

Textkodierung mit TEI — Aufgabe 03a: Erste Schritte der Textkodierung mit XML/TEI (Beispiel: Postkarte)

Christof Schöch

DARIAH-DE

Lizenz: Creative Commons Attribution 4.0 (CC-BY).

Das vorliegende Dokument wurde für DARIAH-DE neu erstellt. Das vorliegende Dokument beruht teilweise auf den folgenden Materialien: James Cummings, Renée Baalen, Ylva Berglund-Prytz: "An Introduction to XML and the Text Encoding Initiative", Digital.Humanities@Oxford Summer School 2012, University of Oxford, 2012, <http://digital.humanities.ox.ac.uk/dhoxss/2012/>, CC-BY-SA.

Überblick

Hinweise: Diese Aufgabe setzt die Kenntnis von Referat 03, "TEI und das Modul core", voraus. Außerdem setzt diese Aufgabe voraus, dass Sie jEdit mit dem Plugin "XML" auf Ihrem Computer installiert haben. (Für die Installation von jEdit und der notwendigen Plugins, siehe separate Installationshinweise.)

1. Lernziele

Nach dieser Aufgabe sollten Sie Folgendes erreicht haben:

- * Die für die Kodierung relevante Strukturen in Dokumenten erkennen
- * Einige relevante Funktionen des XML-fähigen Editors "jEdit" kennen
- * Eigene Erfahrung mit der Kodierung von Textstrukturen gemacht haben
- * Erste Elemente des Standards der Text Encoding Initiative verwenden können

2. Materialien

- * Zwei Bilddateien: digitale Faksimiles einer Postkarte `Postkarte_recto.jpg` und `Postkarte_verso.jpg`
- * Eine Textdatei mit der Transkription des auf der Postkarte befindlichen Textes `Postkarte_Transkription.txt`
- * Eine TEI-Datei, die fast keinen Inhalt enthält `Postkarte_Start.xml`
- * Ein Stylesheet zur Visualisierung der Kodierung `Postkarte_Stylesheet.css`
- * Eine Schema-Datei zur Überprüfung der Kodierung `TEI-Schema_allgemein.rng`
- * Installationshinweise für jEdit und das Plugin "XML" `Installation_jEdit.pdf`

3. Vorbereitung

- * Den Ordner mit den Materialien herunterladen, entpacken und öffnen

- * jEdit starten

Aufgabenstellung

1. Analyse der Struktur der Postkarte

Beginnen Sie mit einer Analyse der Struktur der Postkarte und notieren Sie Ihre Antworten auf die folgenden Fragen in dem freien Bereich auf dem Blatt:

- * Welche größeren oder kleineren Teile und Bereiche des Dokuments, welche sonstigen Merkmale und Informationen, kann man unterscheiden?
- * Wie unterscheiden sich die Teile/Bereiche/Merkmale? Wie kann man sie in Bezug zueinander setzen (bspw. hierarchisch, als Sequenz, etc.)?
- * Welche Bereiche/Teile/Merkmale würden Sie im Rahmen einer (digitalen) Textedition explizit festhalten wollen, welche können implizit oder unberücksichtigt bleiben?

2. Identifikation der passenden TEI-Elemente für die Grobstruktur

Im vorausgegangenen Input-Referat haben Sie eine ganze Reihe grundlegender TEI-Elemente aus den Modulen 2, 3 und 4 der Guidelines der Text Encoding Initiative kennengelernt.

- * Legen Sie nun zunächst für 4-6 größere Bereiche, die Sie unterscheiden wollen fest, welche Elemente aus TEI Sie jeweils dafür verwenden wollen. (Ein Beispiel ist bereits eingetragen.)

Vorderseite: `div` evtl. auch genauer: `<div type="Vorderseite">`

3. Transkription öffnen und bearbeiten

Unter den Materialien befindet sich neben den Bild-Dateien folgende weitere Datei: `Postkarte_Transkription.txt`. Es handelt sich um eine Textdatei, die nur eine ganz einfache, lineare Transkription des Textes enthält, den man auf der Postkarte lesen kann.

- * Öffnen Sie die Datei mit dem Editor jEdit (Rechtsklick auf die Datei, „Öffnen mit“, jEdit auswählen. Alternativ: jEdit starten, dann File > Open... > zur Datei navigieren, auf „Open“ klicken).
- * Vergleichen Sie die Transkription mit der Postkarte: Welche Informationen oder Aspekte sind bereits implizit oder explizit repräsentiert? Welche Informationen über die Postkarte fehlen? Notieren Sie einige Aspekte, die Ihnen auffallen.

... ..

4. Schema mit der Kodierung verbinden

Der Ordner enthält außerdem eine Datei, die eine fast „leere“ TEI-Datei darstellt `Postkarte_Start.xml`. Öffnen Sie auch diese Datei und schauen Sie sich die Grundstruktur zunächst einmal ein bisschen an. Welche beiden übergeordneten Bereiche kann man unterscheiden?

- * Verbinden Sie jetzt ein Schema mit der leeren TEI-Datei. Plugins > XML > Set schema Datei "TEI-Schema_allgemein.rng" auswählen, Mit „OK“ bestätigen
- * Speichern Sie die gerade geöffnete Datei unter einem neuen Namen ab, bspw. Postkarte_Kodierung1.xml (File > Save as > Namen vergeben, Save).
- * Stellen Sie sicher, dass nach dem Abspeichern keine Fehlermeldung erscheint. (Nur dann ist die Datei in dieser Form valide.)

