Einführung in die *Text Encoding Initiative* (TEI)

Edirom-Summerschool

Peter Stadler

27.–29. September 2010

Was wird in diesem Kurs behandelt?

„Einführung in die *Text Encoding Initiative* (TEI)“

- Wer oder was ist der/die/das TEI?
- Wofür, warum, wie viel Markup?
- Arbeiten mit digitalen Texten
- Eigene XML-Schema-Anpassungen

Kursablauf

Tag 1 Einführung, Aufbau einer TEI-Datei, typographisches und inhaltliches Auszeichnen, interne Verweise

Tag 2 Texte in Versform, Apparat und Textkritik

Tag 3 TEI-Maßschneiderung mit ODD, digitale Editionsrichtlinien

Was ist ‚TEI'?

TEI = Text Encoding Initiative

Institution Die TEI Community, vertreten durch das TEI Konsortium

Codierungsschema Die Auszeichnungssprache TEI, bzw. die durch TEI definierte Menge von (XML-)Elementen

Guidelines Die formale Beschreibung des Kodierungsschemas, hrsg. durch das TEI Konsortium

Die Institution TEI

- 1987 Konstituierendes Treffen im Vassar College, Poughkeepsie, NY
- 1988 Beginn der Arbeit der TEI als internationales, mehrsprachiges Projekt zur Entwicklung von Richtlinien für die Vorbereitung elektronischer Texte für die wissenschaftliche Forschung
- 1990 Veröffentlichung des ersten Vorschlags der Richtlinien P1 (= „Proposal 1“)
- 1994 (Erste) Veröffentlichung der *Guidelines for Electronic Text Encoding and Interchange* in der Version P3
- 2000 Bildung eines TEI-Konsortiums als Forum zur Weiterführung der Arbeit

Die TEI ist eine Initiative der Wissenschaft für die Wissenschaft. Zu Beginn maßgeblich getragen durch die *Association for Computers and the Humanities*, die *Association for Computational Linguistics*, und die *Association for Literary and Linguistic Computing*.

Die TEI Community

Seit Ende 2000 existiert das TEI-Konsortium als non-profit-Organisation, das demokratisch konstituiert und akademisch und ökonomisch unabhängig ist.
Die Community partizipiert und kommuniziert über diverse Kanäle:

Mailinglisten allen voran die TEI-L, aber auch alle SIGs unterhalten eigene Mailinglisten

SIG (Special Interest Group) existieren für diverse Spezialgebiete wie Correspondence, Education, Manuscripts etc.

Mitgliedschaft Jeder kann für einen jährlichen Beitrag von 50 $ persönliches Mitglied der TEI werden. Daneben gibt es auch institutionelle Mitgliedschaften

Die TEI Guidelines

- beschreiben ein Standardformat für den Datenaustausch
- stellen eine Hilfestellung für das Codieren von Texten in diesem Format dar
- decken alle textuellen Merkmale aller Textsorten, die von Wissenschaftlern untersucht werden, ab
- sind softwareunabhängig
- ermöglichen benutzerdefinierte Erweiterungen
- sind konform zu aktuellen und zukünftigen Standards

Das Codierungsschema TEI

- besteht aus diversen *Modulen*, die alle eine bestimmte Anzahl von XML Elementen und Attributen deklarieren
- Elemente können zu einer oder mehreren *Klassen* gehören
- Abhängig von diesen Klassen wird der *Content* und die Attribute der Elemente definiert

Die TEI Module

Modulname	Kurzbeschreibung
analysis	Analysis and Interpretation
certainty	Certainty and Uncertainty
core	Common Core
corpus	Metadata for Language Corpora
dictionaries	Print Dictionaries
drama	Performance Texts
figures	Tables, Formulae, Figures
gaiji	Character and Glyph Documentation
header	Common Metadata
iso-fs	Feature Structures
linking	Linking, Segmentation, and Alignment
msdescription	Manuscript Description
namesdates	Names, Dates, People, and Places
nets	Graphs, Networks, and Trees
spoken	Transcribed Speech
tagdocs	Documentation Elements
tei	TEI Infrastructure
textcrit	Text Criticism
textstructure	Default Text Structure
transcr	Transcription of Primary Sources
verse	Verse

