Textauszeichnung

Prof. Dr. Christof Schöch

Modul Auszeichnungssprachen
MSc. Digital Humanities, Universität Trier

Universität Trier

Überblick

1. Markup oder Tabelle?
2. Grundlagen von XML
3. XML und das OHCO-Modell von Text
4. Textkodierung in der Praxis: Editoren

(1) Markup oder Tabelle?

Beispiel: plain text

Harry Potter and the Sorcerer's Stone
CHAPTER ELEVEN - QUIDDITCH
As they entered November, the weather turned very cold. The mountains around the school became icy gray and the lake like chilled steel.
"Library books are not to be taken outside the school," said Snape.

Beispiel: tabellarisch (Annotation)

	A	B	C	D	E
1	wordform	pos	lemma		
2	Harry	NP0	harry		
3	Potter	NP0	potter		
4	and	CJC	and		
5	the	AT0	the		
6	Sorcerer	NN1	sorcerer		
7	's	POS	's		
8	Stone	NN1	stone		
9	CHAPTER	NN1	chapter		
10	ELEVEN	CRD	eleven		
11	-	PUN	-		
12	QUIDDITCH	AJ0	QUIDDITCH		
13	As	CJS	as		
14	they	PNP	they		
15	entered	VVD	enter|entered		
16	November	NP0	november		
17	,	PUN	,		
18	the	AT0	the		

Beispiel: Markup (Metadaten, Struktur)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="eng">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="book">Harry Potter and the Sorcerer's Stone</title>
        <author>Rowling, Joan K.</author>
      </titleStmt>
      <publicationStmt>
        <p>For internal use only.</p>
      </publicationStmt>
    </fileDesc>
  </teiHeader>
  <text>
    <body/>
    <div type="chapter">
      <head>CHAPTER ELEVEN - QUIDDITCH</head>
      <p>As they entered November, the weather turned very cold. The mountains
      around the school became icy gray and the lake like chilled steel.</p>
      <p><said who="Snape">"Library books are not to be taken outside the school,"
      </said> said Snape.</p>
    </div>
  </text>
</TEI>
```

Beispiel: Tabelle (Annotation + Struktur)

	A	B	C	D	E
1	wordform	pos	lemma	sentence	direct-speech
2	Harry	NP0	harry	1	0
3	Potter	NP0	potter	1	0
4	and	CJC	and	1	0
5	the	AT0	the	1	0
6	Sorcerer	NN1	sorcerer	1	0
7	's	POS	's	1	0
8	Stone	NN1	stone	1	0
9	CHAPTER	NN1	chapter	2	0
10	ELEVEN	CRD	eleven	2	0
11	-	PUN	-	2	0
12	QUIDDITCH	AJ0	QUIDDITCH	2	0
13	As	CJS	as	3	0
14	they	PNP	they	3	0
15	entered	VVD	enter|entered	3	0
16	November	NP0	november	3	0
17	,	PUN	,	3	0
18	the	AT0	the	3	0

Beispiel: Markup (Struktur + Annotation)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="eng">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="book">Harry Potter and the Sorcerer's Stone</title>
        <author>Rowling, Joan K.</author>
      </titleStmt>
      <publicationStmt>
        <p>For internal use only.</p>
      </publicationStmt>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <div type="chapter">
        <head>
          <s>
            <w pos="NN1" lemma="chapter">CHAPTER</w>
            <w pos="CRD" lemma="eleven">ELEVEN</w>
            <w pos="PUN" lemma="-">-</w>
            <w pos="AJ0" lemma="quidditch">QUIDDITCH</w>
          </s>
        </head>
        <p>
          <s>
            <w pos="CJS" lemma="as">As</w>
            <w pos="PNP" lemma="they">they</w>
            <w pos="VVD" lemma="enter">entered</w>
            <w pos="NP0" lemma="november">November</w>
```

(2) Grundlagen von XML

Eine Definition von Markup

We define markup, or (synonymously) encoding, as any means of making explicit an interpretation of a text. [...] Encoding a text for computer processing is, in principle, like transcribing a manuscript from *scriptio continua*; it is a process of making explicit what is conjectural or implicit, a process of directing the user as to how the content of the text should be (or has been) interpreted.

