XML und TEI

oder Die drei magischen Buchstaben

Überblick

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

Text Encoding Initiative – TEI

- Gründung
  - Frühjahr 1987: European workshops on standardisation of historical data (J.P. Genet, M Thaller)
  - Herbst 1987: NEH funds an exploratory international workshop on the feasibility of defining "text encoding guidelines"

Vassar College, Poughkeepsie

TEI

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

- „ is a consortium which collectively develops and maintains a standard for the representation of texts in digital form”
- Synonym für Quasi-Standard für die Kodierung von Texten

TEI Angebote

- TEI Guidelines zur Kodierung von Texten
  - Aktuelle Version **TEI P5**
    - Version 2.0.2 of TEI P5 published on 2012-02-02
  - 6-monatiger Release Zyklus
  - Dokumentation der ältere Versionen sind noch erreichbar
- Werkzeuge
  - zur Erstellung eines eigenen TEI-Schemas: **Roma** / ODD
  - Stylesheets zur Transformation der TEI-Dokumente (Freitag)
- Aktivitäten
  - Special Interest Groups, TEI Members ‘ Meeting
  - Mailingliste http://listserv.brown.edu/archives/cgi-bin/wa?A0=tei-l

XML ⟺ TEI?

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

TEI Guidelines

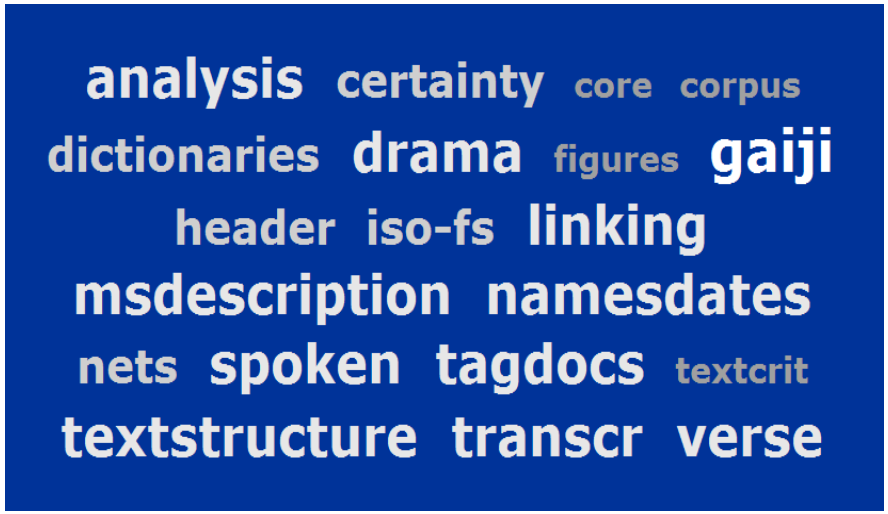
- ist eine Einschränkung der im Prinzip unendlichen Möglichkeiten von XML
- formalisierte Schemata zur Validierung von XML-Dateien
- Fragen, die geklärt werden müssen:
  - Welche Tags und Attribute werden bereitgestellt?
  - Wie dürfen die Tags verschachtelt werden?
  - Wie kann ich die gegebenen Möglichkeiten um eigene Regeln erweitern?

TEI Customizations

- TEI-Standard stellt mehrere hundert Elemente (Tags) und Attribute bereit, z.B. für
  - ‚Normale' Textkodierung
  - Textkritische Editionen
  - Linguistische Corpora
  - Bibliographische Beschreibung von Handschriften
  - Verknüpfung von Texten mit digitalen Bildern
- In den seltensten Fällen alle benötigt

TEI Customizations

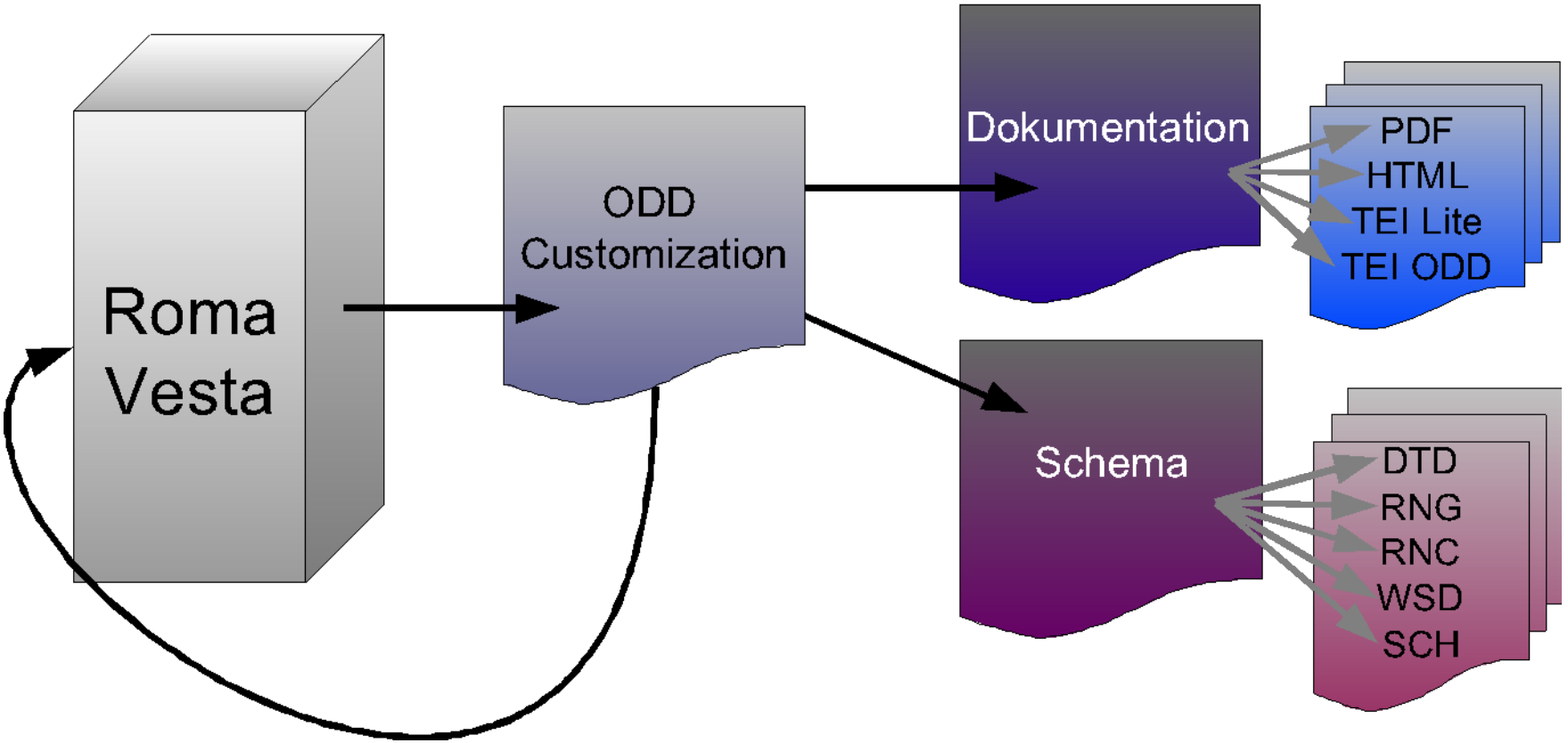
- Modularer Aufbau der TEI erlaubt Definition von Untermengen des TEI-Tagsets
- d.h. mein Schema muss nicht alle Elemente und Attribute der TEI enthalten (customisations)