Aufbau einer TEI Datei

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Title</title>
      </titleStmt>
      <publicationStmt>
        <p>Publication information</p>
      </publicationStmt>
      <sourceDesc>
        <p>Information about the source</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p>Some text here.</p>
    </body>
  </text>
</TEI>
```

Der Header

Der TEI Header ist deutlich von dem *front matter* des Textes selbst zu trennen. Ersterer beinhaltet Metainformationen über den Text, zweiterer die Originaltitelei des Quelldokuments.

fileDesc (file description) beinhaltet die vollständige bibliographische Beschreibung der elektronischen Datei (und der zugrundeliegenden Vorlage(n))

encodingDesc (encoding description) dokumentiert die Richtlinien der Übertragung in die elektronische Form

profileDesc (text-profile description) ermöglicht die detaillierte Beschreibung nicht-bibliographischer Textmerkmale, besonders der auftretenden Sprachen, Domäne, Setting etc.

revisionDesc (revision description) dokumentiert die Dateirevisionen

Der Body

front (front matter) enthält alle einleitenden Texte wie Titelei, Vorwort, Widmungen etc.

body (text body) enthält den gesamten Textkörper eines definierten Textes, ausschließlich der front bzw. back matters.

group enthält den Textkörper eines zusammengesetzten Textes, der als Sammlung distinkter Einzeltexte angesehen wird, bsplw. eine Gedichtsammlung o. ä.

back (back matter) enthält alle Anhänge etc., die den Text beschließen

Erweiterte Textstruktur

div (text division) Eine Einheit (z. B. Kapitel) innerhalb von `front`, `body`, oder `back`

head (heading) Eine Überschrift

p (paragraph) Markiert einen (Prosa-) Absatz

Erweiterte Textstruktur