By markup language we mean a set of markup conventions used together for encoding texts. A markup language must specify how markup is to be distinguished from text, what markup is allowed, what markup is required, and what the markup means. XML provides the means for doing the first three; documentation such as these Guidelines is required for the last.

(Quelle: "A Gentle Introduction to XML", in: Guidelines of the TEI)

Was ist XML?

- eXtensible Markup Language (eXtensible Meta-Language?)
- Metasprache / Metasyntax zur Definition von XML-Formaten
- Standard für digitale Repräsentation von Daten
- Prinzipien + Syntax (aber kein Vokabular)
- einfach (wenige, allgemeine, mächtige Mechanismen)
- anwendungs- und plattformunabhängig
- W3C-Standard

XML-basierte Sprachen

XML-basierte Sprachen

- Grundprinzipien
  - jede XML-basierte ML respektiert die Syntax von XML
  - jede XML-basierte ML definiert ein eigenes Vokabular
  - Das Vokabular umfasst die Menge der spezifischen Elementtypen und Attribute;
  - Das Vokabular kann auch zusätzliche syntaktische Regeln bestimmen
  - die XML-basierte ML eignet sich für (einen oder mehrere) bestimmte Dokumenttypen
  - die XML-basierte ML ist in einem Schema oder einer DTD definiert

XML-basierte Sprachen

- Grundprinzipien
  - jede XML-basierte ML respektiert die Syntax von XML
  - jede XML-basierte ML definiert ein eigenes Vokabular
  - Das Vokabular umfasst die Menge der spezifischen Elementtypen und Attribute;
  - Das Vokabular kann auch zusätzliche syntaktische Regeln bestimmen
  - die XML-basierte ML eignet sich für (einen oder mehrere) bestimmte Dokumenttypen
  - die XML-basierte ML ist in einem Schema oder einer DTD definiert
- Beispiele für XML-basierte ML
  - TEI, MEI, CEI
  - MathML
  - MusicML
  - XHTML
  - SVG
  - uvm.

Bestandteile von XML: Elemente, Attribute, Werte, Strings

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="eng">
3    <teiHeader>
4      <fileDesc>
5        <titleStmt>
6          <title type="book">Harry Potter and the Sorcerer's Stone</title>
7          <author>Rowling, Joan K.</author>
8        </titleStmt>
9        <publicationStmt>
10         <p>For internal use only.</p>
11       </publicationStmt>
12     </fileDesc>
13   </teiHeader>
14   <text>
15     <body/>
16     <div type="chapter">
17       <head>CHAPTER ELEVEN - QUIDDITCH</head>
18       <p>As they entered November, the weather turned very cold. The mountains
         around the school became icy gray and the lake like chilled steel.</p>
19       <p><said who="Snape">"Library books are not to be taken outside the school,"
         </said> said Snape.</p>
20     </div>
21   </text>
22 </TEI>
```

Weitere Bestandteile

- Processing instructions (im Prolog)
- Entities (Platzhalter im Text, oder weil Teil des Markups)
- Leere Elemente
- Kommentare

```
<?Anweisung?>

<!-- Kommentar -->

Ich k&ouml;nnte glauben, da&szlig; Sie ...

<pb></pb> = <pb/>

&amp;

