

Referat 03: TEI und das Modul "Core"

DARIAH-DE Tutorials *Digitale Textkodierung mit TEI*

Redaktion: Christof Schöch (Univ. Würzburg)

Version. 1.0, 02/2014

Grundlage der Folien: DH@Oxford 2012

Lizenz: Creative Commons Attribution 4.0 International (BY-SA)



Überblick

Dieses Input-Referat beinhaltet:

1. Hintergrundinformationen zur Text Encoding Initiative
2. Informationen zu grundlegenden Mechanismen der TEI

1. Hintergrundinformationen zur Text Encoding Initiative

Was ist die Text Encoding Initiative?

Ein Standard zur Kodierung von Texten

- Guidelines und Schemata
- Tools (Roma, OxGarage)

Eine Institution zur Pflege des Standards

- Mitglieder (Institutionen, Privatpersonen)
- Board of Directors, (Technical) Council, SIGs

Eine Community rund um den Standard

- Nutzer rund um die Welt
- Members' Meeting and Conference
- Mailing-Listen
- TEI Journal

Relevanz der TEI: warum das alles?

Weil Datenaustausch nützlich ist

- zwischen Menschen / Projekten
- (zunehmen) zwischen Maschinen

Weil Datenintegration nützlich ist

- von verschiedenen Datentypen
- aus verschiedenen technischen Kontexten

Weil wir Daten langfristig speichern wollen

- (zunehmend komplexe) Daten
- Metadaten

Ziele der TEI

Datenaustausch

- "interchange" oder "interoperability"

Universelle Einsetzbarkeit

- Texte in allen Sprachen (und Schriftsystemen)
- aus allen Epochen

Hilfestellung

- für Anfänger: Anwendung der Guidelines
- für Spezialisten: Anpassung, Weiterentwicklung der Guidelines
- Ergebnis: flexibler, modularer, vielschichtiger Standard

Geburtsstunde der TEI

1987 in Europa

- Treffen bzgl. Standards für historische Quellen und Daten

1987 in den USA

- Treffen zur Idee von "text encoding guidelines"



Entwicklung der TEI : P1-P3

1988-1990

- Erste Runde der Entwicklung: TEI P1
- Konsultationen; etwa 50 Personen

1990-1992

- Zweite Runde der Entwicklung: TEI P2
- Arbeitsgruppen; mehrere hundert Personen

1993-1994:

- Beide Vorschläge werden verbunden: TEI P3

1995-1999:

- Dissemination der Guidelines (ohne Institution oder Budget)

Entwicklung der TEI : P4-P5

1999:

- Gründung der TEI als Konsortium

2001-2003:

- Transformation von SGML (P3) zu XML (P4) ;
- Beginn einer kompletten Überarbeitung

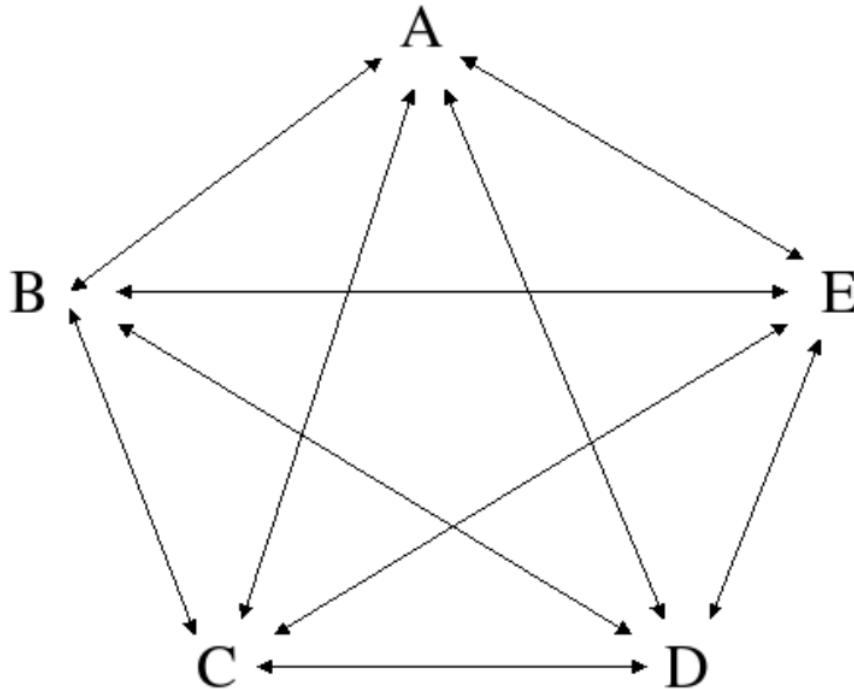
2003:

- TEI P5 erscheint
- regelmäßig Überarbeitungen

Webseite: <http://www.tei-c.org/>

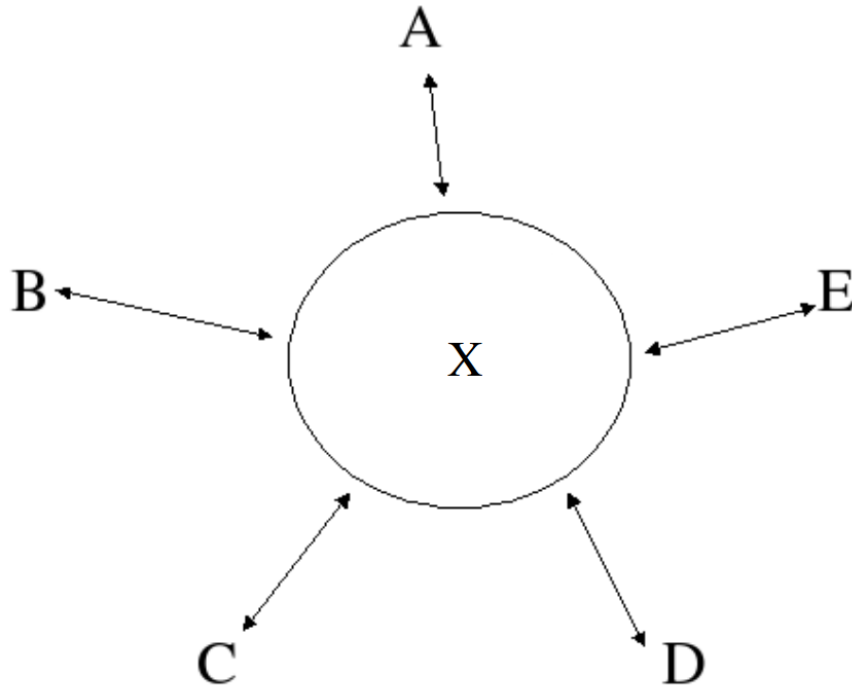
TEI als "pivot format"

Ohne "pivot format": für 5 Formate braucht man 20 Transformationen ($n \cdot n - 1$)



TEI als "pivot format"

Mit "pivot format": für 5 Formate braucht man 10 Transformationen ($n*2$)



TEI ist XML

(siehe Input 1)

TEI "erbt" Eigenschaften von XML

- Inhalt und Funktion zählt
- Darstellung folgt später (mit weiteren "X-Technologien)
- generische Lösungen
- muss wohlgeformt und valide sein

Außerdem

- XML-Software kann für TEI verwendet werden (bspw. Oxygen)

Die TEI ist mehr als ein Schema!

Die TEI-Architektur bietet:

- Begriffe und Definitionen für hunderte textkritische Phänomene und Unterscheidungen
- Nutzungshinweise für alle Elemente, die in den Modulen genannt werden
- Modulares System von Modulen, um angepasste Kodierungs-Schemata zu generieren
- Mechanismus, um alle Elemente auszuwählen, zu verändern, und zu kombinieren
- Ein spezifisches Set von Stylesheets, um Schemata und Dokumentation zu generieren

Was leisten die Guidelines letztlich?

