Einführung in die TEI

Text Encoding Initiative

Christiane Fritze

Leopoldina
Nationale Akademie
der Wissenschaften

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

AKADEMIE
WISSENSCHAFTEN UND KÜNSTE
NORDRHEIN-WESTFALEN

i|d|e

Wiederholung: warum Texte kodieren?

- um maschinenlesbar (explizit) zu machen, was vom Leser implizit verstanden wird
- um Texte mit Informationen anzureichern
- um Texte nachzunutzen
- in unterschiedlichen Formaten
- in unterschiedlichen Kontexten
- von unterschiedlichen Nutzern

Übersicht

- Was ist die TEI?
  - Ein Konsortium.
  - Ein Kodierungsstandard.
- Was bietet die TEI?
  - Guidelines.
  - Werkzeuge.
  - Und Mehr.
- Wie gehe ich mit der TEI um?
  - Grundaufbau des TEI Datenmodells.
  - Guidelines benutzen.

http://tei-c.org

Die Gründung der Text Encoding Initiative

- Frühjahr 1987: European workshop on standardisation of historical data (J.P. Genet, M. Thaller)
- Herbst 1987: NEH funds an exploratory international workshop on the feasability of defining „text encoding guidelines“
- Juni 1990: Release of te hfirst draft (known as „P1“) of the Guidelines

http://www.tei-c.org

Vassar College, Poughkeepsie, Herbst 1987

- “The TEI was established in 1987 to develop, maintain, and promulgate hardware- and software-independent methods for encoding humanities data in electronic form.”

- “Over nearly three decades the TEI has been extraordinarily successful at achieving its objective and it is now widely used by scholarly projects and libraries around the world.”

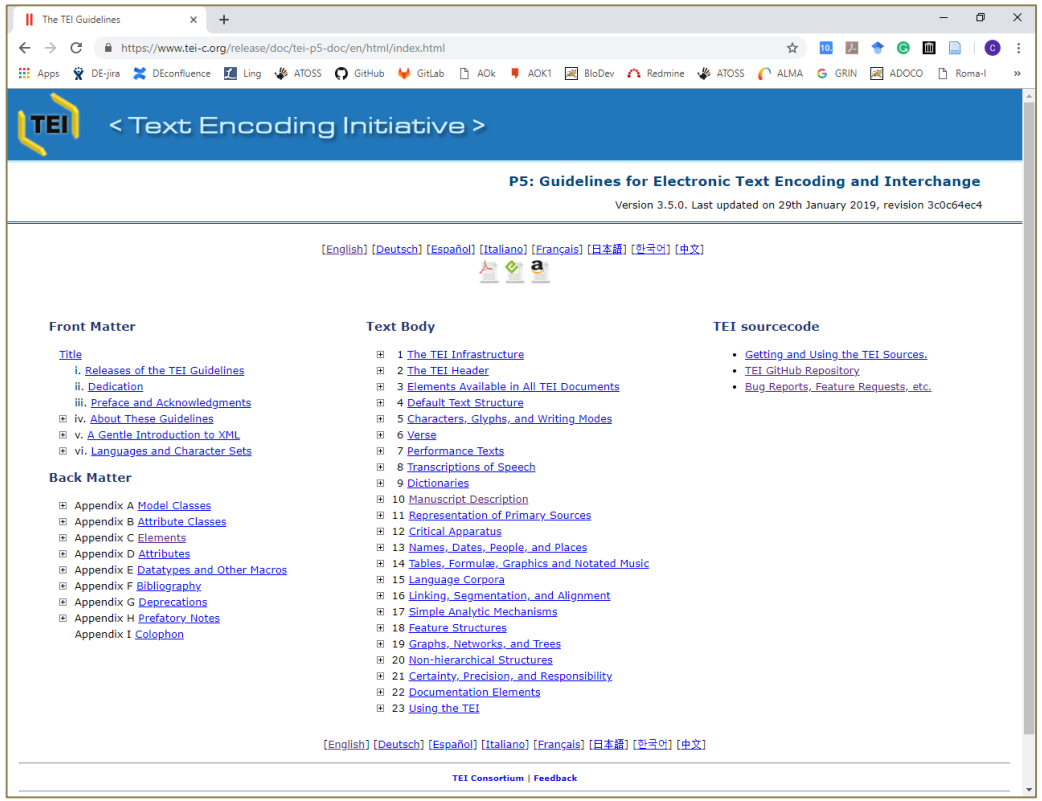
- “TEI is the de facto standard for text encoding.”

