

# Einführung in die Textauszeichnung mit TEI



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

- 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: `<i>Tim Berners-Lee</i>`

=> 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 = Extensible Markup Language

- vom [W3](#)-Konsortium empfohlener Standard für einen an Nachhaltigkeit und Interoperabilität orientierten Umgang mit digitalen Dokumenten
- Fundamentalere 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: **<i>**Tim Berners-Lee**</i>**

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



**CARL MARIA VON WEBER AN CAROLINE BRANDT IN PRAG  
BERLIN, MITTWOCH, 1. JANUAR BIS SAMSTAG, 4. JANUAR 1817  
(NR. 15)**

Home > Carl Maria von Weber > Korrespondenz > A041001 (bearbeitet) ← Zurück

An Mademoiselle  
[Caroline Brandt](#),  
Wohlgebohren  
Mitglied des Ständischen Theaters  
zu  
[Prag](#),  
Wohnhaft am Juden-  
TandelMarkt im Hause  
des Herrn PostOffizianten  
[Schwarz](#),

Nr: 15  
d 1<sup>er</sup> Januar 1817

Nachts 2 Uhr, oder vielmehr [Früh](#) 2 Uhr.  
Prost, [Neujahr](#)!! Millionen [Bußen](#), und gute Nacht gute Nacht, mein einig vielgeliebter [Maks](#).

Abends 1/2 11 Uhr.  
Ich wollte daß der Teufel alle überfüllten Freunde über 400000 Hektoliter weg mit ihren Nasenstücken, ungewaschenen,  
vorlauten einfältigen [Flaspermäulern](#) führte. Ich bin so böse ich könnte alles [frakaffern](#) vor [Wuth](#). No, nur erschrick  
nur nicht zu sehr, ich bin schon wieder gut, und beim Lichte besehen hat ja mein [Maks](#) dadurch einige Tage früher  
Freude gehabt, ¶ die man nie früh genug haben kann. Aßber so ist es, ich meine immer alle Freude für meine [Lina](#) dürfte  
nur von mir aus gehen.

Nun will ich aber ordentlich erzählen, sonst glaubst du ich bin toll geworden. Nachdem ich Gestern meinen [Brief No. 14](#)  
abgeschickt hatte, lauerte ich vergebens auf einen Brief von [Maks](#). Gleng um 6 Uhr auf die Akademie, von da zu [Krausens](#),  
und von da um 10 Uhr zu [Jordans](#) ¶, wo [Lichtensteins](#) und beinahe alle meine Bekannten waren. an beiden Orten mußte  
ich spielen, wie das wohl hier geht wie du weißt. Es war recht lustig, ich saß zwischen der [Wollank](#) und [Lork](#), und  
plauderte recht viel von dir, was mich wie immer sehr heiter und aufgeräumt machte. Punkt 12 Uhr tranken wir deine  
Gesundheit, dann wurde Prost [Neujahr](#) gejubelt, und und – und – dein [Maks](#) mußte viel [Bußen](#), aber du hättest nichts  
dagegen gehabt, und alle wünschten dich unzählige mal herbey. Müde matt und abgetragen kam ich um 2 Uhr nach  
Hause, und konnte nicht in Bett gehen ohne dir wenigstens 2 Worte zu schreiben, wie oben zu lesen. Heute habe viel  
Vörten geschnitten, Mittag bey [Hothos](#) gegessen, und dann in die [Armide](#) von [Gibik](#) gegangen die bis nach 10 Uhr dauerte.  
Wie ich nach Hause kam, so habe ich dasjenige haben Brief No. 1 und 2 der das erste Gatte heißt Isachard von [Friedrich](#), heißt.

**Dokument**

- Text
- Apparat
- Faksimile
- Rückverweise
- XML-Vorschau
- Download

**Kontext**

**Personen**

- Apitz, Herr
- Beer, Amalie
- Beer, Jacob Hertz
- Brandt, Christiane Sophia Henr...
- Clement, Franz
- Friedrich August L. König von...
- Gänsbacher, Johann
- Gerle, Wolfgang Adolph

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

[illegible]

```
<div>
<opener>
  <dateline>
    <hi rend="latintype">Nro</hi>: 15</b>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>
...
```

# FRANZ ANTON VON WEBER AN FRANZ ANTON HOFFMEISTER (HOFFMEISTER & KÜHNEL) IN LEIPZIG CARLSRUHE IN OBERSCHLESISIEN, MONTAG, 26. MAI 1806

Home > Franz Anton von Weber > Korrespondenz > A040167 (bearbeitet)

Zurück

Karlsruhe in Oberschlesien

d 26<sup>te</sup> May 1806.