Über den rein pragmatischen Aspekt hinaus

- konsens-basierte Strategie, um Text-Ressourcen zu organisieren und zu dokumentieren
- Dadurch auch: eine enzyklopädische Kodifikation aktueller editorischer Praxis
- Eine Art, darüber nachzudenken, was "Text" eigentlich ist, und was eine "Edition" leisten soll

Der TEI-Werkzeugkasten: Software

Von der TEI für TEI:

- Roma – Webservice für Schemata: <http://www.tei-c.org/Roma>
- Oxgarage – Dokument-Konversion nach oder aus TEI
<http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient/>

Weitere Tools

- Publikationstools: Versioning Machine, SADE, TEI Boilerplate, TEICHI for Drupal, TEI Display for Omeka, etc.
- Sonstige Tools: TextGrid (Arbeitsumgebung), TEUCHOS (msDesc), eLaborate (Transkription), TILE (Text/Bild), CollateX (Stemmata), etc.
- ...und natürlich können alle gegenwärtigen und zukünftigen XML-Editoren und -Tools mit TEI umgehen, denn es ist 100% XML.

2. Grundlegende Mechanismen: Elemente aus dem Modul "Core"

Was kodieren?

(*Punch* ist eine berühmte, humoristische Zeitschrift, die in England von 1841 bis 1992 wöchentlich erschien.)

July 1, 1914.]

PUNCH, OR THE LONDON CHARIVARI.

1



Vol. CXLVII.

PROGRESS.

[“Giving evidence recently before a Select Committee of the House of Commons, Miss C. E. Collet, of the Home Office, said the commercial laundry was killing the small hand laundry.”—*Evening News*.]

The little-crafts! How soon they die!
In cottage doors no shuttle clicks;
The hand-loom has been ousted by
A large concern with lots more sticks.

The throb of pistons beats around;
Great chimneys rise on Thames's
banks;

The same phenomena are found
In Sheffield (Yorks) and Oldham
(Lancs).

No longer now the housewife makes
Her rare preserves, for what's the
good?

The factory round the corner fakes
Raspberry jam with chips of wood.

'Tis so with what we eat and wear,
Our bread, the boots wherein we splash;
'Tis so with what I deemed most fair,
Most virginal of all—the Wash.

'Tis this that chiefly, when I chant,
Fulfills my breast with sighs of ruth,
To think that engines can supplant
The Amazons I loved in youth.

That not with tender care, as erst
By spinster females fancy-free,
These button-holes of mine get burst
Before the shift comes back to me;

That mere machines, and not a maid
With fingers fatuously plied,
The collars and the cuffs have frayed
That still excoriate my hide;

That steam reduces to such states
What once was marred by human skill,
That socks are sundered from their mates
By means of an electric mill;

That not by Cupid's coy advance
(Some crone conniving at the fraud),
But simply by mechanic chance,
I get this handkerchief marked
“Maud.”

This is, indeed, a striking change;
I sometimes wonder if the world
Gets better as the skies grow strange
With coils of smoke about them
curled.

If the old days were not the best
Ere printed formulas conveyed
Sorrow about that silken vest
For all eternity mislaid;

Ere yet the unwieldy motor-van
Came clattering round the kerbstone's
brink,
Its driver dreaming some new plan
To make my mauve pyjamas shrink.
Evon.

Was sollte/will man kodieren?

Informationen zur Identifikation

- Jahr, Nummer der Ausgabe; Seitenzahl

Teile des Dokuments

- Teile mit einem Bild, Gedicht, Prosa, Drama, oder andere Dinge

Weitere Aspekte

- Bild, Überschrift
- Strophen, Zeilen
- Absätze
- Wortwechsel, Bühnenanweisungen
- und vieles mehr!

Modul 4: Default Text Structure

Grundlegende Elemente, um Textstrukturen zu kodieren.

- Dokument-Abschnitte: `<text>`, `<front>`, `<body>`, `<back>`
- Abschnitte: `<div>` und `<div1>`, `<div2>`, `<div3>`, etc.
- Briefe u.a.: `<head>`, `<salute>`, `<opener>`, `<epigraph>`, `<closer>`, `<byline>`, `<dateline>`, `<salute>`, `<signed>`
- Zitate u.a.: `<cit>`, `<quote>`

Makrostruktur

Alle Ausgaben eines Jahres von Punch bilden gemeinsam einen Band. Wie könnte man dies strukturieren?