TEI Angebote - Community

- Special Interest Groups
- Annual members‘ meetings
  - September 2019 in Graz (Österreich)
- TEI-L Mailingliste
  - tei-l@listserv.brown.edu
  - Fragen von allen für alle auf allen Niveaus
  - Mail-Archiv:
    - http://listserv.brown.edu/archives/cgi-bin/wa?A0=TEI-L
    - http://markmail.org/list/edu.brown.listserv.tei-l
- Zeitschrift: jTEI
  - http://journal.tei-c.org/journal/index

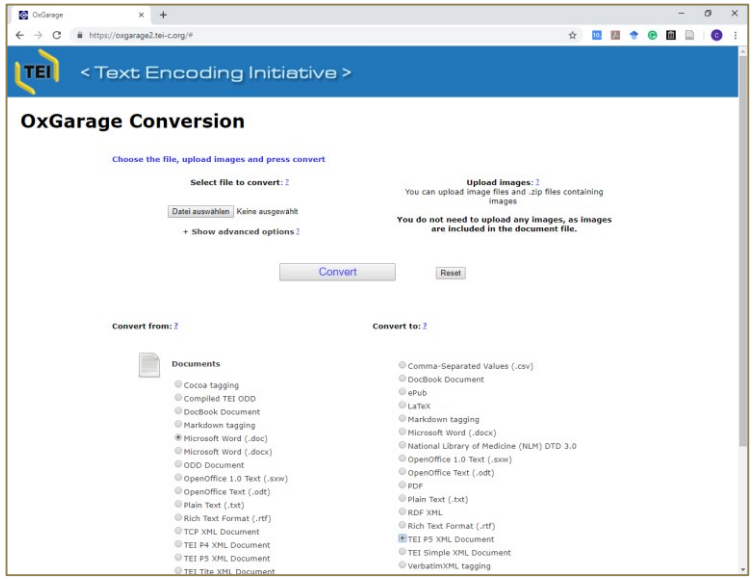
TEI Angebote - Guidelines

- **TEI P5 Guidelines for Electronic Text Encoding and Interchange**
  - Version 3.5.0.
  - Last updated on 29th January 2019, revision 3c0c64ec4
  - derzeit 1926 Seiten
- Online https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html
- EPUB und MOBI Format
- PDP-Format
- Build-In im oXygen Editor

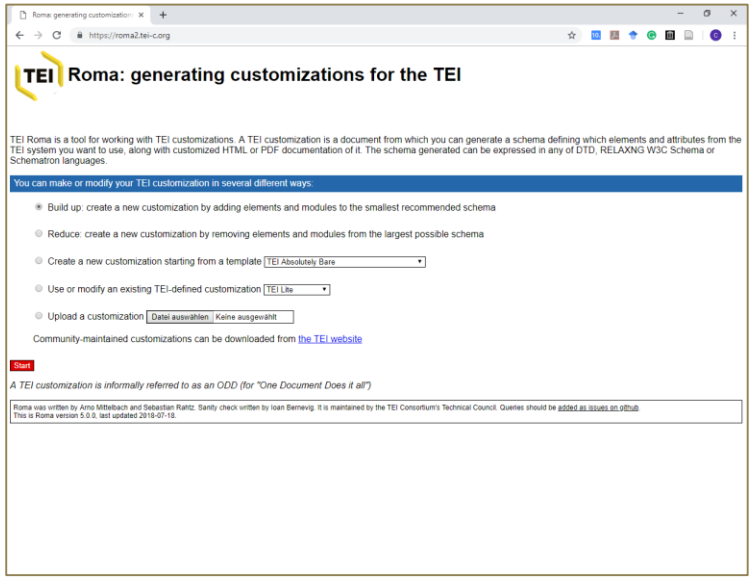


TEI Angebote – Tools

- **OxGarage**: Konversion von und nach TEI
- **ROMA**: Erstellen von Schemata und Dokumentationen
- **Stylesheets**: Konversion von TEI-Dokumenten in diverse Formatvorlagen