```
16  <text>
17    <body>
18      <div>
19        <head>Chapter 1</head>
20        <p>Some text …</p>
21        <figure>
22          <head>The TEI batch</head>
23          <graphic url="http://www.tei-c.org/logos/TEI-glow.png"/>
24        </figure>
25        <p>Some more text …</p>
26      </div>
27      <div>
28        <head>Chapter 2</head>
29        <p>Some text …</p>
30      </div>
31    </body>
32  </text>
```

Übung 1 – Typographisches Auszeichnen

Ziel: Analysieren der Vorlage, Erstellen der notwendigen Meta-Informationen im `teiHeader` und Erfassen von typographischen Auszeichnungen.

201 202

ALLGEMEINE

MUSIKALISCHE ZEITUNG.

Den 19ten März. Nº. 12. 1817.

Ueber die Oper, Undine, nach dem Märchen gleiches Namens von Fried. Baron de la Motte Fouqué selbst bearbeitet, mit Musik von E. T. A. Hoffmann, und zuerst auf dem königl. Theater zu Berlin erschienen.

Als ich den Vorsatz fasste, etwas über dieses schöne Werk öffentlich zu sagen, wandelte unwillkürlich die Form der Anzeigen, Recensionen, oder wie man es immer nennen will, gleichen Zweckes, vor meinem Innern vorüber, indem sich mir zugleich vergegenwärtigte, wie ungemein schwierig es sey, ein bestimmtes Bild des beurtheilten Gegenstandes durch sie zu erhalten; oder etwas dem Eindrucke ähnliches, den das Werk zu machen fähig ist. Es schien mir dabey fast immer, entweder auf die gewöhnlichen Gesellschaftsurtheile hinauszulaufen, wo ohne weitere Beweisführung eine Partey es gut, die andere schlecht findet, die gemässigtere es weder verwirft noch erhebt, und alles nur Gewicht und Glaubwürdigkeit durch die schreibung nur höchst selten ganz fühlbar zu machen. Es versteht sich von selbst, dass diese Meynung auch vielfältiger Beschränkung unterliegt, und namentlich bey schon allgemein verbreiteten Kunstwerken, deren Bau und Structur zu zergliedern, nur heilbringend für die Belehrung-Suchenden seyn kann. In vorliegendem Falle aber, wo es blos Zweck ist, das Publicum auf ein Werk aufmerksam zu machen, indem man die geistige Region anzudeuten sucht, in der es sich bewegt, und die Gestalt, die der Componist ihm verliehen hat, in bezeichnenden Umrissen darstellen will, scheint es mir nothwendig erst auseinander