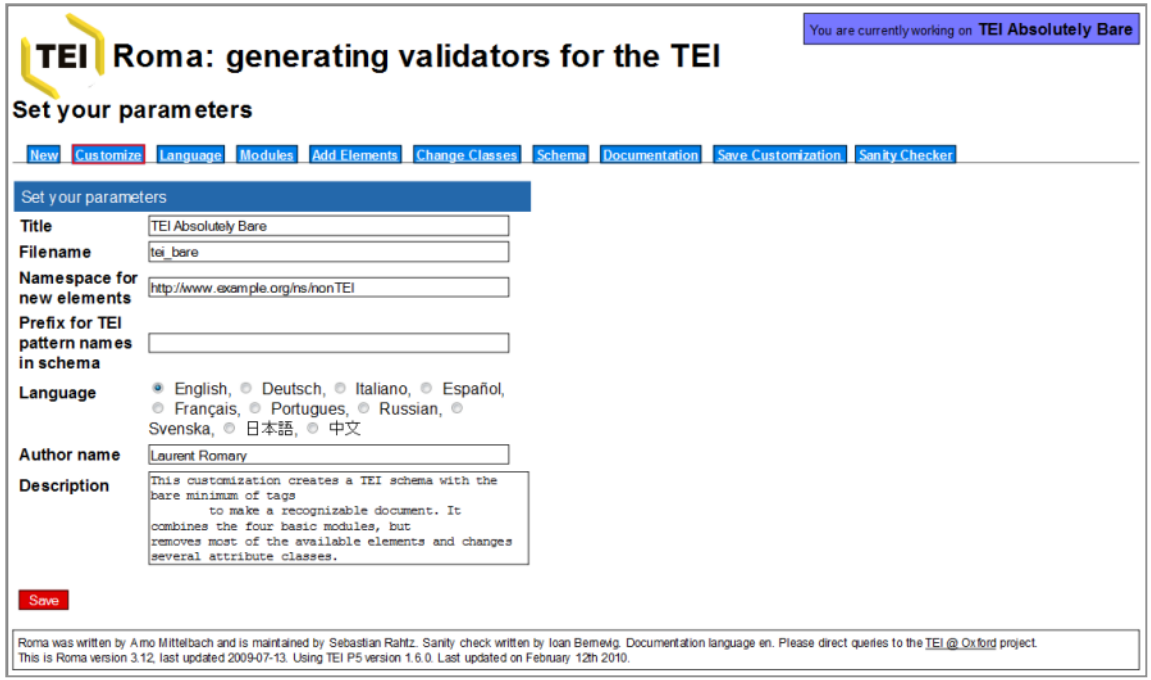
TEI Customizations

- Möglichkeit, angepasste Schemata zu generieren mit dem Tool Roma (http://www.tei-c.org/Roma/)
- oXygen ermöglicht Einbindung einiger vorgefertigter Schemata, u.a.
  - TEI All (alle Elemente, Maximalschema)
  - TEI Bare (nur das allernotwendigste)
  - TEI Lite (die wichtigsten Elemente)

Roma – ODD – Schema



- New
- Customize
- Language
- Modules
- Add Elements
- Change Classes
- Schema
- Documentation
- Save Customization
- Sanity Checker



Exkurs ODD: Spezifikation

- ODD bedeutet „**O**ne **d**ocument **d**oes it all.“
- ODD ist XML.
- ODD ist TEI.
  - ist für TEI-Datenmodellbeschreibung spezifisches Vokabular
  - TEI Guidelines Kapitel 22 „Documentation Elements“
  - Modul **tagdocs**

Exkurs ODD: So siehts aus

```
<?xml version="1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>My TEI Extension</title>
            <author>generated by Roma 3.12</author>
         </titleStmt>
         <publicationStmt>
            <p>for use by whoever wants it</p>
         </publicationStmt>
         <notesStmt>
            <note type="ns">http://www.example.org/ns/nonTEI</note>
         </notesStmt>
         <sourceDesc>
            <p>created on Sunday 28th February 2010 08:53:41 AM</p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <front>
         <divGen type="toc
      </front>
      <body>
         <p>My TEI Custo
         <schemaSpec ide
            <moduleRef key="core"/>
            <moduleRef key="tei"/>
            <moduleRef key="header"/>
            <moduleRef key="textstructure"/>
         </schemaSpec>
      </body>
   </text>
</TEI>
```

```
<schemaSpec ident="myTEI" docLang="en" prefix="tei_" start="TEI" xml:lang="en">
   <moduleRef key="core"/>
   <moduleRef key="tei"/>
   <moduleRef key="header"/>
   <moduleRef key="textstructure"/>
</schemaSpec>
```

