Beispiele

- /person/name/
- /person/name/vorname
- /person/name/vorname/text()
- /person/child::text()
- /person/beruf[@type="main"]/text()
- /person/beruf[2]

# XML + XSLT = Dokument



# **Literaturhinweise**

# Lektürehinweise

## Referenzlektüre

- Georg Vogeler und Patrick Sahle: „XML“, in: Digital Humanities: Eine Einführung, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, 2017, 128-146.

## Weitere Empfehlungen

- Vonhoegen, Helmut. *Handbuch: Einstieg in XML. Grundlagen, Praxis, Referenz*. Bonn, 2015.
- Allen Renear: „Text Encoding“. In: Susan Schreibman et. al., eds.: *Companion to Digital Humanities*. 2006.
- Christof Schöch: „Ein digitales Textformat für die Literaturwissenschaft. Die Richtlinien der Text Encoding Initiative und ihr Nutzen für Textedition und Textanalyse“, *Romanische Studien* 4 (2016), S. 325–364, URL: <http://www.romanischedestudien.de/index.php/rst/article/view/58>.
- Melissa Terras, Edward Vanhoutte, Ron Van den Branden, hg.: *TEI by Example*. URL: <http://teibyexample.org/> (interaktives Tutorial)