UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Einführung in die Textauszeichnung mit TEI

<?xml version="1.0"?>

Agnes Brauer
a.brauer@ub.uni-frankfurt.de

Kursinhalt

- Was ist Textauszeichnung und wozu ist sie gut?
- Was ist XML?
- Was ist TEI?
- Wie geht man bei der Textauszeichnung mit TEI vor?
- Welche Ressourcen und Werkzeuge können hierfür verwendet werden?
- Wie annotiert man bestimmte Textsorten? (TEI-Guidelines Kap. 2, 4, 7)

Textauszeichnung

Textauszeichnung =

Kodieren der Strukturdaten eines Dokuments
– visueller oder logisch-semantischer Art –
mithilfe sogenannter Markup-Sprachen

Markup-Sprache?

Markup-Sprachen geben die **Struktur eines Dokuments** mithilfe geeigneter **Metadaten** wieder („Daten über die Daten“).

Diese Metadaten werden in Form von sog. **<tags>** (eigtl. 'Etikett', 'Schild') in die eigentlichen (Text-)Daten eingebracht

= **Annotation** oder eben **Auszeichnung** der Daten

Markup-Sprache?

Prominentestes Beispiel für eine Markup-Sprache:
HTML

HTML-Tags zielen dabei vordergründig auf die Abbildung visueller Aspekte eines Dokuments.

Beispiel:
HTML-Quellcode: ****Tim Berners-Lee****
=> Darstellung im Browser: *Tim Berners-Lee*

TEI?

Die Auszeichnungssprache TEI stellt eine spezielle, für bestimmte Zwecke definierte Markup-Sprache dar.

Sie beruht auf der **universellen Markup-Sprache XML**.

XML?

XML = E**x**tensible **M**arkup **L**anguage

- vom W3-Konsortium empfohlener Standard für einen an Nachhaltigkeit und Interoperabilität orientierten Umgang mit digitalen Dokumenten

- Fundamentaler Unterschied zu HTML: mithilfe von XML wird eine **inhaltliche Auszeichnung** angestrebt

- XML liegt das Prinzip der Trennung des Informationsgehalts eines Dokuments von dessen äußerer Form zugrunde
  = Konzept des **generic markup**

XML?

Das **generic markup** trifft Aussagen über die logische oder semantische **Bedeutung** der annotierten Textstelle:

Vgl.:

HTML: ****Tim Berners-Lee****

XML: **<name>**Tim Berners-Lee**</name>**

Wozu Markup-Sprachen?

Vorteil von Markup-Sprachen:

die annotierten Daten liegen als einfache, **plattformunabhängige Textdokumente** vor

Dies gewährleistet:

- langfristige Verfügbarkeit
- Unabhängigkeit von proprietärer Software
- flexiblen Datenaustausch

Wozu Markup-Sprachen?

… und speziell im Falle von XML:

- vielfältige Verwendbarkeit der Daten:

  flexibel zu definierende **Übersetzungssprachen** ermöglichen es, aus einem XML-Dokument mit geringem Aufwand ein
  - **HTML-Dokument** oder ein
  - **druckfertiges Manuskript** zu generieren, ebenso kann eine
  - **Datenbank** mit Inhalt angereichert werden

XML – eine Metasprache!

XML stellt jedoch nicht nur einfach eine Markup-Sprache dar, sondern fungiert darüber hinaus auch als **Metasprache**:

> XML „is a general-purpose markup language for creating special-purpose markup languages“ (BREITMAN et al. 2007, 30).

Mit der Syntax von XML können beliebig viele andere Markup-Sprachen definiert werden

Ein solches **XML-Derivat** ist zum Beispiel auch TEI

TEI?

- TEI ist eine **Auszeichnungssprache**, die sich als Standard für das digitale Auf- und Verarbeiten von Texten etabliert hat