&lt;
```

Wohlgeformtheit und Validität

Wohlgeformtheit und Validität

- „well-formed“ (wohlgeformt)
  - Dokument entspricht den allgemeinen Prinzipien von XML
  - die Kriterien sind immer gleich
  - die Kriterien sind allgemein

Wohlgeformtheit und Validität

- „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

”Wohlgeformtheit” im Detail

- 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

”Validität” im Detail

- Dokument ist wohlgeformt (siehe oben)
- 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) XML und das OHCO-Modell von Text

Vorbemerkung: Vorteile deskriptiven Markups

Vorbemerkung: Vorteile deskriptiven Markups

- For creation / composition
  - "Composition is simplified"
  - "Structure-oriented editing is supported"
  - "More natural editing tools are supported"
  - Alternative document views are facilitated"

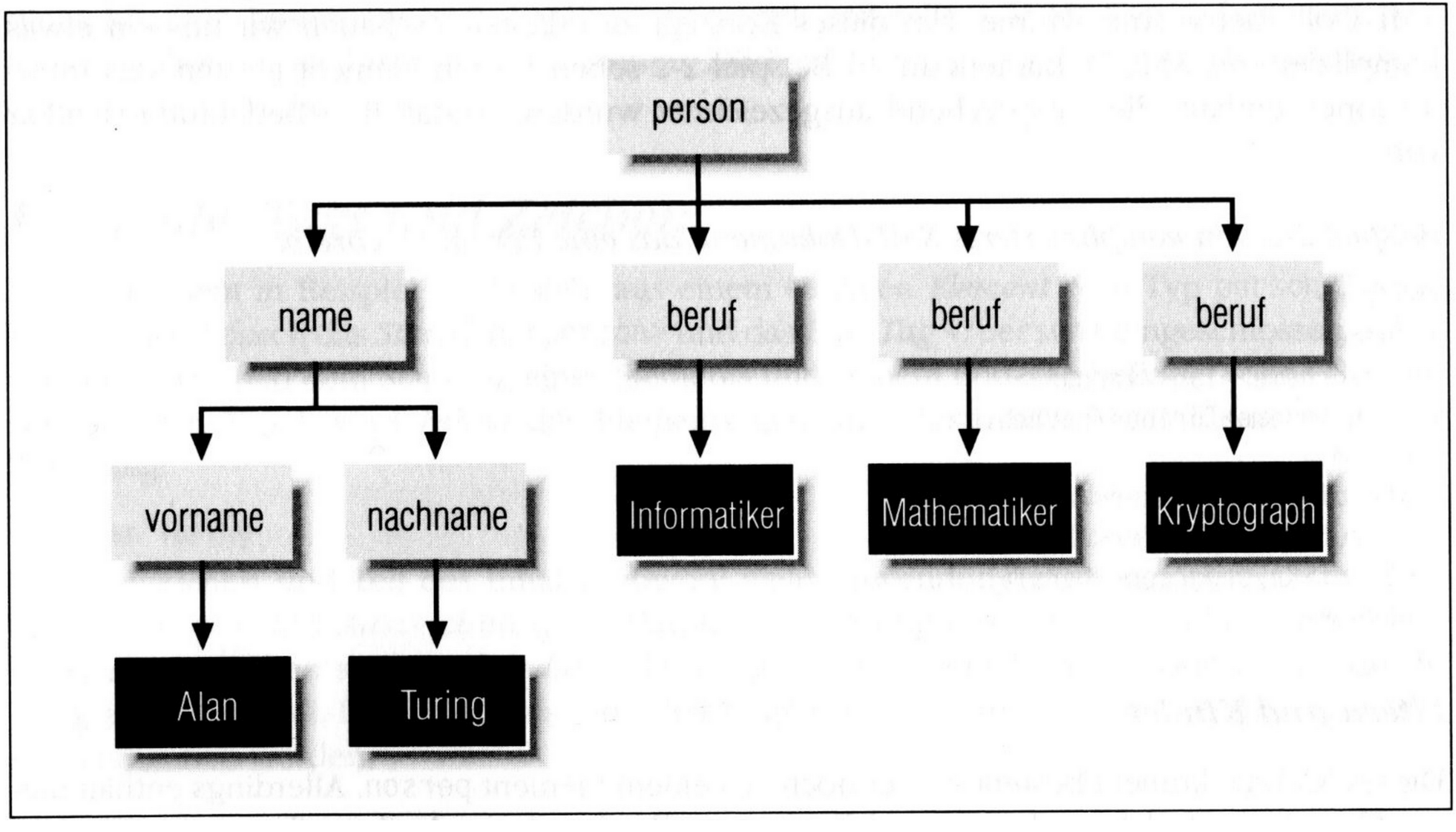
Vorbemerkung: Vorteile deskriptiven Markups

- For creation / composition
  - "Composition is simplified"
  - "Structure-oriented editing is supported"
  - "More natural editing tools are supported"
  - Alternative document views are facilitated"