https://oxgarage.tei-c.org/

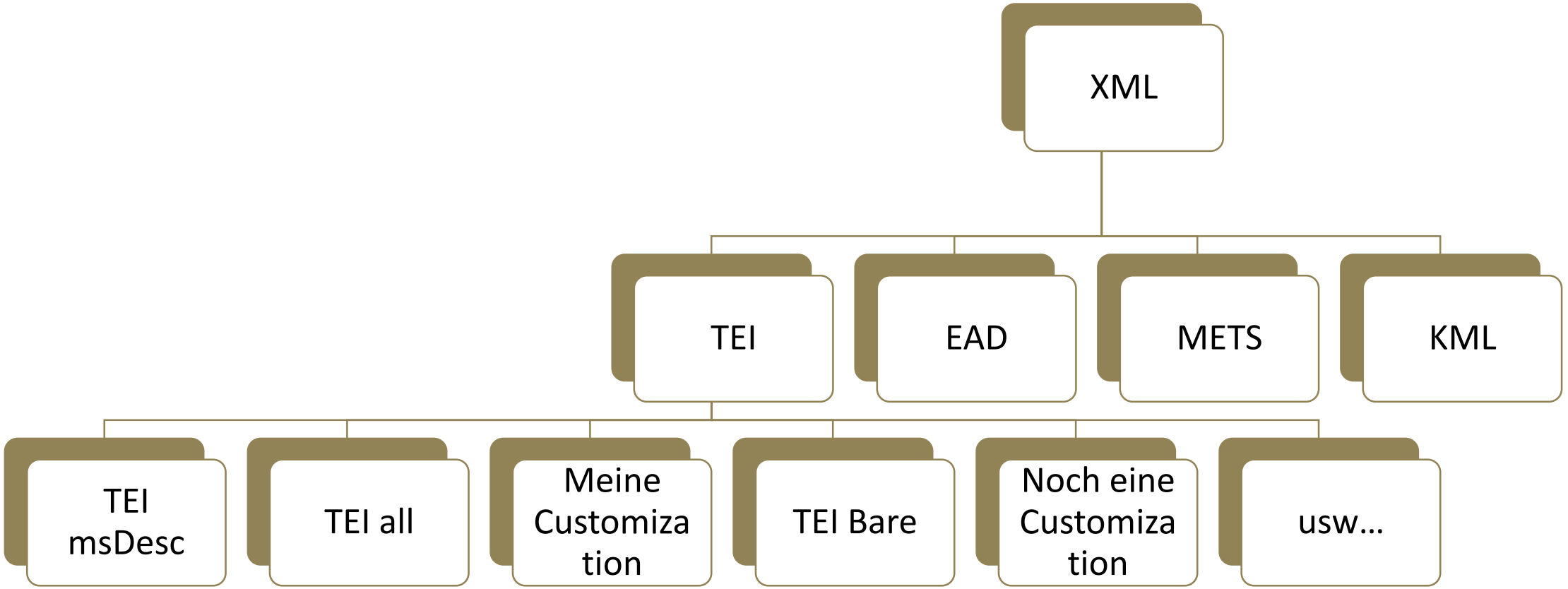


https://roma.tei-c.org/

XML ⟺ TEI?

- XML: stellt nur die grundlegenden Regeln bereit:
  - Baumstruktur (ein Wurzelelement, korrekte Schachtelung)
  - Konventionen für die Kodierung von
    - Elementen `<element/>`
    - Attributen `<element attribut="wert"/>`
    - Entity-Referenzen `&entity;` z.B. `&lt;`
    - Kommentaren `<!-- … -->` usw.
  - Benennungsregeln für Elemente und Attribute

TEI ist XML


XML
TEI
EAD
METS
KML
TEI msDesc
TEI all
Meine Customiza tion
TEI Bare
Noch eine Customiza tion
usw…

TEI Guidelines

- … sind eine Einschränkung der im Prinzip unendlichen Möglichkeiten von XML.
- … bieten formalisierte Schemata zur Validierung von XML-Dateien.

- Fragen, die geklärt werden müssen:
  - Welche Tags für Elemente und Attribute werden bereitgestellt?
  - Wie dürfen die Tags verschachtelt werden?
  - Wie kann ich die gegebenen Möglichkeiten um eigene Regeln erweitern?