Exkurs ODD: Spezifikation

```
<schemaSpec ident="mein_projekt_schema"
  docLang="en" prefix="tei_" start="TEI"
  xml:lang="en">
```

- <schemaSpec> – formale Schemaspezifikation
- @ident – Dateiname
- @docLang – Dokumentationssprache
- @prefix – Prefix für TEI-Elementpattern
- @start – Einstiegspunkt (Wurzelelement)

Exkurs ODD: Spezifikation

```
<moduleRef key="textstructure"/>
<moduleRef url="http://www.tei-c.org/
  release/xml/tei/custom/schema/relaxng/
  svg11.rng">
```

- <moduleRef> bindet Module ins Datenmodell ein
- @key – Name des TEI Moduls
- @url – bezieht externes Schema in Relax NG Notation ein

Exkurs ODD: Elemente spezifizieren

```
<elementSpec module="textstructure"
  ident="div1" mode="delete"/>
```

- <elementSpec> ändert Standardverhalten von Elementen
- @module – Herkunftsmodul des Elements
- @ident – Elementname
- @mode – Aktion
  - delete | change | add | replace
- @ns – Namespace des Elements

Exkurs ODD: Elemente ändern

- Element aus Modul löschen

```
<elementSpec ident="div1" mode="delete" module="textstructure">
```

- Element ändern: Attribute löschen

```
<elementSpec ident="gap" mode="change" module="core">
  <attList>
      <attDef ident="hand" mode="delete"/>
      <attDef ident="agent" mode="delete"/>
      <attDef ident="cert" mode="delete"/>
     <attDef ident="resp" mode="delete"/>
  </attList>
</elementSpec>
```

Exkurs ODD Spezifikation

```
<classSpec ident="att.global" type="atts"
  mode="change">           <attList>
      <attDef ident="rend" mode="delete"/>
  </attList>
</classSpec>
```

- <classSpec> – spezifiziert Änderungen an Klassen von Attributen oder Elementen
  - Attribute einer Klasse löschen oder hinzufügen
  - Attributwerte ändern
- att.global – Klasse der globalen Attribute
- @type – gibt Klassenart an (hier Attributklasse)

Exkurs ODD: Dokumentieren

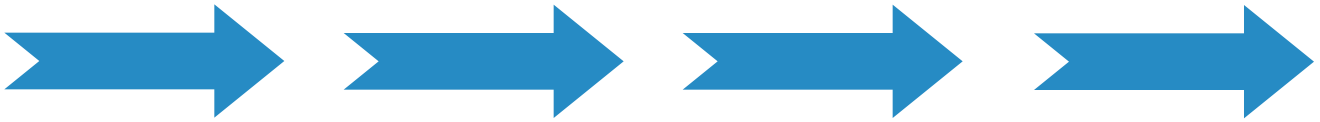
- Dokumentation der Änderung in der formalen Spezifikation

```
<classSpec ident="att.global" type="atts" mode="change"  module="tei">
   <attList>
      <attDef ident="rend" mode="change">
              <valList type="closed" mode="replace">
                  <valItem ident="b">
                      <desc xml:lang="de">b steht für fett (bold).</desc>
               </valItem>
               <valItem ident="i"/>
           </valList>
        </attDef>
  </attList>
</classSpec>
```

Exkurs ODD: Dokumentieren

```
<p>Vom Element <gi>gap</gi> sind die Attribute
  <att>hand</att> und <att>agent</att> gelöscht
  worden. Das globale Attribut <att>rend</att> darf
  nur die Werte <val>b</val> und <val>i</val>
  annehmen. So <tag>moduleRef key="figures"</tag>wird
  ein Modul eingebunden.</p>
```

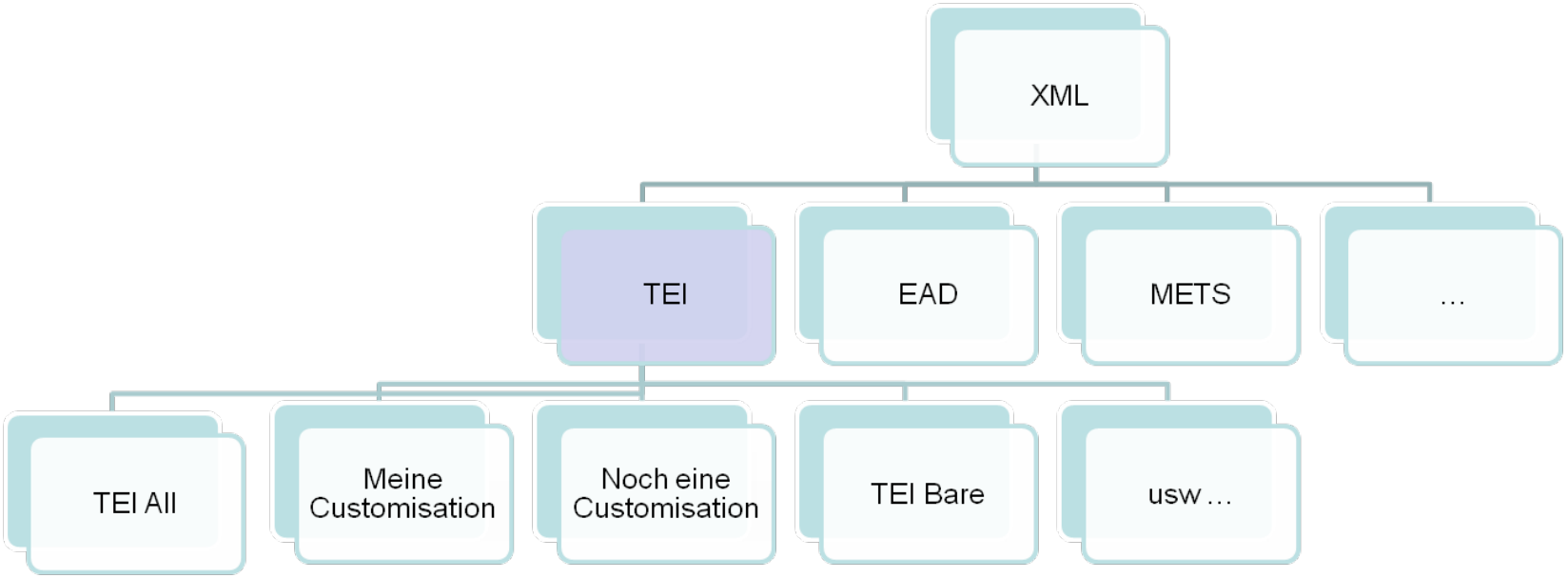
- <gi> – Elementnamen dokumentieren
- <att> – Attribute dokumentieren
- <val> – Attributwerte dokumentieren
- <tag> – Tags dokumentieren