zu setzen, wie der Beurtheiler selbst sehe, glaube, und denke, woraus dann leicht das Resultat für Jeden zu ziehen ist, in wiefern er seinen hieraus entspringenden Urtheilen beypflichten könne. In dieser Hinsicht glaube ich das folgende Bruchstück aus einer grössern Arbeit von mir *), noch der eigentlichen Anzeige der Oper voranschicken zu müssen, weil es auch überdies die Gestaltung der Oper *Undine* grösstentheils ausspricht. —

Hinweise

Vorgabe: TEI-Lite-Schema

Guidelines:

- Minimal and Recommended Headers
- Highlighting and Quotation
- Milestone Elements

Übung 2 – Inhaltliches Auszeichnen

Ziel: Erfassen von Personen, Orten, Werken und Rollen sowie Anlegen und Verlinken von Personenlisten in der Profile Description

Vorgabe: TEI-All-Schema

Guidelines:

- Names
- The Profile Description

Übung 3 – Einfache XPath-Abfragen

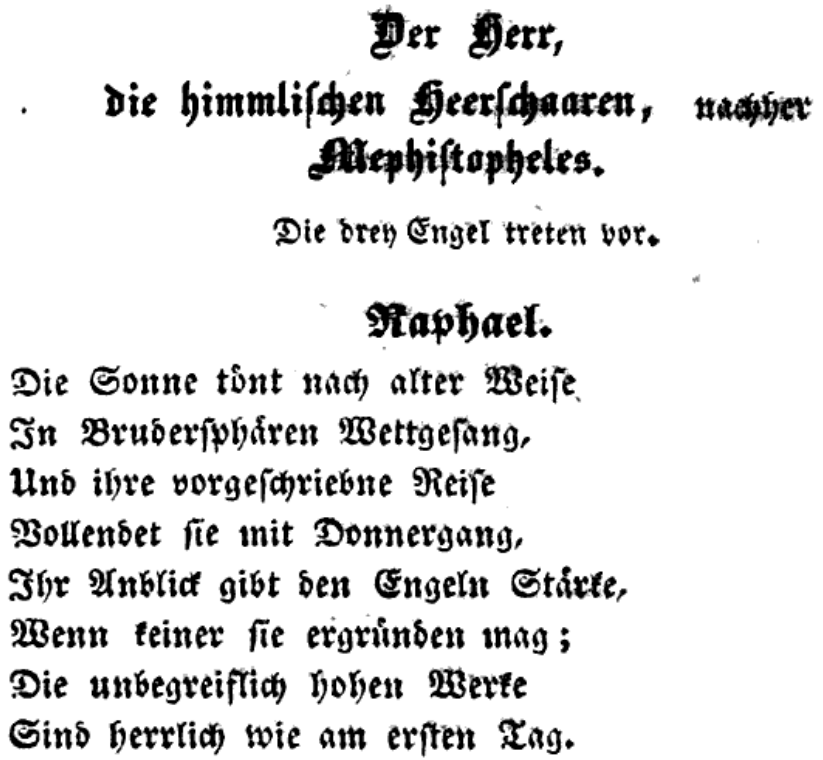
Ziel: Abfragen von Elementen innerhalb einer Datei mittels der oXygen XPath-Direkteingabe bzw. innerhalb mehrerer Datein mittels oXygens „Finden/Ersetzen in Dateien“.

XPath:

- `//element` – findet alle Elemente namens `element`
- `//element[contains(., 'suchwort')]` – findet alle `element` die `suchwort` enthalten

Übung 4 – Auszeichnen von Versen

Ziel: Analysieren der Vorlage, Erstellen der notwendigen Meta-Informationen im `teiHeader` und Erfassen der Versstruktur.



Der Herr,

die himmliſchen Heerſchaaren, nachher Mephiſtopheles.

Die drey Engel treten vor.

Raphael.

Die Sonne tönt nach alter Weiſe
In Bruderſphären Wettgeſang,
Und ihre vorgeſchriebne Reiſe
Vollendet ſie mit Donnergang,
Ihr Anblick gibt den Engeln Stärke,
Wenn keiner ſie ergründen mag;
Die unbegreiflich hohen Werke
Sind herrlich wie am erſten Tag.

Hinweise

Vorgabe: TEI-Drama-Schema

Guidelines:

- Performance Texts

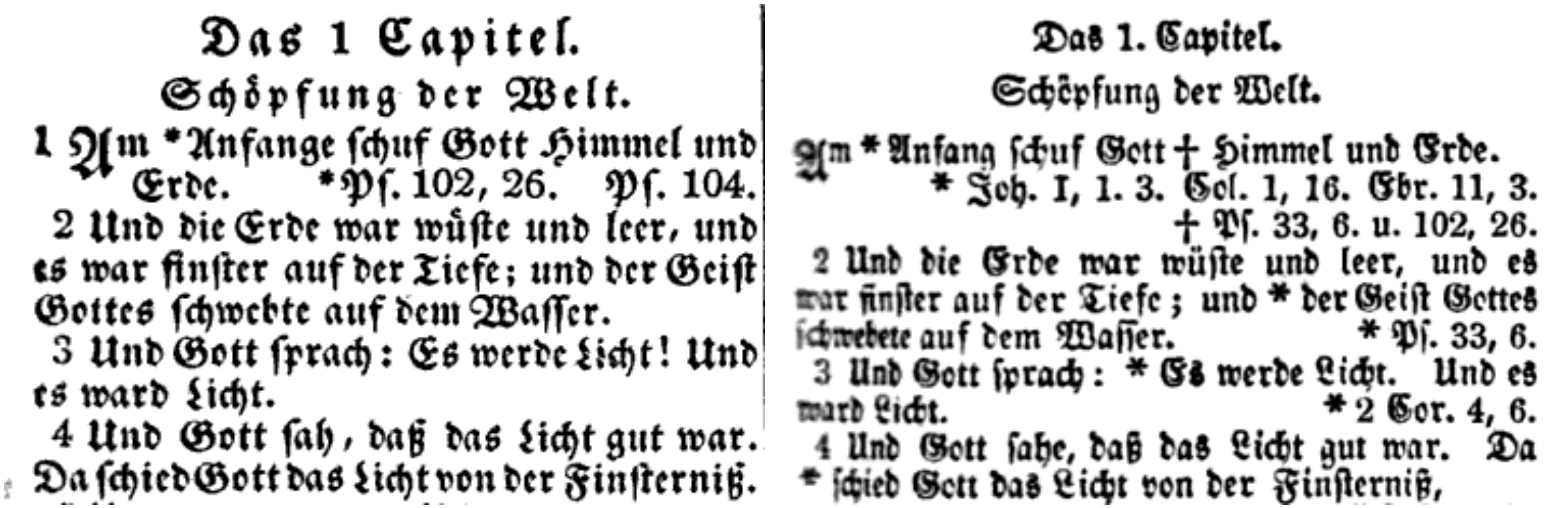
Übung 5 – Textkritischer Apparat

Ziel: Erfassen von mehreren Zeugen im `teiHeader`, Auszeichnen von Lesarten im `text`.

Vorgabe: TEI-All-Schema

Guidelines:

- Critical Apparatus



Das 1 Capitel.

Schöpfung der Welt.

1 Am *Anfange ſchuf Gott Himmel und Erde. *Pſ. 102, 26. Pſ. 104.

2 Und die Erde war wüſte und leer, und es war finſter auf der Tiefe; und der Geiſt Gottes ſchwebte auf dem Waſſer.