Eine Möglichkeit

- Band und jede Ausgabe = `<text>`
- Anfang (Titelblatt) des Bands = `<front>`
- Inhaltsverzeichnis am Ende = `<back>`
- Ausgaben des Bands = `<group>`
- Teile einer Ausgabe = `<div>`

Makrostruktur: Beispiel

```
<text xml:id="v147">
  <front>
    <!-- introd. materials for volume 147-->
  </front>
  <group>
    <text xml:id="I1914-07-01">
      <body><!-- first issue (1 July) --></body>
    </text>
    <text xml:id="I1914-07-15">
      <body><!-- second issue (15 July) --></body>
    </text>
    <!-- etc... -->
  </group>
  <back><!-- index, appendix etc. --></back>
</text>
```

Innherhalb eines Abschnitts <div>

Abschnitte (<div>) können kleinere Einheiten beinhalten

- <head> (heading) – Titel, Untertitel, etc.
- <p> (paragraph) - Absatz
- <sp> (speech) – Rede in einem Drama; beinhaltet <p>, aber auch <stage> und <speaker>
- <list> - Liste: beinhaltet <head>, <label>, <item>
- <table> - Tabelle: beinhaltet <row>, darin <cell>
- <l> (verse line) – Verszeilen; können in <lg> (Strophen) gruppiert sein
- <figure> - Abbildung, kann <graphic>, <figDesc>, <head> beinhalten

Außerdem können Abschnitte weitere Abschnitte beinhalten

- <div>Abschnitt<div>Unterabschnitt</div></div>

Innerhalb eines Absatzes <p>

Innerhalb der Makrostruktur-Elemente stehen zahlreiche Elemente zur Verfügung

- Grenz-Elemente: Seiten-, Spalten-, Zeilenumbruch
- Hervorhebung, Emphase, Zitate
- Editorische Intervention: Korrekturen, Normalisierung
- Namen, Zahlen, Daten, Personen, Orte
- Links und Querverweise
- Kommentare, Annotationen, Indizes
- Abbildungen und Grafiken
- Bibliografische Angaben
- Linguistische Analysen, bspw. auf Wortebene

Ein einfacher Dialog aus "Punch"



WHEN THE SHIPS COME HOME.

GREECE. "ISN'T IT TIME WE STARTED FIGHTING AGAIN?"

TURKEY. "YES, I DARESAY. HOW SOON COULD YOU BEGIN?"

GREECE. "OH, IN A FEW WEEKS."

TURKEY. "NO GOOD FOR ME. SHAN'T BE READY TILL THE AUTUMN."

Beispiel: Dialog

```
<div type="cartoon">
  <figure>
    <head>When the ships come home</head>
    <figDesc>A man in Turkish dress lounges on a sofa;
another man, in traditional Greek costume, stands beside
him.</figDesc>
    <graphic url="../Punch/147_03-003.png"/>
  </figure>
  <sp>
    <speaker>Greece.</speaker>
    <p>Isn't it time we started fighting again?</p>
  </sp>
  <sp>
    <speaker>Turkey.</speaker>
    <p>Yes, I daresay. How soon could you begin?</p>
  </sp>
  <!-- ... --!>
</div>
```

Eine Intervention des Editors



Ein Beispiel aus "Punch"

“Excuse me sir, but would you like to buy a nice little dawg?”

Wir können mit TEI...

Festhalten, dass dort tatsächlich "dawg" steht: mit `<orig>`

```
<p>Excuse me sir, but would you like to buy a nice little  
<orig>dawg</orig>?</p>
```

die Stelle korrigieren und festhalten, dass "dog" eine Regularisierung des Editors ist; mit `<reg>`

```
<p>Excuse me sir, but would you like to buy a nice little  
<reg>dog</reg>?</p>
```

Wir können mit TEI...