TEI Customizations

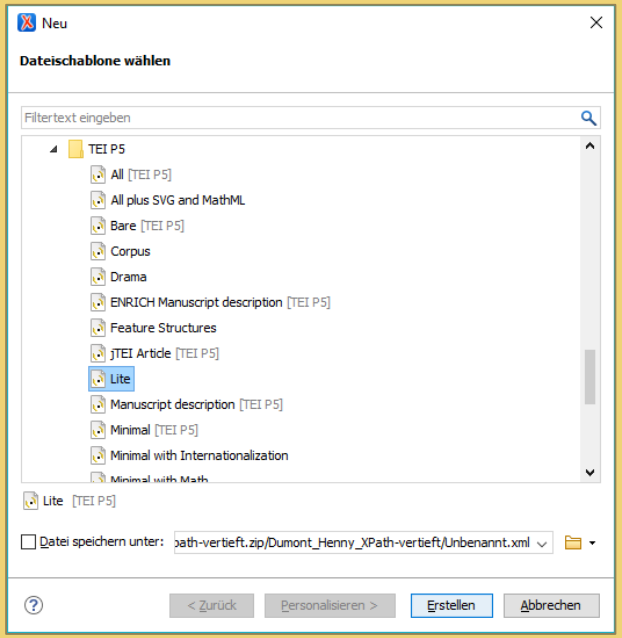
- TEI-Standard stellt mehrere hundert Elemente und Attribute bereit
- In den seltensten Fällen werden alle benötigt
- Modularer Aufbau der TEI erlaubt Definition von Untermengen des TEI-Tagsets/ Teilgruppen von TEI-Elementen zu definieren bzw. auszuwählen

TEI Customizations

- Diese Schemata können mit Hilfe des ROMA-Tools erstellt und modifiziert werden: http://www.tei-c.org/Roma/
- oXygen XML editor kann Dokumente mit vordefinierten Schemata verbinden, z.B.:
  - TEI All (umfasst alle Elements, maximales Schema)
  - TEI Bare (umfasst nur absolut notwendige Elemente)
  - TEI Lite (umfasst alle wichtigen und am häufigsten gebrauchten Elemente)
  - TEI Minimal (noch weniger Elemente?)

Übung

- Öffnen Sie im Oxygen-XML-Editor ein neues XML-TEI P5-Dokument (TEI Minimal):
  - [STRG + N] → Framework-Vorlagen → TEI P5 → Minimal
- Ersetzen Sie die Platzhalter (Title, Publication information, Information about the source, Some text here) mit passenden Inhalten.



TEI ist XML

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_minimal.
3  <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_minimal.
4     schematypens="http://purl.oclc.org/dsdl/schematron"?>
5  <TEI xmlns="http://www.tei-c.org/ns/1.0">
6    <teiHeader>
7        <fileDesc>
8           <titleStmt>
9              <title>Title</title>
10          </titleStmt>
11          <publicationStmt>
12             <p>Information about publication or distribution</p>
13          </publicationStmt>
14          <sourceDesc>
15             <p>Information about the source</p>
16          </sourceDesc>
17       </fileDesc>
18   </teiHeader>
19   <text>
20       <body>
21          <p>Some text here.</p>
22       </body>
23   </text>
24 </TEI>
```

Repräsentation eines Dokuments

- Eine TEI-Datei repräsentiert ein „real world object“, durch
  - Metadaten (<teiHeader>, u.a. <msDesc>)
  - digitale Abbilder (<facsimile>)
  - Transkription/„Edition“ (<text>)

TEI Grundstruktur

- Wurzelelement **<TEI>**
- Enthält mindestens zwei Unterelemente, nämlich
  - **<teiHeader>**

    für Metadaten und **muss** immer vorhanden sein und
  - **<text>** für Texte aller Art und/oder
  - <sourceDoc> für genetische Edition / topographische Transkriptionen und /oder
  - <facsimile> für Verknüpfungen und/oder
  - <fsDecl> v.a. für Textanalysen, linguistische Merkmale u.ä.

TEI Grundstruktur - Sonderfall teiCorpus

- Sonderfall **`<teiCorpus>`** besteht aus
  - `/teiCorpus/teiHeader`
  - 1..n `/teiCorpus/TEI`
  - Vorteil: Trennung von Metadaten, die sich auf das Gesamtkorpus beziehen („Goethes Briefe“), und Metadaten, die sich auf die Teile beziehen („Brief an Eckermann v. 14.8.1830“)
  - Geeignet z.B. für:
    - sprachwissenschaftliche Korpora
    - Sammeleditionen aus mehreren Quellen
    - Briefeditionen
    - Nachlässe