Hinweis: Die Schema-Datei legt fest, welche Elemente in der Kodierung erlaubt sind, und an welcher Stelle sie jeweils vorkommen dürfen. Das heißt, das Schema steuert oder kontrolliert die Kodierung, um sicherzustellen, dass diese stets valides TEI ist.

5. Von der Transkription zur Kodierung

Jetzt geht es darum, die Transkription (wie sie in der ersten Datei vorliegt) in die TEI-Datei einzufügen und mit einigen TEI-Elementen zu strukturieren. Hierfür können Sie die TEI-Elemente nutzen, die Sie in Abschnitt 2 der Übung festgelegt haben.

- * Markieren Sie zunächst die gesamte Transkription und kopieren Sie sie an der richtigen Stelle in die in die TEI-Datei hinein. Die richtige Stelle ist der Bereich zwischen dem öffnenden und dem schließenden `p`-Element innerhalb des `div`-Elements, das sich im `body`-Element befindet. Den dort bereits vorhanden Platzhaltertext können Sie löschen.
- * Speichern Sie die Datei und stellen Sie sicher, dass keine Fehlermeldung erscheint, d.h. dass die Datei weiterhin valide ist.
- * Jetzt umgeben Sie die gewünschten größeren Bereiche der Transkription mit passenden Tags (d.h. Anfangs- und End-Tags). Das Hinzufügen von Markup bedeutet auch, dass die Transkription nun in eine Kodierung verwandelt wird.

Beachten Sie, dass die unmittelbar vor und unmittelbar nach der Transkription bereits vorhandenen `div`- und `p`-Tags eventuell nicht angemessen sind; modifizieren Sie diese Tags entsprechend. Zur Erinnerung: Es gibt öffnende und schließende Tags: `<beispiel>` (öffnend) und `</beispiel>` (schließend). Gemeinsam mit dem umschlossenen Text bildet ein öffnendes und ein schließendes Tag jeweils ein Element. Beachten Sie außerdem, dass Elemente hierarchisch angeordnet werden müssen, d.h. zwar verschachtelt werden können, aber nicht überlappen dürfen.

- * Beobachten Sie, während Sie Elemente hinzufügen, wie sich jEdit verhält. Was macht jEdit automatisch? Werden Tags korrekt visuell hervorgehoben?
- * Speichern Sie die Datei zwischendurch immer wieder (Symbol links oben in der Symbolleiste, mit dem kleinen blauen Pfeil). Werden Sie gewarnt, wenn Sie fehlerhaftes XML schreiben? Nutzen Sie ggfs. die Hinweise, die jEdit gibt, um ihr XML-Datei zu korrigieren.

Hinweis: Man kann die Kodierung automatisch etwas übersichtlicher gestalten, indem man eine automatische Einrückung der Elemente veranlasst. Klicken Sie hierzu in der Menüleiste auf "Plugins > XML > Indent XML".

6. Hinzufügen weiterer Kodierungs-Informationen

Bisher haben Sie nur eine relativ grobe Vorstrukturierung der Transkription durch TEI-Elemente vorgenommen. Jetzt sollten Sie die einzelnen Bereiche des Textes genauer kodieren.

- * Speichern Sie die Datei noch einmal unter einem neuen Namen ab, bspw.

`Postkarte_Meine-Kodierung2.xml`.

- * Machen Sie erneut eine Liste der Abschnitte, die sie explizit benennen möchten und wählen Sie dafür jeweils passende TEI-Elemente aus. Klammern Sie den Bereich der Postkarte mit den Briefmarken zunächst aus.

Hinweis: Schauen Sie im Input-Referat zu den Modulen 2-4 nach, wenn Sie sich nicht mehr sicher sind. Oder lesen Sie in den TEI-Guidelines, Kapitel 2-4 nach; die Guidelines sind verfügbar unter: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.

7. Informationen im Header

- * Ersetzen Sie jetzt noch die Platzhalter-Texte im `teiHeader` durch Angaben, die im Kontext ihrer digitalen Kodierung der Postkarte sinnvoll erscheinen. (Die Postkarte stammt aus der Postkartensammlung von Zeno.org.)

8. Optional: Details ergänzen mit Attributen

Für einige wichtige Informationen haben Sie vielleicht kein passendes oder ausreichend präzises Element unter den TEI-Elementen gefunden. Kein Grund zum Verzweifeln! Mit dem Mechanismus der Attribute kann man zusätzliche Informationen strukturiert festhalten, ohne das Element-Inventar der TEI zu erweitern. Ein klassisches Beispiel ist das Element `hi` (highlighting, also Hervorhebung). Es sagt an sich nichts darüber aus, auf welche Weise der Text hervorgehoben wurde. Hierfür gibt es das Attribut "rend", das Werte wie "kursiv", "unterstrichen" oder "gras-let-large" annehmen kann (die Werte sind nicht vordefiniert).

- * Ergänzen Sie das Element `hi` um das Attribut "rend" mit einem geeigneten Wert. Prüfen Sie, ob Ihr Dokument weiterhin valide ist.

9. Ende der Aufgabe

Wenn Sie vorerst zufrieden sind und ihr Dokument keine Fehler enthält, sind Sie fast am Ende dieser Aufgabe angekommen.

- * Gemeinsam vergleichen wir jetzt mehrere Ihrer Kodierungen um herauszufinden, inwiefern Sie sich einig sind und welche Aspekte des Dokuments problematisch sind.

- * Alternativ können wir gemeinsam eine Beispielkodierung anschauen.

- * Und wir können uns anschauen, wie die Kodierung nun mit einem Stylesheet verknüpft werden kann, dass sie in einem Browser korrekt angezeigt wird.