Oder beides festhalten, und dafür `<choice>` verwenden

```
<p>Excuse me sir, but would you like to buy a nice little  
<choice>  
  <orig>dawg</orig>  
  <reg>dog</reg>  
</choice>?</p>
```

Modul 2: Header (für Metadaten)

Für das Punch-Projekt und viele andere, folgende Informationen müssen festgehalten werden:

- Identifikation der Ressource ("Um was handelt es sich?")
- Zuordnung von Verantwortlichkeiten ("Wer hat wann was gemacht?")
- Angabe der Quelle(n) ("Was wurde woher genommen?")
- Publikations-Angaben: ("Wie wird diese Ressource zugänglich gemacht?")
- Dokumentation der Kodierungspraxis ("Was bedeutet der Markup?")

Der TEI Header erlaubt all dies und einiges mehr

Der <teiHeader>

Oberstes Element

- <teiHeader> mit <fileDesc>

Drei obligatorische Bereiche:

- <titleStmt> - Angaben zum Titel, mit: <title>, <author>, <editor>
- <publicationStmt> - Angaben zur Publikation, mit: <p> oder strukturierten Informationen
- <sourceDesc> - Angaben zur Quelle, auf dem das Dokument beruht, mit <p> oder strukturierten Informationen

Viele optionale Dinge, bspw:

- <msDesc> - Manuskript-Beschreibung
- <revisionDesc> - Bearbeitungen des TEI-Dokuments, mit <change>

Ein einfacher Header für "Punch"

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Punch, or the London Charivari, Vol. 147, July 1,
1914</title>
    </titleStmt>
    <publicationStmt>
      <idno type="gutenberg">24357</idno>
      <availability><p>This text is freely available for re-use
under US and UK law [...].</p>
      </availability>
    </publicationStmt>
    <sourceDesc><p>This text is a TEI version of a Project Gutenberg
text originally located at <ptr
target="http://www.gutenberg.org/dirs/2/4/3/5/24357/" />.</p>
    </sourceDesc>
  </fileDesc>
  <revisionDesc>
    <change when="2008-07-26T23:49:55.968+01:00" />
  </revisionDesc>
</teiHeader>
```

Modul 3: "core"

Rund 50 Elemente, die man fast immer braucht; der Kern ("core") des Systems.

Grundlegende Texteigenschaften

- Prosa: `<p>` (Absatz) und `<q>` (Quotation)
- Vers: `<lg>` (linegroup), `<l>` (line)
- Drama: `<sp>` (speech), `<speaker>`, `<stage>` (Bühnenanweisung)

Weitere grundlegende Elemente

- Leere Elemente: `<pb/>`, `<lb/>`, `<milestone/>`, etc.
- Bibliographische Angaben: `<bibl>`, `<monograph>`, `<author>`, `<title>`, `<publisher>`, `<pubPlace>`, und viele mehr

Beispiel: Versdrama

```
<speaker>Mephistopheles</speaker>
<lg type="stanza">
  <l>Ich möcht' mich gleich dem Teufel übergeben,</l>
  <l>Wenn ich nur selbst kein Teufel wär'!</l>
</lg>
</sp>
<sp>
  <speaker>Faust</speaker>
  <lg type="stanza">
    <l>Hat sich dir was im Kopf verschoben?</l>
    <l>Dich kleidet's, wie ein Rasender zu toben!</l>
  </lg>
</sp>
```

Beispiel: Bibliographische Angabe

```
<biblFull>
  <titleStmt>
    <title>Envisioning Information</title>
    <author>Tufte, Edward R[olf]</author>
  </titleStmt>
  <extent>126 pp.</extent>
  <publicationStmt>
    <publisher>Graphics Press</publisher>
    <pubPlace>Cheshire, Conn. USA</pubPlace>
    <date>1990</date>
  </publicationStmt>
</biblFull>
```

Zum Schluss: globale Attribute

Einige Merkmale lassen sich auf (fast) alles anwenden

- Identität
- Sprache
- Aussehen

TEI bietet dafür "globale Attribute" an

- `@xml:id` - "unique identifier"
- `@n` – Name oder Nummer
- `@xml:lang` – Sprache (nach ISO Standard)
- `@rend` und `@rendition` – visuelles Erscheinungsbild