- For publishing
  - Formatting can be generically specified and modified
  - Apparatus can be automated
  - Output device support is enhanced
  - Portability and interoperability are maximized

Vorbemerkung: Vorteile deskriptiven Markups

- For creation / composition
  - "Composition is simplified"
  - "Structure-oriented editing is supported"
  - "More natural editing tools are supported"
  - Alternative document views are facilitated"
- For publishing
  - Formatting can be generically specified and modified
  - Apparatus can be automated
  - Output device support is enhanced
  - Portability and interoperability are maximized
- For archiving, retrieval, and analysis
  - Information retrieval is supported
  - Analytical procedures are supported

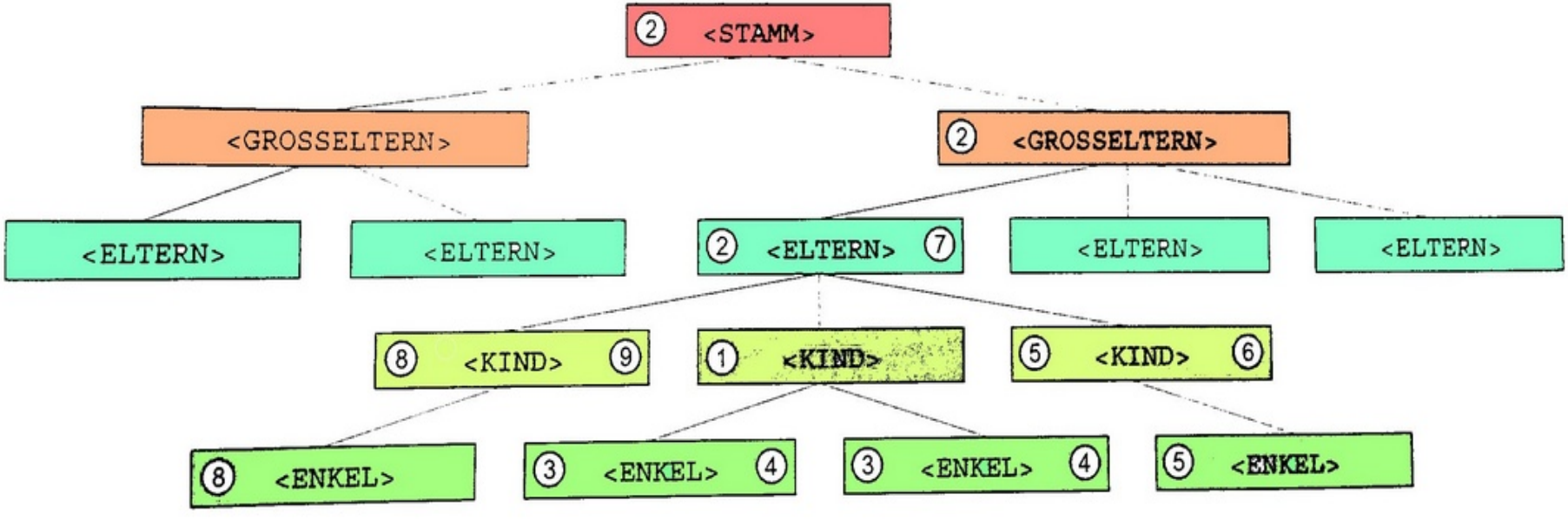
Die Baumstruktur von XML



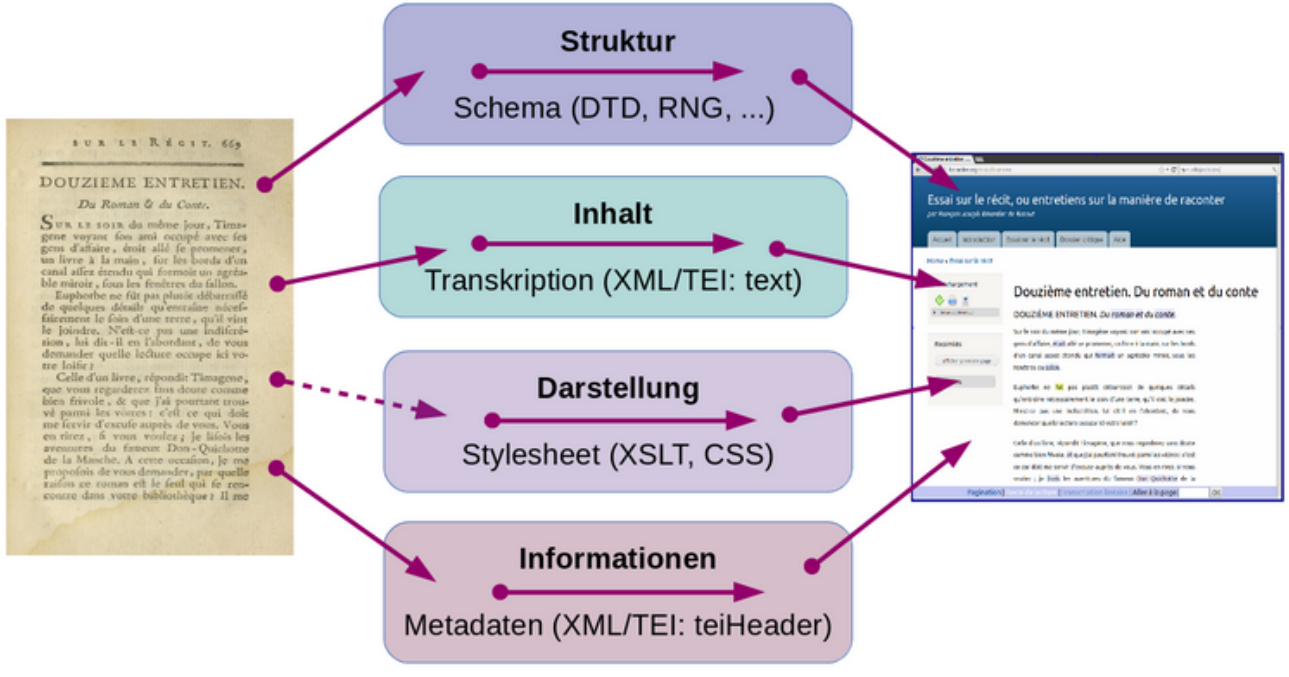
In XML kodiert

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

Baumstruktur abstrakt



Bestandteile eines Dokuments



Konzeptuelles Modell: Was ist Text?

The model [of text] in question postulates that text consists of objects of a certain sort, structured in a certain way. The nature of the objects is best suggested by example and contrast. They are chapters, sections, paragraphs, titles, extracts, equations, examples, acts, scenes, stage directions, stanzas, (verse) lines, and so on. But they are not things like pages, columns, (typographical) lines, font shifts, vertical spacing, horizontal spacing, and so on. The objects indicated by descriptive markup have an intrinsic direct connection with the intellectual content of the text; they are the underlying "logical" objects, components that get their identity directly from their role in carrying out and organizing communicative intention. The structural arrangement of these "content objects" seems to be hierarchical – they nest in one another without overlap. [...].

On this account then text is an "Ordered Hierarchy of Content Objects" (OHCO), and descriptive markup works as well as it does because it identifies that hierarchy and makes it explicit and available for systematic processing.

(Alan Renear, "Text Encoding", 2006)

Herausforderung: Überlappende Hierarchien