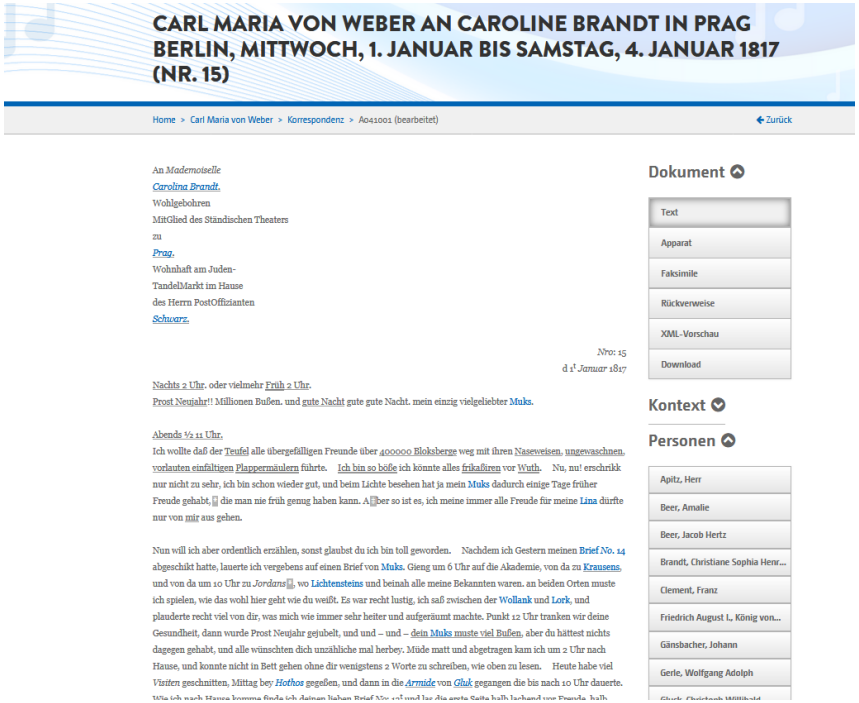
- TEI ist nach der **Text Encoding Initiative** benannt, einem Konsortium, das sich der kollaborativen Entwicklung und Pflege eines Standards zur Repräsentation von Texten in digitaler Form widmet

- Es stellt in seinen ***Guidelines*** Richtlinien für die Encodierung (= Auszeichnung) maschinenlesbarer Texte zusammen - insbesondere für die Zwecke der Geistes- und Kulturwissenschaften

Verwendungsbeispiel für TEI

TEI-annotierte Textdaten bilden die Grundlage
für anspruchsvolle, im wissenschaftlichen Kontext
entstandene Onlineprojekte

Beispiel:
Carl-Maria-von-Weber-Gesamtausgabe

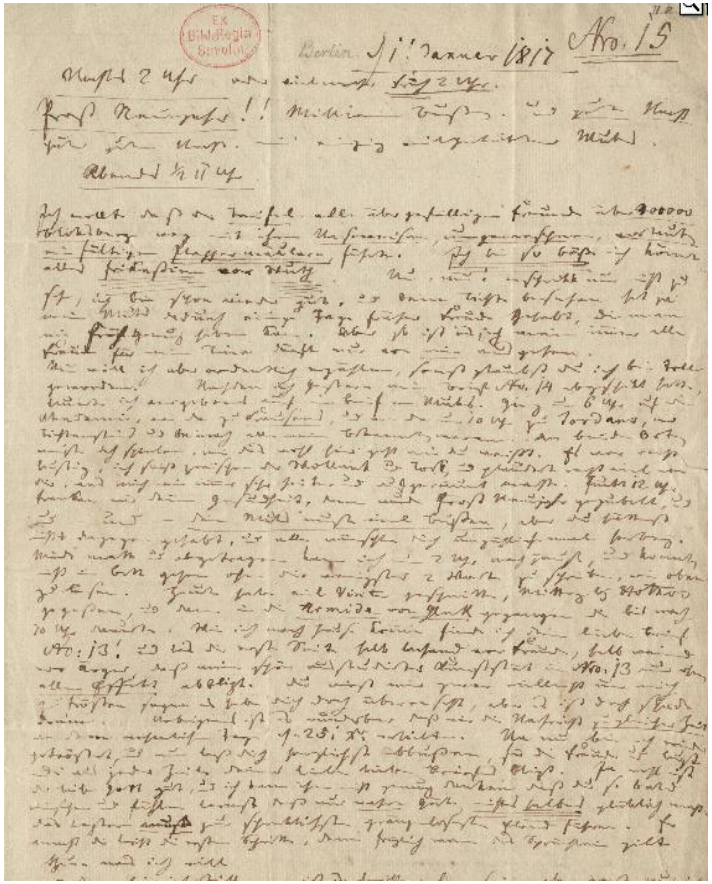


https://weber-gesamtausgabe.de/de/A002068/Korrespondenz/A041001.html

Verwendungsbeispiel für TEI

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Vorlage der Digitalisierung



TEI-Annotation

```
<div>
<opener>
   <dateline>
      <hi rend="latintype">Nro</hi>: 15<lb/>d 1
      <hi rend="superscript">t</hi>
      <hi rend="latintype">Januar</hi>1817
   </dateline>
   <dateline rend="left">
      <hi rend="underline" n="1">Nachts 2 Uhr</hi>. oder
vielmehr <hi rend="underline" n="1"><hi rend="underline"
n="1">Früh</hi>2 Uhr</hi>.
   </dateline>
</opener>
<p n="1">
   <hi rend="underline" n="1">Prost Neujahr</hi>!!
Millionen Bußen. und<hi rend="underline" n="1">gute
Nacht</hi>gute gute Nacht. mein einzig vielgeliebter
   <rs type="person" key="A000213">Muks</rs>.
</p></div>
…
```

Was ist XML denn nun genau?