Übung

- Starte bei http://www.tei-c.org/Roma/
- Wähle eine vordefinierte Customization
- Füge notwendige Module hinzu
- Entferne überflüssigen Elemente
- Speichere ODD-File, Schema und Dokumentation
- Prüfe eine TEI-Datei auf Validität mit diesem Schema
  - oder
- Lege eine neue TEI-Datei an, die auf dieses Schema „hört“

TEI und …


XML
TEI
EAD
METS
…
TEI All
Meine Customisation
Noch eine Customisation
TEI Bare
usw …

Los geht’s!

TEI Grundstruktur

- Rootelement **<TEI>**
- Enthält mindestens zwei Unterelemente, nämlich
  - **<teiHeader>** (muss immer vorhanden sein)
  - **<text>** und/oder
  - **<facsimile>** (für Verknüpfung mit Bildern) oder
  - **<fsdDecl>** ‘Feature Structure Declaration‘ (v.a. für Textanalysen, Linguistische Merkmale u.ä.)

TEI Grundstruktur

- 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

<teiHeader>

TEI Guidelines: 2 The TEI Header

- enthält Metadaten zum TEI-Text:
  - Autor, Titel usw.
  - wann erstellt?
  - Quelle(n), Editionsrichtlinien, Versionsgeschichte
  - …
- vier Teile im Header:
  - **<fileDesc>:** notwendig
  - **<encodingDesc>**: fakultativ
  - **<profileDesc>**: fakultativ
  - **<revisionDesc>**: fakultativ

Header 1: <fileDesc>

- Bibliographische Beschreibung des TEI-Dokuments (Autor, Titel, Editor, Projekt, Erstellungsdatum usw.)
- Beschreibung der Quelle(n), z.B. einer Druckausgabe, einer Handschrift, eines Archivguts usw.
- muss enthalten:
  <titleStmt>, <publicationStmt>, <sourceDesc>
- kann enthalten:
  <editionStmt>, <extent>, <seriesStmt>, <notesStmt>

Header 1: <fileDesc>

- <titleStmt>
  - Angaben zu Autor, Titel usw., bezogen auf die digitale Edition (nicht die edierte Vorlage o.ä.)

```
<TEI>
  <teiHeader>
    <fileDesc>
      <titleStmt>
```

Header 1: <fileDesc>

- <publicationStmt>
  - Publikationsdaten **der elektronischen Ausgabe** (nicht der Vorlage)

```
<TEI>
  <teiHeader>
    <fileDesc>
      <publicationStmt>
```

Header 1: <fileDesc>

```
<TEI>
  <teiHeader>
    <fileDesc>
      <sourceDesc>
```

- **<sourceDesc>**
  - Beschreibung **der edierten Quelle**
  - Im einfachsten Fall `<p>digital erstellt</p>`
  - Freie Beschreibung möglich, z.B. `<p>Brief von Heine an Nicolai, Berlin SBB-PK, Nachlass Nicolai, Kasten 7</p>`
  - Bibliographische Aufnahmen mit: **<bibl>**, **<biblStruct>** oder **<biblFull>**
  - Handschriftenbeschreibungen mit **<msDesc>**

Header 1: <fileDesc>

- <bibl>, <biblStruct>, <biblFull> nicht allein in der <sourceDesc>, sondern auch im Dokument selbst (z.B. bei Fußnoten, Bibliographien o.ä.) möglich
  - <bibl> lässt unstrukturierte Beschreibung zu
  - <biblStruct> geeignet für Beschreibung gedruckter Vorlagen
  - <biblFull> ursprünglich entwickelt für Beschreibung von digitalen Ressourcen, bei Beschreibung von Druckvorlagen häufig problematisch
  - <msDesc> speziell für mittelalterliche Handschriften

Header 2: <encodingDesc>

- Beschreibung der editorischen Praxis, u.a.
  - Projektbeschreibung: <projectDesc>
  - Editionsrichtlinien (Normalisierung u.ä.): <editorialDecl>
  - Für Korpora: <samplingDecl>
  - Beschreibung der Verwendung von Tags und ggf. Verknüpfung mit bestimmter Darstellungsweise: <tagsDecl>
  - Evtl. Definition eigener Sonderzeichen

```
<TEI>
   <teiHeader>
     <fileDesc>
       ...
     <encodingDesc>
```