```
1
2 <lg>
3   <l><s>Er streckt ins Dunkel seine Fleischerfaust.</s></l>
4   <l><s>Er schüttelt sie.</s><s>Ein Meer von Feuer jagt</l>
5   <l>Durch eine Straße.</s><s> Und der Glutqualm braust</l>
6   <l>Und frißt sie auf, bis spät der Morgen tagt.</s></l>
7 </lg>
8
```

Lösungsansätze

- Zerlegen und Verbinden

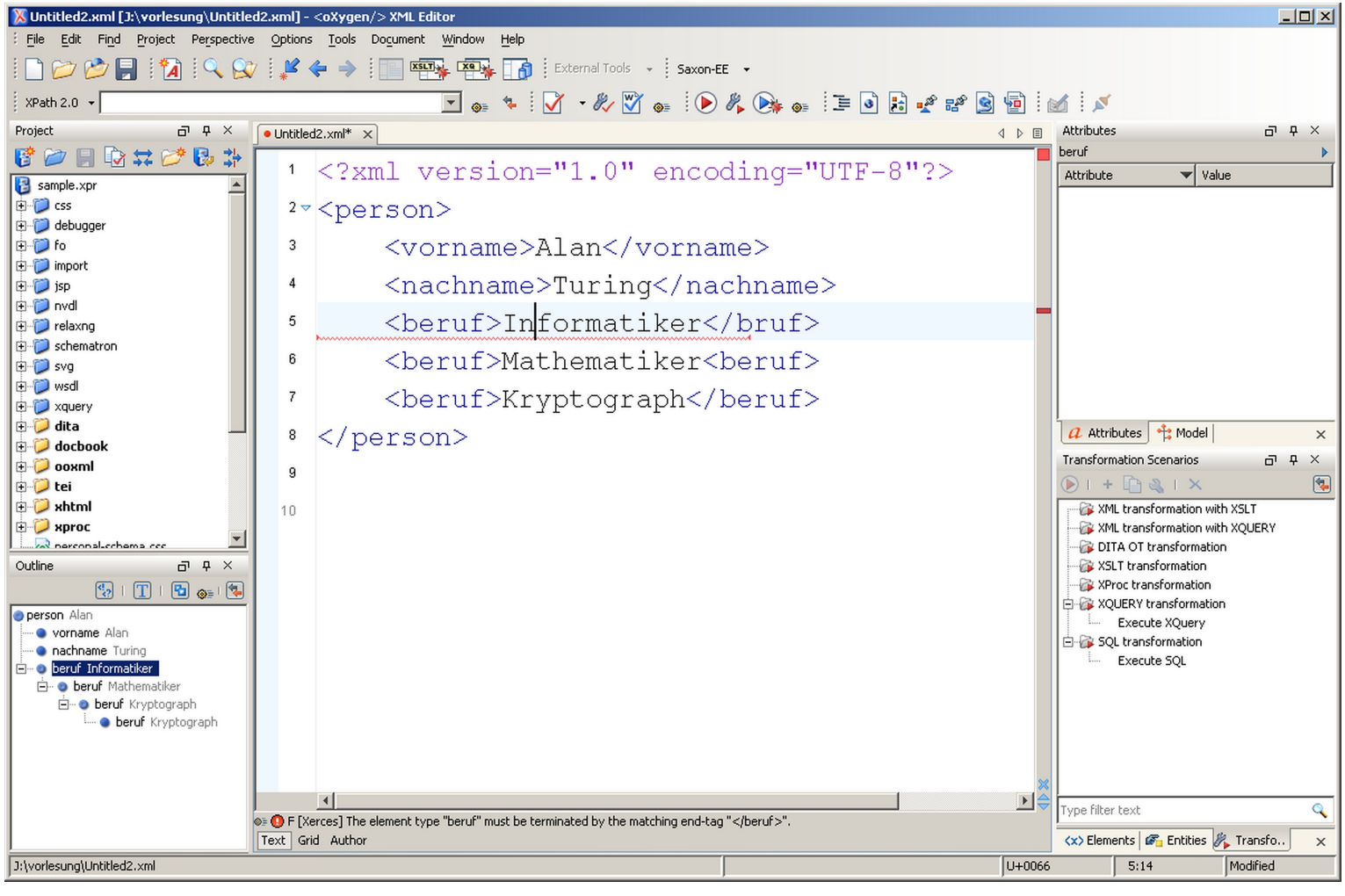
```
11 <l><s n="1">Er schüttelt sie.</s><s n="2">Ein Meer von Feuer jagt</s> </l>
12 <l><s corresp="#2">Durch eine Straße.</s><s n="3">Und der Glutqualm braust</s></l>
13
```

- Leeres Element mit Verweis