3 Und Gott ſprach: Es werde Licht! Und es ward Licht.

4 Und Gott ſah, daß das Licht gut war. Da ſchied Gott das Licht von der Finſterniß.

Das 1. Capitel.

Schöpfung der Welt.

Am * Anfang ſchuf Gott † Himmel und Erde.
* Joh. I, 1. 3. Col. 1, 16. Ebr. 11, 3.
† Pſ. 33, 6. u. 102, 26.

2 Und die Erde war wüſte und leer, und es war finſter auf der Tiefe; und * der Geiſt Gottes ſchwebete auf dem Waſſer. * Pſ. 33, 6.

3 Und Gott ſprach: * Es werde Licht. Und es ward Licht. * 2 Cor. 4, 6.

4 Und Gott ſahe, daß das Licht gut war. Da * ſchied Gott das Licht von der Finſterniß,

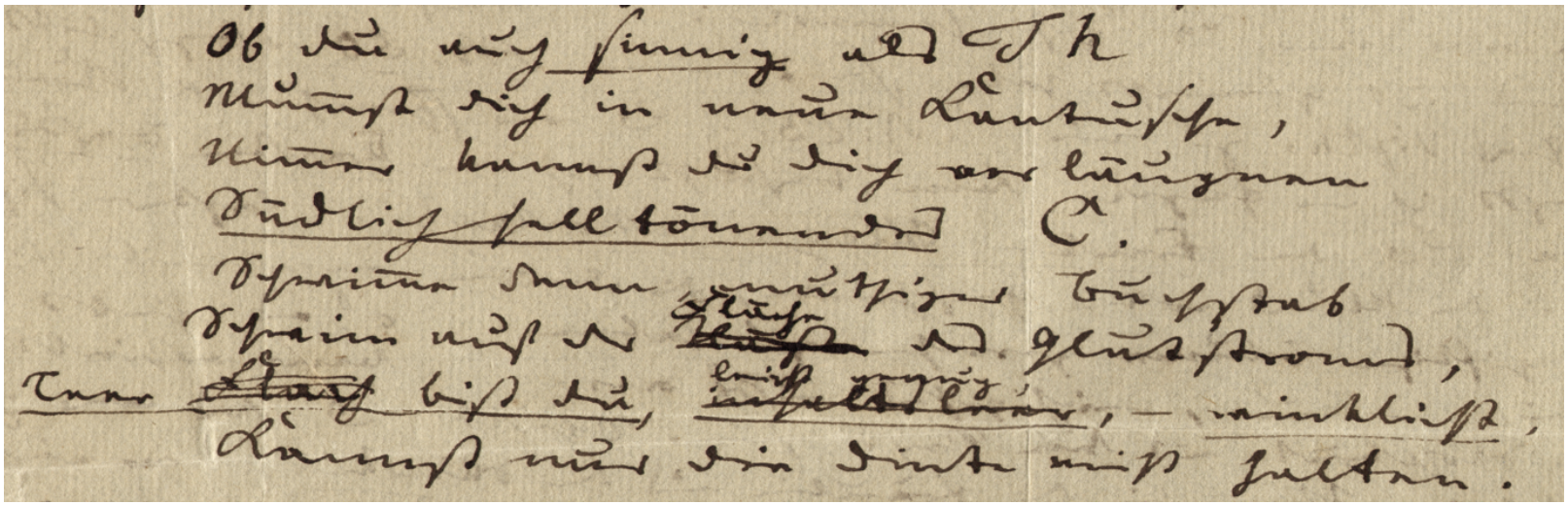
Übung 6 – Ersetzungen, Streichungen, Hinzufügungen

Ziel: Auszeichnen von textuellen Ersetzungen, Streichungen und Hinzufügungen

Vorgabe: TEI-All-Schema

Guidelines:

- Additions, Deletions, and Omissions
- Substitutions



XSLT

Extensible Stylesheet Language Transformation (XSLT)

- Tutorial auf w3schools
- TEI XSLT Stylesheets

TEI Maßschneiderung

„From the start, the TEI was intended to be used as a set of building blocks for creating a schema suitable for a particular project. This is in keeping with the TEI philosophy of providing a vocabulary for describing texts, not dictating precisely what those texts must contain or might have contained. This means that it is likely, not just possible, that you will want to have a tailored view of the TEI.“

Beispielkonfigurationen

Es gibt nicht das TEI-Schema, sondern ein bestimmtes Schema ist eine Auswahl aus den verfügbaren Modulen, wobei Elemente und Attribute aus diesen Modulen weiter modifiziert sein können.

tei_bare	TEI Absolutely Bare
teilite	TEI Lite
tei_corpus	TEI for Linguistic Corpora
tei_ms	TEI for Manuscript Description
tei_drama	TEI with Drama
tei_speech	TEI for Speech Representation
tei_odds	TEI for authoring ODD
tei_allPlus	TEI with maximal setup, plus external additions
tei_svg	TEI with SVG
tei_math	TEI with MathML
tei_xinclude	TEI with XInclude (experimental)
tei_dictionaries	TEI for Dictionaries (experimental)

Die TEI Module

Modulname	Kurzbeschreibung
analysis	Analysis and Interpretation
certainty	Certainty and Uncertainty
core	Common Core