Header 3: <profileDesc>

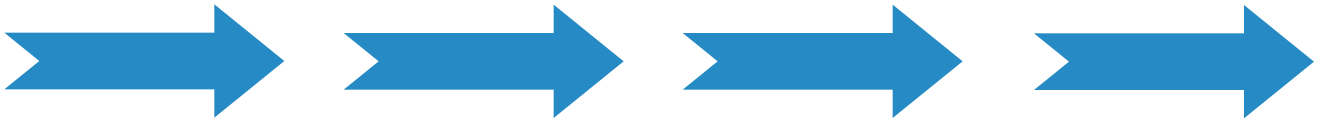
- Kodierung inhaltlicher Informationen über den Text
  - Entstehungszeit
  - Sprache
  - Textsorte
  - v.a. für Sprachcorpora von Interesse, bei 'normalen' Editionen eher selten verwendet

```
<TEI>
   <teiHeader>
     <fileDesc>
       ...
     <encodingDesc>
     <profileDesc>
```

Header 4: <revisionDesc>

- Informationen über die 'Geschichte' des Dokuments
  - Wann erstellt?
  - Wann wurden welche Veränderungen vorgenommen?
  - evtl. Begründung der Veränderungen usw.
  - v.a. sinnvoll bei großen Editionsprojekten mit mehreren MitarbeiterInnen
  - Beispiel

```
<TEI>
   <teiHeader>
     <fileDesc>
       ...
     <encodingDesc>
     <profileDesc>
     <revisionDesc>
```



Übung

- Erstellen Sie in oXygen ein TEI-Dokument mit der Vorlage "TEI All"
- Erstellen Sie ein `<titleStmt>` und `<publicationStmt>` für Ihr eigenes Projekt unter Angabe von Autor, Titel, Herausgeber
- Setzen Sie entweder eine Quellenangabe für Ihr eigenes Projekt oder die folgende Angabe mit `<biblStruct>` um:
- J.W.v.Goethe: Die Leiden des jungen Werthers. Erste Fassung. In: Goethes poetische Werke. Hg. von Liselotte Lohrer. Stuttgart: Cotta 1950. Bd 6, S. 7-130.
- Alternative: Bearbeiten Sie den Header Ihres mit OxGarage erzeugten Dokuments.

Zum Inhalt: <text>

- 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

Strukturelle Gliederungselemente

TEI-Guidelines: 4 Default Text Structure

- **`<div>`** (Division): Abschnitte im Dokument (z.B. Buch, Kapitel, einzelne Gedichte, Akte/Szenen u.ä.)
- Wichtige Attribute:
  - **`@n`** (Nummerierung, z.B. "1.1.2.a", entweder aus der Quelle übernommen oder selbst erstellt)
  - **`@type`** (z.B. "book", "chapter", "poem")
  - **`@xml:id`** (eindeutiger Identifikator, muss dokumentenweit eindeutig sein, und mit einem Buchstaben beginnen, i.d.R. selbst vergeben oder automatisch erzeugt)

Strukturelle Gliederungselemente

- `<div n="x">` vs. `<div1>` bis `<div7>`
- Alternativ zu `<div>`, bei dem die Hierarchiestufe durch die Schachtelung innerhalb des XML-Dokuments definiert ist, können die Hierarchiestufen auch explizit gemacht werden mit `<div1>`, `<div2>` usw.
- Entscheidung durch den Editor
- `<div>` ist flexibler
- `<div1>` usw. z.T. leichter in der nachfolgenden Verarbeitung
- Mischung nicht möglich

<front> und <back>

- Spezialelemente für Vorstücke (Titelblatt, Vorwort, Inhaltsverzeichnis u.ä.) und Nachstücke (Register, Nachwort usw.)
- v.a. bei der Umsetzung gedruckter Vorlagen wichtig
- Enthaltene Elemente können z.B. sein:
  - **<titlePage>**, **<docImprint>**, <byline>, ** ** usw.
- Im Prinzip alle Elemente verfügbar, die auch in **<body>** verfügbar sind. Eher eine vom Herausgeber bestimmte Gliederung

Grundlegende Elemente

TEI-Guidelines: 3 Elements Available in All TEI Documents

- **<p>** (paragraph): Absatz
- **<ab>** (anonymous block): Irgendein Textblock
- **<head>** (head line): Überschrift
- **<lb/>** (line break): Zeilenumbruch (z.B. wenn Zeilenumbrüche der Vorlage mit transkribiert werden)
- **<pb/>** (page break): Seitenumbruch, normalerweise der der Vorlage
- Mit **@n, @type, @xml:id** spezifizierbar

Hervorhebung und wörtliche Rede

- **<hi>** (highlighted): allgemeiner Tag für Hervorhebungen, z.B. Kursiv o.ä., spezifizierbar durch **@rend**
- **<foreign>, <emph>, <distinct>**: Verschiedene Hervorhebungsarten bzw. Markierung ‚ungewöhnlicher' Textteile (Fremdsprachiges, Slang, Archaismen)
- **<q>** für wörtliche Rede (in Anführungsstrichen)
- **<quote>** und **<cit>** für Zitate

Gedichte, Dramen

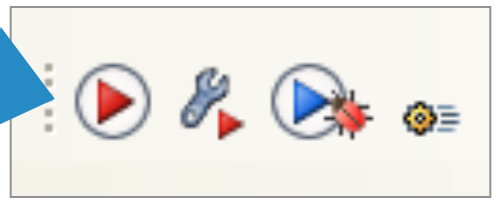
- **<lg>** und **<l>** für Gedichte, Versdramen, gebundene Sprache
- **<l>** bezeichnet die metrische Zeile, **<lb/>** markiert den gedruckten oder handschriftlichen Zeilenumbruch
- Für Dramen stellt die TEI ein umfangreiches Vokabular zur Auszeichnung von Sprechern, Sprechtexte, Bühnenanweisungen usw. zur Verfügung, die wichtigsten sind:
  - **<sp>, <speaker>, <stage>**

Häufig gebraucht:

- **<list>**, **<item>**, **<label>** für Listen
- **<listBibl>** für Literaturlisten
- **<note>** für Anmerkungen (z.B. Fußnoten, Marginalien)
- **<figure>** und **<graphic>** für Illustrationen u.ä.