Geliebtester alter [Herzensfreund!](#) und Or[dens]br[uder]!!!

Ich hoffe noch Anspruch machen zu dürfen auf unsre alte vieljährige Bekanntschaft und Freundschaft, und wenn Sie mich also noch dafür anerkennen, wie ich hoffe, so zeigen dieselben mir es dadurch, das sie mir, wo möglich, mit umgehendem Postwagen, die *Partitur* des [Vater unsers](#) vom seel. Kapelmeister [Naumann](#), versteht sich gegen Baar Zahlung anhero zu schiken belieben wolten, es solte zum Beschluß *Benefice* meines [Karls](#) als *Opern Director* in [Bressl](#) seyn.

Kann ich Ihnen in [Bressl](#) oder sonsten in hiesigen Gegenden Gefälligkeit erzeigen, so befehlen Sie Geliebter Br: demjenigen, so sich Achtungsvoll nennt.

Dero

g[an]z ergeb[en]ster D[iene]r und o[r]dens[br]uder

[Freyherr von Weber.](#)

Kammerherr

## Einstellungen

Zeige Markierungen im Text

- ☒ Lesarten Δ
- ☒ Textkonstitution +
- ☒ Herausgeberergänzungen []
- ☒ Einzelstellenerläuterung \*
- ☒ Themenkommentare T

## Dokument

Text

Apparat

Faksimile

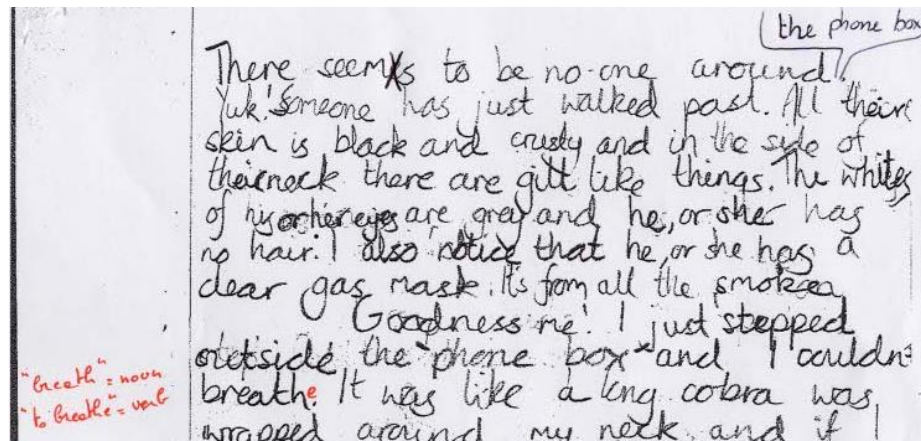
XML-Vorschau

Download

## Kontext

Die in der Beilage Nr. 72

Alle hervorheben Groß-/Kleinschreibung Akzente Ganze Wörter 1 von 1 Übereinstimmung



```
<p>There seem<del rend="crossout" hand="#HR">e</del>s to be no-one around <add place="above" hand="#HR">the phone box</add>. Yuk! <subst hand="#HR" resp="#TBE">
<del rend="writeover">s</del>
<add>S</add>
</subst>omeone has just walked past. All their skin is black and crusty and in the side of their neck are <unclear reason="illegible" resp="#TBE">gitt</unclear> like
things. The whites of his or her eyes are grey and he, or she has no hair. I also notice that he, or she has a clear gas mask. It's from all the smoke.</p>
<p>Goodness me! I just stepped outside the phone box and I couldn't<unclear reason="cutoff while xeroxing" resp="#TBE">t</unclear> breath<add hand="#teacher">e</add>.
<add hand="#teacher" place="margin">"breath" = noun <lb/> "to breathe" = verb</add> It was like a <choice>
<sic>kng</sic>
<corr>king</corr>
</choice> cobra was wrapped around my neck and if I was to get a little bit of air in it was horrible, like when you stop at red man beside the traffic lights and a
```

Abb.: <https://teibyxample.org/tutorials/TBEDo6v00.htm>



### Concepts



Abb.: [https://wvp.northeastern.edu/outreach/seminars/\\_utils/gfx/concept\\_syntax\\_language.png](https://wvp.northeastern.edu/outreach/seminars/_utils/gfx/concept_syntax_language.png)

## XML

### Syntax

```

<element>
  <element attribute="value">
    content
  </element>
</element>

```

## TEI

### Language: vocabulary and grammar

```

<p>
<note type="foot">
<head>

```

Textauszeichnung /  
Annotation macht  
implizites Wissen  
über den Text  
durch Metadaten  
explizit!