corpus	Metadata for Language Corpora
dictionaries	Print Dictionaries
drama	Performance Texts
figures	Tables, Formulae, Figures
gaiji	Character and Glyph Documentation
header	Common Metadata
iso-fs	Feature Structures
linking	Linking, Segmentation, and Alignment
msdescription	Manuscript Description
namesdates	Names, Dates, People, and Places
nets	Graphs, Networks, and Trees
spoken	Transcribed Speech
tagdocs	Documentation Elements
tei	TEI Infrastructure
textcrit	Text Criticism
textstructure	Default Text Structure
transcr	Transcription of Primary Sources
verse	Verse

Modifizieren der Module

- Elemente entfernen
- Elemente umbenennen
- Attribute von existierenden Elementen entfernen, hinzufügen oder ändern
- Neue Elemente hinzufügen

Und wie??

ODD – „One Document does it all“

Eine ODD-Datei ist eine normale TEI-Datei, die die Verwendung von Modulen und Elementen spezifiziert, dabei gleichzeitig diese formal dokumentiert und Prosa-Dokumentation im Sinne von elektronischen Editionsrichtlinien enthält.

Übung 7 – Mein erstes ODD

Ziel: Ein Minimal-ODD erstellen und via Roma ein RelaxNG Schema (XML Syntax) generieren. Dieses mit der Undine XML-Datei verknüpfen.

Guidelines:

- Getting Started with P5 ODDs
- Roma Web Service

Übung 8 – Hinzufügen von Modulen

Ziel: Das Minimal-ODD erweitern bis die Undine Datei dagegen validiert.

Guidelines:

- Getting Started with P5 ODDs
- Roma Web Service

Übung 9 – Einschränken von Attributwerten

Ziel: Die ODD Datei modifizieren so dass für `hi@rend` nur noch „italic“ und „spaced“ erlaubt sind.

Guidelines:

- Getting Started with P5 ODDs
- Roma Web Service

Übung 10 – Entfernen von Elementen

Ziel: Mindestens `email`, `gloss` und `said` aus dem ODD entfernen.
Besser noch: Nicht genutzte Model Klassen ganz entfernen.

Guidelines:

- Getting Started with P5 ODDs
- Roma Web Service

Übung 11 – Hinzufügen von Elementen

Ziel: `characterName` dem Model `model.nameLike.agent` hinzufügen. Evtl. auch ein `workName` einbauen.

Guidelines:

- Getting Started with P5 ODDs
- Roma Web Service

Übung 12 – Maßschneiderung

Ziel: Ausgehend von tei_bare ein Schema für unser Faust-XML erstellen.

Guidelines:

- Getting Started with P5 ODDs
- Roma Web Service