```
5  <teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
6      <teiHeader> [13 lines]
20     <TEI xml:id="MyTextNumber1">
21         <teiHeader> [13 lines]
35         <text> [6 lines]
42     </TEI>
43     <TEI xml:id="MyTextNumber2">
44         <teiHeader> [13 lines]
58         <text>
59             <body> [4 lines]
64         </text>
65     </TEI>
66
67 </teiCorpus>
```

<text> und Sonderfall <group>

- Das <text>-Element enthält den eigentlichen Text
  - Enthält i.d.R. ein <body>-Element
  - dazu fakultativ <front> und/oder <back>
  - oder <group>
- Sonderfall <group>: enthält 1..n <text>-Elemente
- Unterschied zw. <teiCorpus> und <group>: bei <teiCorpus> hat jeder Text einen eigenen Header, bei <group> nicht
- Die Entscheidung zw. <teiCorpus> und <group> hängt v. Editionsaufbau ab

TEI-Header – Mittwoch

- **<teiHeader>** kann aus vier Hauptteilen bestehen:
  - **<fileDesc>** (file description)
    Bibliographische Beschreibung des TEIDokuments
  - <encodingDesc> (encoding description)
    Beschreibung der editorischen Praxis
  - <profileDesc> (profile description)
    Kodierung inhaltlicher Informationen über den Text, v.a. bei Sprachcorpora
  - <revisionDesc> (revision description)
    Informationen über die Änderungsgeschichte des TEI-Dokuments
- N.B.: **Only <fileDesc> is mantatory.**

TEI Grundgerüst

```
<TEI>
  <teiHeader>
      <!--...-->
  </teiHeader>
  <facsimile>
      <!-- Reihe von <graphic> oder <surface> Elementen --
  >
  </facsimile>
  <text>
      <pb facs="page1.png"/>
      <!-- text contained on page 1 is encoded here -->
      <pb facs="page2.png"/>
      <!-- text contained on page 2 is encoded here -->
  </text>
</TEI>
```

<facsimile>

```
<facsimile>
  <graphic url="page1.png"/>
  <graphic url="page2.png"/>
</facsimile>
```

Gruppierung ist möglich

```
<facsimile>
  <graphic url="page1.png"/>
  <surface>
    <graphic url="page2-
  highRes.png"/>
    <graphic url="page2-
  lowRes.png"/>
  </surface>
</facsimile>
```

Ausweis von Zonen auf der Seite ist möglich

```
<facsimile>
  <surface
    ulx="0"    uly="0"
    lrx="200" lry="300">
    <graphic url="page1.png"/>
    <zone
      ulx="25"  uly="25"
      lrx="180" lry="60">
      <desc>Titel</desc>
    </zone>
  </surface>
</facsimile>
```

<text>

- Das <text> Element enthält den eigentlichen Text
- Struktur des Textes: Front Matter, Body, Back Matter

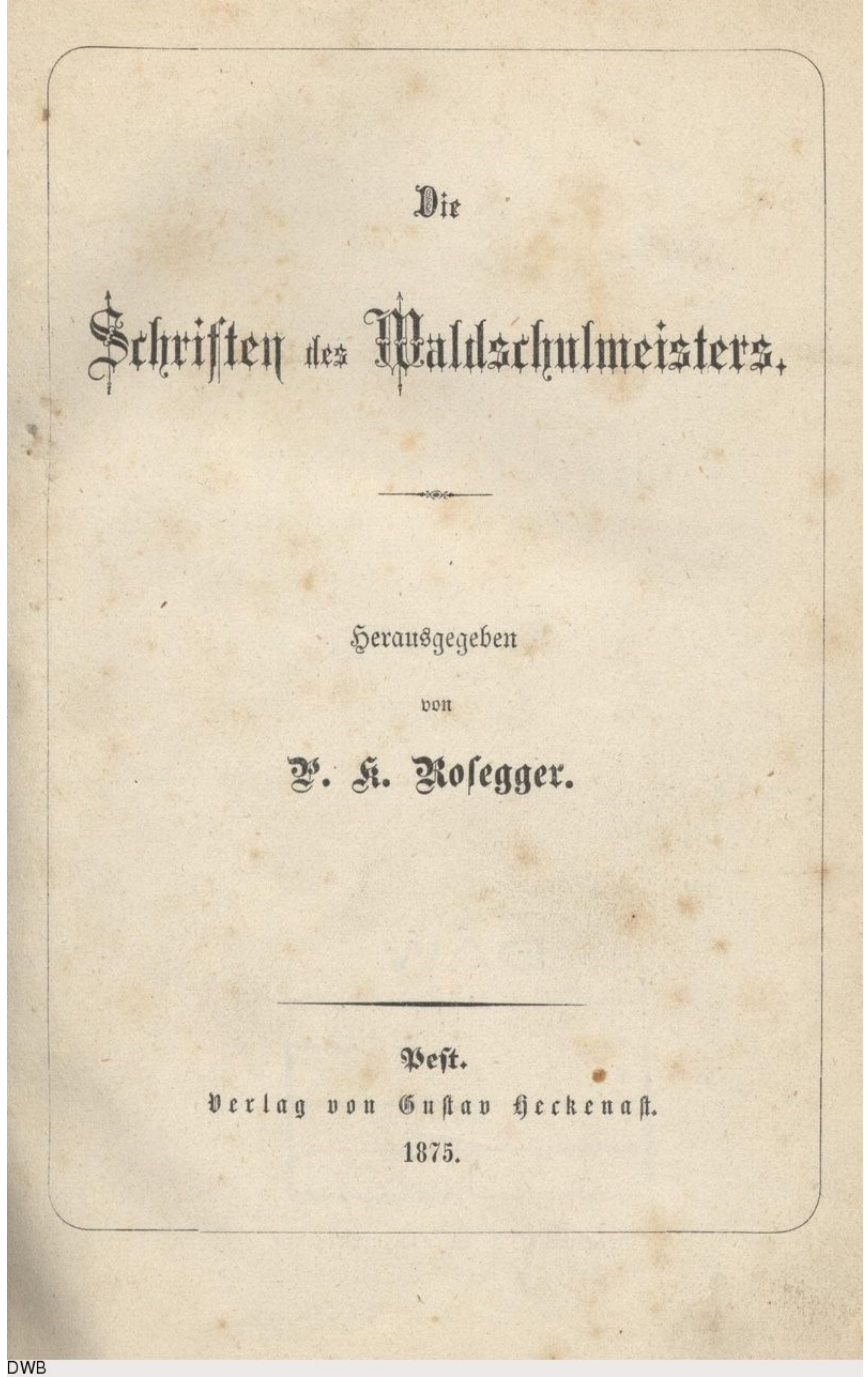
```
<text>
    <pb facs="page1.png"/>
    <front> </front>
    <body> </body>
    <back> </back>
</text>
```

<text> Front matter

- Front Matter
  - Titelseite, Vorwort/Prolog, Widmung, …
  - Funktionen: Identifikation, Einleitung, Aufforderung, Dank …
  - NICHT der TEI-Header!
    - Hier wird das vorliegende Dokument transkribiert
    - Im Header dagegen werden (zusätzliche) Metadaten zur TEI-Datei notiert

- Elemente:

```
<titlePage>
<div type="preface | ack | dedication | abstract
| contents | frontispiece | incipit | prayer"/>
<docTitle> <titlePart>
```


Die

Schriften des Waldschulmeisters.

Herausgegeben

von

P. K. Rosegger.

Pest.

Verlag von Gustav Heckenast.

1875.

DWB

DTA: urn:nbn:de:kobv:b4-200905195026

```
<TEI>
  <text>
    <front>
      <pb facs="#f0007"/>
      <titlePage type="main">
        <docTitle>
          <titlePart type="main"> <hi
rendition="#b">Die<lb/>
Schriften des Waldschulmeisters.</hi>
</titlePart>
        </docTitle><lb/>
        <milestone rendition="#hr"
unit="section"/>
        <byline>Herausgegeben<lb/>
von<lb/><docAuthor><hi rendition="#b">P. K.
Ro&#x017F;egger.</hi></docAuthor></byline><lb/>
        <milestone rendition="#hr"
unit="section"/>
        <docImprint>
          <pubPlace> <hi
rendition="#b">Pe&#x017F;t.</hi> </pubPlace><lb/>
          <publisher> <hi rendition="#g">Verlag
von Gu&#x017F;tav Heckena&#x017F;t.</hi>
</publisher><lb/>
          <docDate>1875.</docDate>
        </docImprint>
      </titlePage><lb/>
    </front>
  </text>
</TEI>
```

<back>

- enthält Anhänge jeglicher Art, die auf den Hauptteil eines Textes folgen

```
<back>
    <div type="appendix">
        <head>Bibliography</head>
        <listBibl>
            <bibl>Bieler, Ludwig, The Works of St. Patrick,
                in Ancient Christian Writers (Westminster, Maryland/ London, 1953).</bibl>
            <bibl>Bieler, Ludwig, "Irish Manuscripts in Medieval Germania",
                Irish Ecclesiastical Record, 87, 5th ser. (1957), 161-169.</bibl>
            <bibl> ... </bibl>
        </listBibl>
    </div>
</back>
```

<body>

- enthält den gesamten, eigenständigen Text, außer Vorspann (front) und Nachspann (back)

```
<text>
    <body>
        <p>
            <!-- Text goes here -->
        </p>
    </body>
</text>
```


TEI-Guidelines: 4 Default Text Structure

- „division“
- Neutrales Element für Textabschnitte
- Kann beliebig tief geschacht
- Kann mit @type und @subty
- @n für Nummerierung
- @xml:id für Identifikation