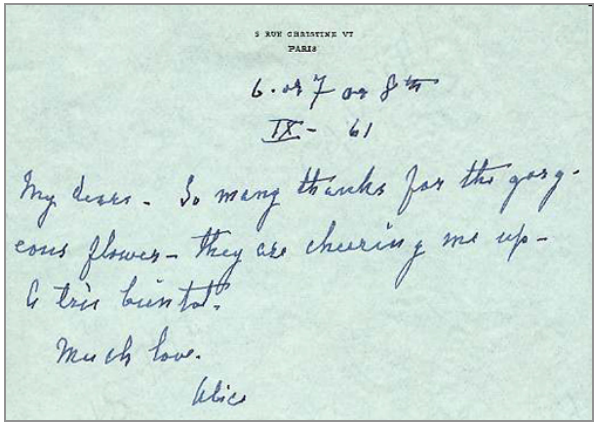
Übung

- Erstellen Sie eine Textgliederung in **<front>**, **<body>** und **<back>**
- Fügen Sie eine fiktive Titelseite ein.
- Kodieren Sie Zeilenumbrüche.
- Zeichnen Sie ein beliebiges Stück mit **<hi>** aus.
- Kodieren Sie ein Stück wörtliche Rede.
- Transformieren Sie nach xHTML per vordefiniertem Szenario in oXygen.

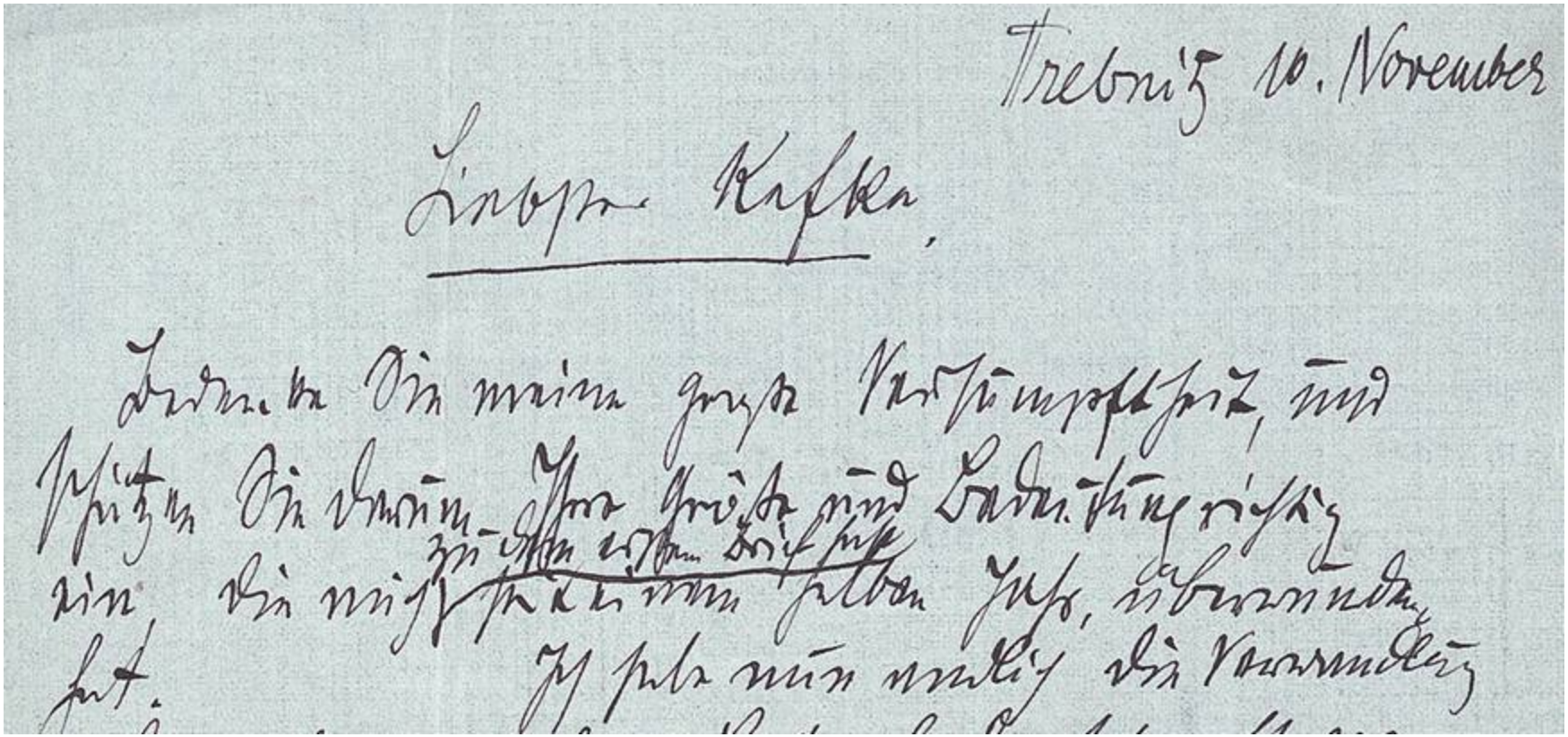


Briefe

- Briefe können als eigenes Dokument, als Texte innerhalb einer `<group>` oder als `<div>`-Elemente ediert werden
- Ggf. mit `<teiCorpus>` gruppieren
- **`<opener>`** und **`<closer>`** als Container-Elemente, z.B. für
  - **`<dateline>`**: Ort und Datum
  - **`<byline>`**: Verfasserangabe
  - **`<salute>`**: Gruß
  - **`<signed>`**: Unterschrift



Beispiel



Transkription pur

Brief von Franz Werfel an Franz Kafka

Trebnitz 10. November

Liebster Kafka.

Bedenken Sie meine große Versumpftheit, und schätzen Sie darum Ihre Größe und Bedeutung richtig ein, die mich (zu dem ersten Brief heute) seit einem halben Jahr, überwunden hat.