```
17 <l><s>Er schüttelt sie.</s><sStart/>Ein Meer von Feuer jagt</l>
18 <l>Durch eine Straße.<sEnd/><sStart/>Und der Glutqualm braust<sEnd/></l>
19
```

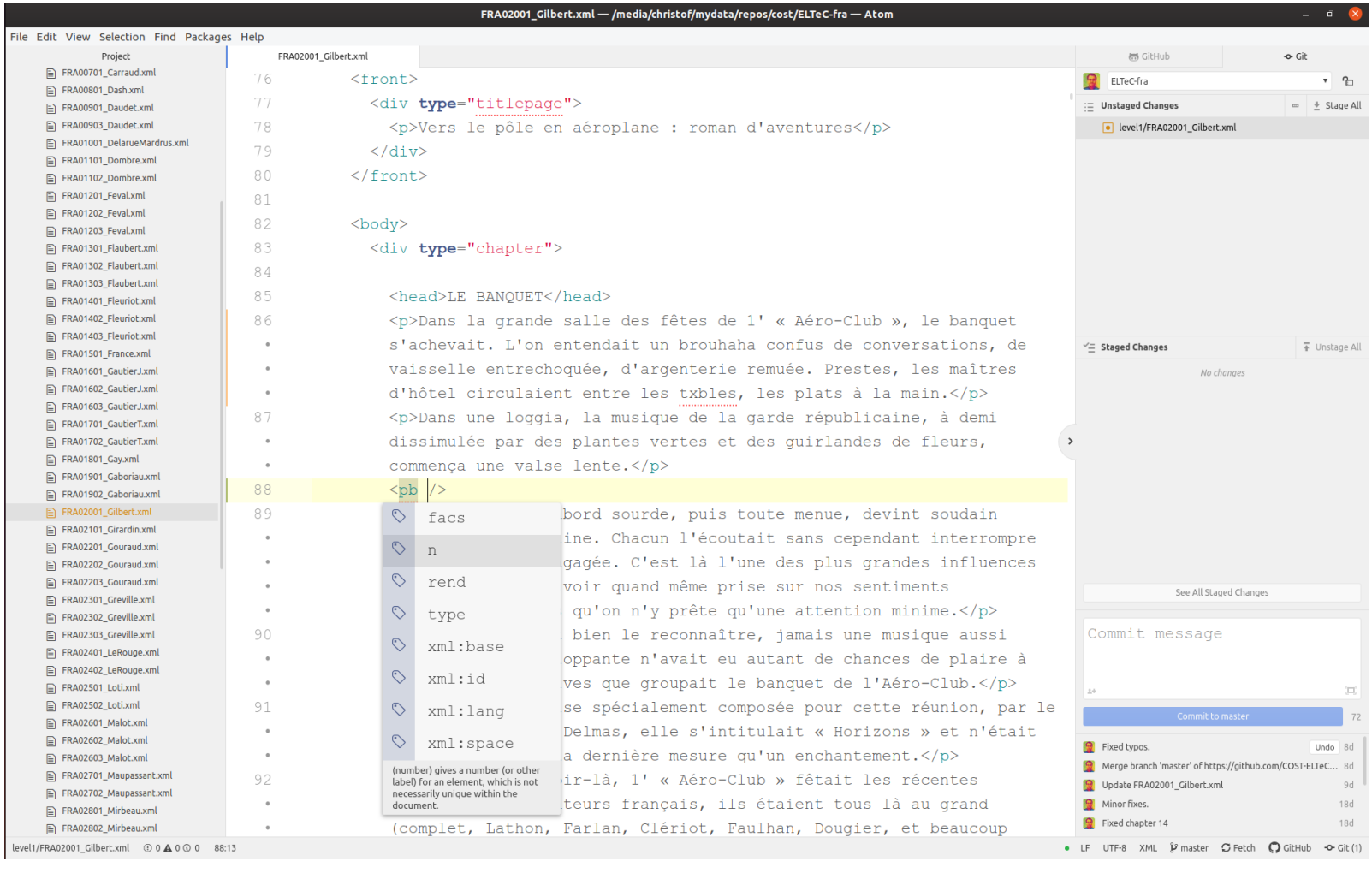
(4) Praxis der Textkodierung: Editoren

Der Klassiker: oXygen ($$$)



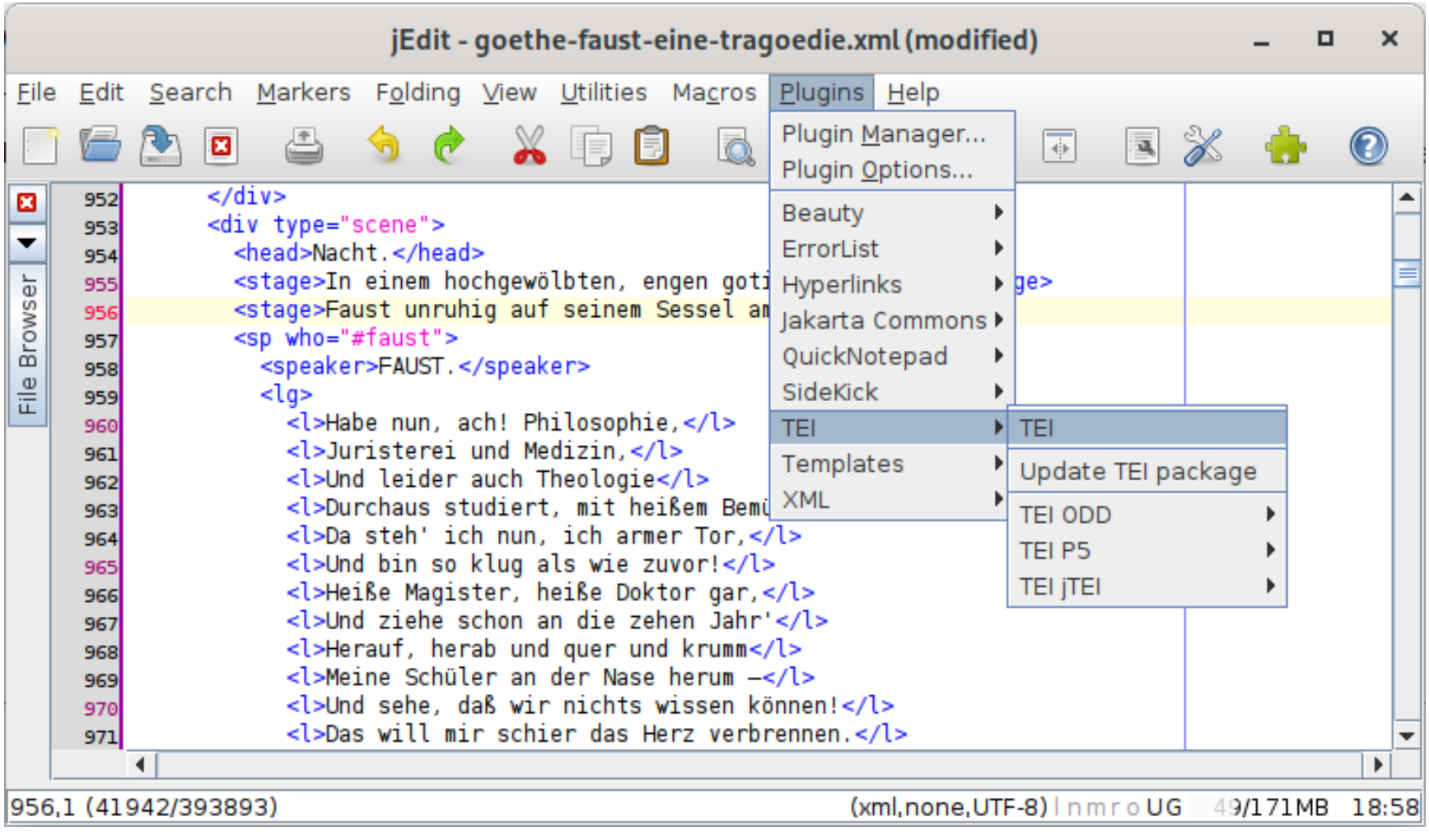
Siehe auch: Ediarum mit oXygen

Der Flexible: Atom (mit Plugins)



Siehe: Can Atom Replace oXygen?

Der Freie: JEdit (mit Plugins)



Siehe: TEI Plugin von DARIAH

Abschluss

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.
- "A Gentle Introduction to XML", in: *P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium, 2020. https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html
- Renear, Allen H.: „Text Encoding“. In: *Companion to Digital Humanities*, ed. Susan Schreibman et al. Oxford: Blackwell, 2006.

Danke!