```
<body>
    <div type="Kapitel" n="1">
        <div type="Unterkapitel" n="1">
            <!-- Hier kommt der Text des ersten Unterkapitels im ersten Kapitel -->
        </div>
        <div type="Unterkapitel" n="2">
            <!-- Hier kommt der Text des zweiten Unterkapitels im ersten Kapitel -->
        </div>
    </div>
    <div type="Kapitel" n="2">
        <div n="1" type="Unterkapitel">
            <!-- Hier kommt der Text des ersten Unterkapitels im zweiten Kapitel -->
        </div>
        <div n="2" type="Unterkapitel">
            <!-- Hier kommt der Text des zweiten Unterkapitels im zweiten Kapitel -->
        </div>
    </div>
</body>
```

<ab> <floatingText>

- **<head> …</head>** Überschrift
- ** … ** (paragraph) Absatz
- **<ab>…</ab>** (anonymous block) irgendein Textblock
  - Textblock, ohne die Semantik von „Absatz“
- **<floatingText>…</floatingText>** schwebender Text
  - Ein bestimmter Text (jeglicher Sorte), der einen anderen Text unterbricht, der ihn umgibt
  - Kann selbst ein Einzeltext oder eine Gruppe von Texten sein
  - Kann überall vorkommen (auf allen Ebenen, beliebig tief)
  - In <floatingText> kann wieder eine komplexe Textstruktur codiert werden

Strukturelle Meilensteine

- `<pb/>` (page break) Seitenumbruch, Seitenumbruch, normalerweise der der Vorlage
- `<lb/>` (line break|beginning) Zeilenumbruch (z.B. wenn Zeilenumbrüche der Vorlage mit transkribiert werden)
  - Diskussion: vor oder nach der Zeile?
- `<cb/>` (column break) Spaltenumbruch
- `<anchor/>` Ankerpunkt (attaches an identifier to a point within a text, whether or not it corresponds with a textual element.)

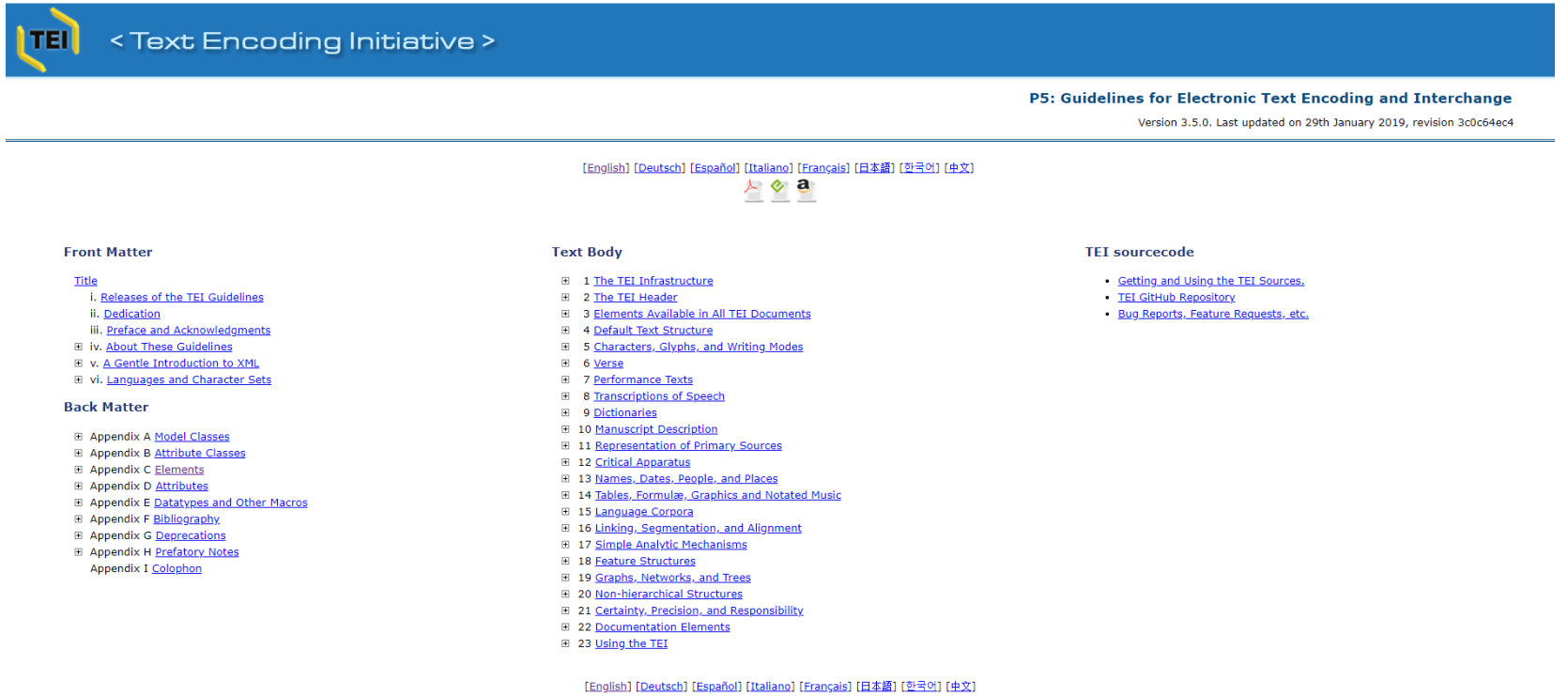
- … mit `@n`, `@type`, `@xml:id` spezifizierbar

Hervorhebung und wörtliche Rede

- `<hi>` (highlighted)
  - allgemeiner Tag für Hervorhebungen
  - z.B. für *kursiv*, **fett**, g e s p e r r t etc.
  - spezifizierbar durch `@rend`
- `<foreign>`, `<emph>`, `<distinct>`
  - verschiedene Hervorhebungsarten bzw. Markierung ‚ungewöhnlicher' Textteile (Fremdsprachiges, Betonung, Archaismen etc.)
- `<q>` (quoted)
  - für wörtliche Rede, Fachausdrücke, „sog." (anstelle von Anführungsstrichen)