TEI-Kodierung

TEI Kodierung

```
<?xml version="1.0" encoding="UTF-8"?>
<div type="letter">
  <head>Brief von Franz Werfel an Franz Kafka</head>
  <opener>
    <dateline rend="right">Trebnitz 10. November</dateline>
    <salute>
      <hi rend="underline">Liebster Kafka.</hi>
    </salute>
  </opener>
  <p>Bedenken Sie meine große Versumpftheit, und <lb/>
    schätzen Sie darum Ihre Größe und Bedeutung richtig <lb/>
    ein, die mich zu dem ersten Brief heute seit einem halben Jahr, überwunden <lb/>
    hat.</p>
  <!-- ... -->
  <closer>
    <salute>Herzlich Ihr </salute>
    <signed>Franz Werfel</signed>
  </closer>
</div>
```

Exkurs: Sonderzeichen

Exkurs: Sonderzeichen

- Bei Transkriptionen älterer und/oder handschriftlicher Texte häufig Sonderzeichen
- Inzwischen zahlreiche Sonderzeichen im Unicode-Standard definiert
- Außerdem bietet die TEI im gaiji-Modul Elemente an, durch die Sonderzeichen definiert, beschrieben und in der Transkription eingesetzt werden können

Was ist Unicode?

- „Internationaler Standard, in dem langfristig für jedes sinntragende Schriftzeichen oder Textelement aller bekannten Schriftkulturen und Zeichensysteme ein digitaler Code festgelegt wird“ (http://de.wikipedia.org/wiki/Unicode)
- Bzw. festgelegt werden soll. (Oliver Duntze)

Warum Unicode

- Ältere Zeichencodierungen konnten lediglich 128 (ASCII, 7 bit) oder 256 (z.B. ISO-8859, 8 bit) Zeichen codieren
- Folge: für unterschiedliche Schriftsysteme mussten verschiedene Zeichencodierungen entwickelt werden und ggf. angegeben werden, in welcher Zeichencodierung eine Datei gespeichert ist (z.B. ISO-8859-1, ISO-8859-5 usw.)

Warum Unicode?

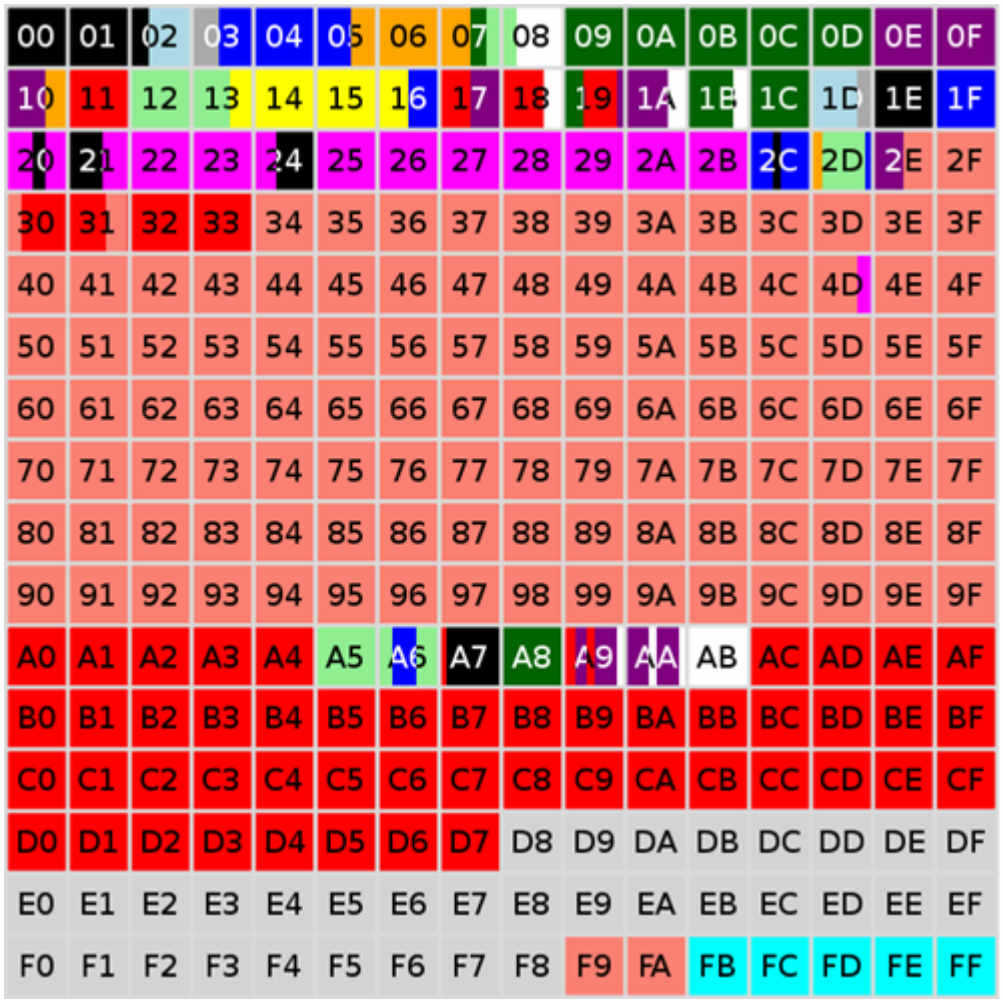
- Unicode soll die verschiedenen miteinander inkompatiblen Zeichenkodierungen ersetzen
- In Unicode 1.0 sollten alle Schriftzeichen der Welt durch 65.536 (2^{16}) sog. „codepoints“ repräsentiert werden
- Inzwischen erweitert auf 17 Bereiche („planes“) von je 65.536 codepoints -> 1.114.112 mögliche Zeichen

Warum Unicode?