Deskriptives Markup:

<head>Dies ist eine Überschrift</head>

Was ist XML denn nun genau?

Komponenten eines XML-Dokuments:

- XML-Deklaration: <?xml version="1.0" encoding="UTF-8"?> (sog. Processing Instructions)
- Wurzelelement
- Elemente: <head>Dies ist eine Überschrift</head>
- Attribute:
  <head type="sub">Dies ist eine untergeordnete Überschrift</head>
- <!-- Kommentare -->
- Entityreferenzen: z.B.: "&" für "&"; „<" für "<" (diese beiden Zeichen sind XML selbst vorbehalten und müssen immer "escaped" werden)
- Namensräume/Namespaces

Was ist XML denn nun genau?

Es werden zwei Grade der Konformität mit dem XML-Standard unterschieden:

- Wohlgeformtheit
- Gültigkeit / Validität

Was ist XML denn nun genau?

- Wohlgeformtheit:
  - es gibt genau ein Wurzelelement, das das gesamte restliche Dokument enthält
  - alle Elemente sind ordentlich geschachtelt ("properly nested")
  - Element- und Attributnamen sind "case sensitive"
  - Zu jedem <start>-Tag gehört ein schließendes Tag: </start>, Ausnahme: <leere/> Tags
  - Attributwerte stehen in Anführungszeichen: <tag attributname="attributwert">

Was ist XML denn nun genau?

– Gültigkeit:

Ein XML-Dokument kann valide sein, d.h.
mit den in einem Schema definierten Regeln übereinstimmen

Gültigkeit / Validität

Ein valides/gültiges XML-Dokument stimmt mit Regeln überein, die in einem zugewiesenen Schema definiert werden.

Ein Schema definiert u.a.:

- die Bezeichnung des Wurzelelements
- die Bezeichnung aller weiteren Elemente
- die Bezeichnung und ggf. Standardwerte für die Attribute
- die Verschachtelung von Elementen

Beispiel einer XML-Datei

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>TEI-Minimal-Beispiel</title>
   </titleStmt>
   <publicationStmt>
    <p>Frei verfügbar</p>
   </publicationStmt>
   <sourceDesc>
    <p>Dieser Text ist digital born.</p>
   </sourceDesc>
  </fileDesc>
 </teiHeader>
 <text> <!-- Ein XML Kommentar -->
  <body>
   <p>Ein Beispieltext von <name>Agnes Brauer</name><lb/>
    für die Übung<hi rend="italic">Textauszeichnung mit TEI</hi>.</p>
  </body>
 </text>
</TEI>
```

Richtig oder falsch?

Übung 1 im Pad (in Breakout-Rooms):

Übung 1: Richtig oder falsch?

```
<p>Lorem ipsum</P>
```

- [ ] richtig, weil:
- [ ] richtig, weil:
- [ ] richtig, weil:
- [ ] falsch, weil:
- [ ] falsch, weil:
- [ ] falsch, weil:

```
<p><wort>Lorem ipsum</p></wort>
```

- [ ] richtig, weil:
- [ ] richtig, weil:
- [ ] richtig, weil:
- [ ] falsch, weil:
- [ ] falsch, weil:
- [ ] falsch, weil:

```
<wort>Lorem <p>ipsum</p></wort>
```

Richtig oder falsch?

Lorem ipsum</P>	Lorem ipsum
<wort>Lorem ipsum </wort>	<wort>Lorem</wort> ipsum
<wort>Lorem ipsum </wort>	wohlgeformt, ggf. nicht valide
Lorem ipsum	Lorem ipsum
	wohlgeformt
< p type="blindtext"/>Lorem ipsum	Lorem ipsum
Lorem ipsum <gap/>	wohlgeformt

Aufbau von TEI-Dokumenten

