TEI - Einführung

TEI = Text Encoding Initiative

- TEI „is a consortium which collectively develops and maintains a standard for the representation of texts in digital form”
- D.h. bei Interesse können Personen oder Organisationen Mitglied werden, an der Weiterentwicklung von Standards mitarbeiten, ‚Special Interest Groups‘ organisieren usw.
- Üblicherweise wird „TEI“ aber auch als Synonym für den Standard, den die TEI entwickelt, benutzt

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

TEI-Angebote

- Community:
  - Mailingliste tei-l
  - Feature requests (@Sourceforge)
  - Members Meetings
- Aktivitäten
  - Special Interest Groups (SIG)
- Tools
  - Source Code (Open Source)
  - Stylesheets
  - OxGarage

TEI Guidelines

- TEI-Standard ist eine Einschränkung der im Prinzip unendlichen Möglichkeiten von XML
- wichtigstes ‚Produkt' der TEI sind die ‚Guidelines' und formalisierte Schemata zur Validierung von XML-Dateien
- Fragen, die geklärt werden müssen:
  - Welche Tags und Attribute werden bereitgestellt?
  - Wie dürfen die Tags verschachtelt werden?
- Erste Version der Guidelines wurde 1988 entwickelt (SGML-basiert), derzeit Version P5 (proposal 5) aktuell

http://www.tei-c.org/Guidelines/P5/ (als PDF, epub, mobi)

TEI und ‚Customisations'

- 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)
- Module u.a.
  - core (Basiselemente)
  - header (Metadaten)
  - textstructure (grundlegende Textstrukturen)
  - msdescription (Handschriftenbeschreibungen)
  - gaiji (Sonderzeichen)

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

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

- (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

Übung

- In welchen Kapiteln werden Elemente
  - zur sprachwissenschaftlichen Auszeichnung erläutert?
  - für die Edition erläutert?
- In welchem Modul sind die Elemente <abbr>, <app>, <g>, <incipit>, <person> und <w> erläutert?
- Welcher Modellklasse gehören die Elemente <msDesc>, <persName>, <term> an?
- Welche Attribute enthalten die Attributklassen att.canonical und att.pointing?
- Wodurch unterscheiden sich die Datentypen data.word und data.text?

Repräsentation eines Dokuments

- Eine TEI-Datei repräsentiert ein „real world object“, durch
  - Metadaten (<teiHeader>, u.a. <msDesc>)
  - digitale Abbilder (<facsimile>)
  - Transkription/„Edition“ (<text>)
- Theoretisch kann ein nicht existierendes Objekt nicht beschrieben werden. (Problem: <msIdentifier>)
- Theoretisch können zur Zeit nur Handschriften als Objekte beschrieben werden. (Bezeichner: <msDesc>)

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
  - 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

Auszeichnungsstrukturen

- Wegen des Verbotes von XML, dass Hierarchien nicht überlappen dürfen, kann immer nur eine Hierarchie durch umschließende Klammern ausgezeichnet werden.
- Alternative Sichten auf das Dokument können nur durch sogenannte 'Milestone'-Elemente gekennzeichnet werden.

Strukturen

- Physikalische Einheiten
  - **<pb/>**, **<cb/>**, **<lb/>**, **<gb/>** → der „break“ beginnt die Einheit
- Strukturelle Texteinheiten
  - ** **, **<p>**, **<list>**, **<lg>**, **<index>**
- Semantische Texteinheiten
  - **<head>** (nicht: <title>!), **<fw>**, **<note>**, **<quote>**, **<term>**
  - **<ref>**, **<bibl>**, **<rs>**
- Entitäten
  - **<persName>**, **<orgName>**, **<placeName>**, **<name>**
  - **<date>**

Ressourcen

- Cheatsheet für Editionen (Marjorie Burghart)
- TEI by Example
- Wolfenbütteler Dokumentation: http://diglib.hab.de/rules/documentation/
  - mit Beispiel-Headern für Projekte, Handschriftenbeschreibungen

- Aktuelle Schemata
  - Für Handschriftenbeschreibungen http://diglib.hab.de/rules/schema/ER/v0.4/europeana-regia.xsd
  - Für Editionen http://diglib.hab.de/rules/schema/tei/P5/v2.0.2/tei-p5-transcr.xsd