- Unicode-Standard wird ständig durch das „Unicode Consortium“ kontinuierlich weiterentwickelt
- Aktuelle Version ist Unicode 6.0.0 (Jan. 2011)
- Lateinisch, Griechisch, Kyrillisch, Arabisch, Hebräisch, CJK
- Aber auch so Schriften wie Balinesisch, Gotisch, Glagolitisch, Ogham, Linear B usw.
- Mehrere „Private Use Areas“ (PUA)
- Ergänzungswünsche können (und sollten) dem Unicode Consortium gemeldet werden

Was gibt es in Unicode?

- „Normale“ Schriftzeichen: a b c δ Д ى א ण
- Satzzeichen „ “ ? ! ,
- Whitespace
- Combining Diacritical Marks: ◌̈ ◌̋ ◌̓
- Vorkombinierte Zeichen á ä ť ů پں ḇ̣ ḛ̆
- Symbole ☭ ♘ ♂ ☿ ∑
- Steuerzeichen Wagenrücklauf, EOF
- …


00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F
10 11 12 13 14 15 16 17 18 19 1A 1B 1C 1D 1E 1F
20 21 22 23 24 25 26 27 28 29 2A 2B 2C 2D 2E 2F
30 31 32 33 34 35 36 37 38 39 3A 3B 3C 3D 3E 3F
40 41 42 43 44 45 46 47 48 49 4A 4B 4C 4D 4E 4F
50 51 52 53 54 55 56 57 58 59 5A 5B 5C 5D 5E 5F
60 61 62 63 64 65 66 67 68 69 6A 6B 6C 6D 6E 6F
70 71 72 73 74 75 76 77 78 79 7A 7B 7C 7D 7E 7F
80 81 82 83 84 85 86 87 88 89 8A 8B 8C 8D 8E 8F
90 91 92 93 94 95 96 97 98 99 9A 9B 9C 9D 9E 9F
A0 A1 A2 A3 A4 A5 A6 A7 A8 A9 AA AB AC AD AE AF
B0 B1 B2 B3 B4 B5 B6 B7 B8 B9 BA BB BC BD BE BF
C0 C1 C2 C3 C4 C5 C6 C7 C8 C9 CA CB CC CD CE CF
D0 D1 D2 D3 D4 D5 D6 D7 D8 D9 DA DB DC DD DE DF
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 EA EB EC ED EE EF
F0 F1 F2 F3 F4 F5 F6 F7 F8 F9 FA FB FC FD FE FF
Lateinische Schriften und Symbole
Lautschriften
Andere europäische Schriften
Nahost- und Südwestasiatische Schriften
Afrikanische Schriften
Südasiatische Schriften
Südostasiatische Schriften
Ostasiatische Schriften
CJK-Ideogramme
Kanadische Silben
Symbole
Diakritika
UTF-16-Surrogates und privater Nutzungsbereich
Verschiedene Zeichen
Nicht belegte Codebereiche

Wie finde ich mein Zeichen?

- Codecharts unter www.unicode.org
- Datenbank unter www.decodeunicode.org
- Oder www.isthisthingon.org/unicode/index.php (The UniSearcher)

Kodierung von Unicode in XML-Dateien

- Entweder Zeichen direkt einfügen, z.B. mit Oxygen:
  - α (intuitiv lesbar, wird aber – je nach Zeichensatz – nicht angezeigt)
- Oder mit Zeichenentitäten:
  - Hexadezimal: `&#x0364;` (gut, entspricht dem Codepoint)
  - Dezimal: `&#945;` (bitte nicht!)

Kombinierende diakritische Zeichen

Z.B. übergestelltes ᵘ (codepoint U+0367)

- oͧ -> oᵘ vs. o□
- Generelles Problem:
  - Ungewöhnliche Zeichen werden nur mit entsprechenden Zeichensätzen und entsprechender Software ordentlich angezeigt
  - Empfehlenswerte Schriften u.a. Arial Unicode MS, Junicode, Code2000, Schriften des GW
  - Weniger empfehlenswert: Mediaevum

Was tun, wenn Unicode nicht weiterhilft?

- Möglichkeit, die Private Use Areas zu verwenden (U +E000-F8FF, Planes 16 u. 17)
- TEI bietet mit den Elementen <char>, <glyph> und <g> eine Methode zur Definition von Sonderzeichen an
- Character -> ein bestimmter „Buchstabe“ (z.B. ein A)
- Glyph -> eine bestimmte Ausführung eines Buchstabens („langes s“, „rundes r“)

<charDecl>

- Teil von teiHeader/encodingDesc
- Enthält <char>- und <glyph>-Elemente
- Darin u.a.:
  - <charName> bzw. <glyphName>
  - <charProp>
  - <desc>
  - <mapping>
  - <figure>

```
<TEI>
  <teiHeader>
    <fileDesc>
      ...
    <encodingDesc>
      <charDecl>
```

Ein Beispiel


Incipit tractatus de efficacia aque bene
dicte, per venerandū magistrū Johannē de turre cremata, sacre
theotogie pfessorem, ordinis predicatorꝝ, tpe concilij Basilienſis
cōpilatus, cōtra Petrū anglicū hereticorꝝ defenſorē in bohemia

Beispiel: 

```
<encodingDesc>
  <charDecl>
    <char xml:id="pstroke">
      <charName>LATIN SMALL LETTER P WITH STROKE</charName>
      <desc>unten durchgestrichenes p, meist als Abbreviatur für per</desc>
      <charProp>
        <unicodeName>general-category</unicodeName>
        <value>Ll</value>
      </charProp>
      <mapping type="standardized">p</mapping>
      <figure>
        <graphic url="min_per01-01.jpg"/>
      </figure>
      <note>ganz häufig verwendet</note>
    </char>
  </charDecl>
</encodingDesc>
```


```
<TEI>
…
<text<
  <body>
    <p>t<g ref="#pstroke">empor</g>e
  concilij Baſilienſis</p>
  </body>
</text>
</TEI>
```