# Was ist XML denn nun genau?

Deskriptives Markup:

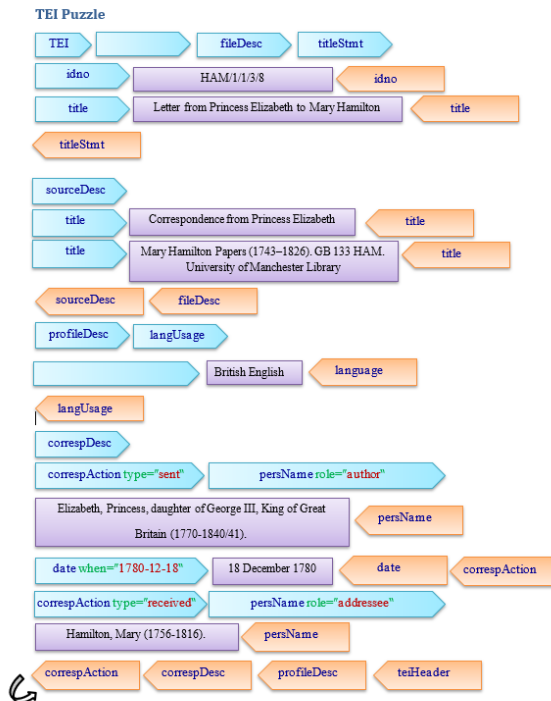
```
<head>Dies ist eine Überschrift</head>
```

# Was ist XML denn nun genau?

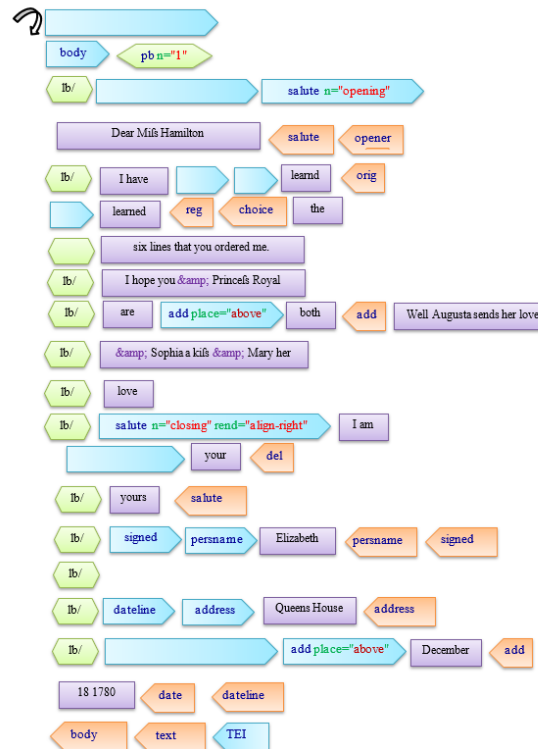
Komponenten eines XML-Dokuments:

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

# Puzzle time!



<https://carpentries-incubator.github.io/tei-xml/aio/index.html>



Gruppe 1:  
[https://miro.com/app/board/uXjVKScd46o=?share\\_link\\_id=617946263309](https://miro.com/app/board/uXjVKScd46o=?share_link_id=617946263309)

Gruppe 2:  
[https://miro.com/app/board/uXjVKRj9R74=?share\\_link\\_id=366563418761](https://miro.com/app/board/uXjVKRj9R74=?share_link_id=366563418761)

Gruppe 3:  
[https://miro.com/app/board/uXjVKRj4UYU=?share\\_link\\_id=85322661196](https://miro.com/app/board/uXjVKRj4UYU=?share_link_id=85322661196)  
Z

# 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 Wurzelement, 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 Wurzelements
- 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 2 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>
```

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

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

# TEI-Schema

## Customizations provided by the TEI Consortium

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

<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/>

# 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](#) auf Ihrem Rechner, Testzugang für einen Monat ist möglich!
- Studierende der GU können für die Version 26 eine Lizenz bis November 2025 erhalten. Bei Interesse bitte eine Mail an mich. Absender\*in muss eine studentische E-Mail-Adresse der GU sein.