```
<TEI>

        <teiHeader>
        …
        </teiHeader>

        <text>
        …
        </text>
</TEI>
```

<teiHeader>

Der TEI Header beinhaltet Metainformationen über den annotierten Text:

fileDesc (file description)	enthält die vollständige biblio-graphische Beschreibung der Datei (und ggf. der zugrundeliegenden Vorlage)
encodingDesc (encoding description)	Richtlinien der Transition in die elektronische Form
profileDesc (text-profile description)	erlaubt die ausführliche Beschreibung nicht-bibliographischer Textmerkmale, z.B. der vorkommenden Sprachen, Personen und ihrer Situierung
revisionDesc (revision description)	verzeichnet Dateirevisionen

<text>

front (front matter)	optional; umfasst alle vorangestellten (Para-)Texte wie Titelei, Vorwort, Inhaltsverzeichnis u.Ä.
body	obligatorisch; enthält den Hauptteil des annotierten Textes (ohne Paratexte)
back (back matter)	optional; enthält Anhänge, die auf den Textkörper folgen
group	bestimmt für Textsammlungen wie z.B. Anthologien; jedes <group>-Element enthält wiederum ein komplettes <text>-Element

Textstruktur

div (text division)	eine Untereinheit (z. B. Kapitel, Abschnitt etc.) innerhalb von front, body oder back
head (heading)	eine Überschrift
p (paragraph)	ein Prosaabsatz

Milestones

Da Seiten-, Zeilen- und Spaltenwechsel nicht immer mit den Struktureinheiten des Textes übereinstimmen, werden sie i.d.R. mithilfe von sog. Milestones gekennzeichnet:

pb/ (page break)	bezeichnet einen Seitenwechsel
lb/ (line break)	markiert Zeilenwechsel
cb/ (column break)	kennzeichnet Spaltenwechsel

Das Kodierungsschema TEI

… besteht aus verschiedenen Modulen, die eine bestimmte Anzahl von XML Elementen und Attributen deklarieren

TEI-Module

Module name	Formal public identifier	Where defined
analysis	Analysis and Interpretation	17 Simple Analytic Mechanisms
certainty	Certainty and Uncertainty	21 Certainty, Precision, and Responsibility
core	Common Core	3 Elements Available in All TEI Documents
corpus	Metadata for Language Corpora	15 Language Corpora
dictionaries	Print Dictionaries	9 Dictionaries
drama	Performance Texts	7 Performance Texts
figures	Tables, Formulae, Figures	14 Tables, Formulæ, and Graphics
gaiji	Character and Glyph Documentation	5 Representation of Non-standard Characters and Glyphs
header	Common Metadata	2 The TEI Header
iso-fs	Feature Structures	18 Feature Structures
linking	Linking, Segmentation, and Alignment	16 Linking, Segmentation, and Alignment
msdescription	Manuscript Description	10 Manuscript Description
namesdates	Names, Dates, People, and Places	13 Names, Dates, People, and Places
nets	Graphs, Networks, and Trees	19 Graphs, Networks, and Trees
spoken	Transcribed Speech	8 Transcriptions of Speech
tagdocs	Documentation Elements	22 Documentation Elements
tei	TEI Infrastructure	1 The TEI Infrastructure
textcrit	Text Criticism	12 Critical Apparatus
textstructure	Default Text Structure	4 Default Text Structure
transcr	Transcription of Primary Sources	11 Representation of Primary Sources
verse	Verse	6 Verse

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ST.html#STMA

TEI-Schema

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Customizations provided by the TEI Consortium

Lite	TEI Lite, the most widely used TEI customization; includes basic elements for simple documents	ODD DTD RNG XSD HTML PDF
TEI Tite	A constrained customization designed for use by keyboarding vendors.	ODD DTD RNG XSD HTML PDF
Bare	TEI Absolutely Bare, a very barebones schema with the absolute minimum of elements	ODD DTD RNG XSD
All	TEI with all modules included	ODD DTD RNG XSD
Corpus	TEI for Linguistic Corpora, includes the modules for encoding linguistic corpora	ODD DTD RNG XSD
MS	TEI for Manuscript Description, includes the elements for describing manuscripts and complex physical aspects of documents	ODD DTD RNG XSD
Drama	TEI with Drama, includes the TEI drama module	ODD DTD RNG XSD
Speech	TEI for Speech Representation, includes the TEI module for spoken language	ODD DTD RNG XSD
Dictionaries	TEI for Dictionaries	ODD DTD RNG XSD

http://www.tei-c.org/Guidelines/Customization/

Links

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

TEI by Example: http://teibyexample.org/

Oxford Teaching Pages: http://tei.oucs.ox.ac.uk/Talks/

Ausblick

- Zur Verarbeitung, Auswertung und Transformation von XML-Daten stehen zahlreiche Technologien bereit

- Zentral sind dabei:

    - die Abfragesprache XPath, die der Suche und Navigation innerhalb von XML-Dokumenten dient
    - Beispiel (ETAHoffmannLite_core.xml): /TEI/teiHeader/fileDesc/sourceDesc/bibl/author

    - die Transformationssprache XSL(T), mit deren Hilfe man XML-Dokumente verändern oder in ein anderes Format (z.B. HTML) überführen kann

ToDos bis zur Hands-on Übung

- Tragen Sie bitte bei Bedarf / Interesse Themenvorschläge für die Hands-on Übung in das kollaborative Dokument ein
- Installieren Sie bitte den Oxygen XML-Editor (Version 23) auf Ihrem Rechner
- Eine Lizenz für **studentische** Teilnehmer*innen wird per E-Mail verschickt, Testzugang für einen Monat ist ebenfalls möglich.

<trailer>Vielen Dank für Ihre Aufmerksamkeit!</trailer>