- `<quote>` (quotation) und `<cit>` (citation) für Zitate

Mehr Elemente

- Strukturelle Texteinheiten
  - `<div>`, `<p>`, `<list>`, `<table>`<lg>, `<line>`, `<index>`
- Semantische Texteinheiten
  - `<head>`, `<fw>`, `<note>`, `<quote>`, `<term>`
  - `<ref>`, `<bibl>`, `<rs>`
- Entitäten
  - `<persName>`, `<orgName>`, `<placeName>`, `<name>`
  - `<date>`, `<geo>`



TEI Guidelines lesen

https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

TEI Klassen und Datentypen

- Module → inhaltlich bzw. formal zusammengestellt
- Elemente → nach semantischen Modellen gruppiert
  - Modellklassen: model.biblLike, model.choicePart, model.quoteLike
- Attribute → nach Inhaltsmodell gruppiert
  - Attributklassen: att.global, att.datable.w3c
  - Datentypen: z.B. data.pointer, data.word
- Module z.B.
  - **core** (Basiselemente)
  - **header** (Metadaten)
  - **textstructure** (grundlegende Textstrukturen)
  - msdescription (Handschriftenbeschreibungen)
  - gaiji (Sonderzeichen)

Globale Attribute

- @xml:id (eindeutiger Identifikator, muss dokumentenweit eindeutig sein, und mit einem Buchstaben beginnen, i.d.R. selbst vergeben oder automatisch erzeugt)
- @xml:lang (Sprache des Inhalts eines Elements)
- @n (Nummerierung, entweder aus Quelle übernommen oder selbst erstellt)
- @rend (Aussehen einer Textstelle in der Quelle!)

TEI Dokumentation lesen – Chapter „C Elemente“

- (kurze) Charakterisierung
- Welche Attribute sind im Element erlaubt?
- Innerhalb welcher Elemente darf das Element verwendet werden?
- Welche Kinderelemente darf das Element haben?
- Technische Beschreibung des Elements
- Klammerung
- Reihenfolgen → 'Komma', |
- Häufigkeiten → 'nichts', +, ?, *
- Beispiele
  → show all

TEI – Hilfe?!

- TEI P5 Guidelines: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html
- IDE: Folien, Mailinglist, Mitglieder
- TEI by Example (tutorial): http://www.tbe.kantl.be
- Andere TEI-Schools z.B. “TEI@Oxford”: http://tei.oucs.ox.ac.uk/Talks/ und http://tei.oucs.ox.ac.uk/Talks/2014-01-toronto/
- TEI-L@listserv.brown.edu
- http://listserv.brown.edu/archives/cgi-bin/wa?A0=TEI-L
- http://markmail.org/list/edu.brown.listserv.tei-l

Alles wird gut