TEI und TUSTEP
Auszeichnungsgrammtik & Universaltool für Germanisten

Matthias Schneider

Universität Trier

Dezember 2012

Gliederung[1]

1. Einführung in TEI als XML-Grammatik
2. Anwendungsbeispiel: TEI-Header für eine Dissertation
3. TUSTEP als Textverarbeitungswerkzeug

[1] Layout: MS LaTeX, Beamerklasse.

Zum Rahmenverständnis

Grundsätzliches

Die *Text Encoding Initiative* (TEI) stellt ein Kompendium an Regeln zur Auszeichnung von XML-Daten dar.

$\longrightarrow$ Regelwerk

Das *Tübinger System von Textverarbeitungsprogrammen* (TUSTEP) besteht aus mehreren Programmen zur wissenschaftlichen Textverarbeitung. Hiermit können u.a. strukturierte Daten (z.B. SGML oder XML) gemäß TEI erzeugt werden.

$\longrightarrow$ Verarbeitungs- und Analysetool

Thesen

1. Die Verarbeitung von großen Mengen an (Text-) Daten bedingt die Verwendung möglichst performaner automatischer Auszeichnungsinstanzen.
2. TEI kann als mächtige XML-Auszeichnungsgrammatik für weitgehend alle Formen von Texten genutzt werden.
3. TUSTEP bietet als System von Textverarbeitungsprogrammen zahlreiche out of the box-Lösung sowie modulare Erweiterungsmöglichkeiten für geisteswissenschaftliche Aufgabenstellungen.

XML – Vorteile

„Eine XML-codierte Datenbasis[2]

[2] Quelle: http://kompetenzzentrum.uni-trier.de/de/arbeitsfelder/standardisierte-auszeichnung/

[3] Weiterhin zu erwähnen: e-Print, Satz, Datenbanken, wissenschaftliche (quantitative) Analysen.

XML – Vorteile

„Eine XML-codierte Datenbasis[2]

- steht langfristig und plattformunabhängig zur Verfügung

[2] Quelle: http://kompetenzzentrum.uni-trier.de/de/arbeitsfelder/standardisierte-auszeichnung/

[3] Weiterhin zu erwähnen: e-Print, Satz, Datenbanken, wissenschaftliche (quantitative) Analysen.

XML – Vorteile

„Eine XML-codierte Datenbasis[2]

- steht langfristig und plattformunabhängig zur Verfügung
- erlaubt den gezielten Zugriff auf alle explizit gekennzeichneten Textelemente

[2] Quelle: http://kompetenzzentrum.uni-trier.de/de/arbeitsfelder/standardisierte-auszeichnung/

[3] Weiterhin zu erwähnen: e-Print, Satz, Datenbanken, wissenschaftliche (quantitative) Analysen.

XML – Vorteile

„Eine XML-codierte Datenbasis[2]

- steht langfristig und plattformunabhängig zur Verfügung
- erlaubt den gezielten Zugriff auf alle explizit gekennzeichneten Textelemente
- ermöglicht die Anreicherung eines Textes mit zusätzlichen Informationen

[2] Quelle: http://kompetenzzentrum.uni-trier.de/de/arbeitsfelder/standardisierte-auszeichnung/

[3] Weiterhin zu erwähnen: e-Print, Satz, Datenbanken, wissenschaftliche (quantitative) Analysen.

XML – Vorteile

„Eine XML-codierte Datenbasis[2]

- steht langfristig und plattformunabhängig zur Verfügung
- erlaubt den gezielten Zugriff auf alle explizit gekennzeichneten Textelemente
- ermöglicht die Anreicherung eines Textes mit zusätzlichen Informationen
- birgt durch die Internationalisierung ein hohes Vernetzungspotenzial

[2] Quelle: http://kompetenzzentrum.uni-trier.de/de/arbeitsfelder/standardisierte-auszeichnung/

[3] Weiterhin zu erwähnen: e-Print, Satz, Datenbanken, wissenschaftliche (quantitative) Analysen.

XML – Vorteile

„Eine XML-codierte Datenbasis[2]

- steht langfristig und plattformunabhängig zur Verfügung
- erlaubt den gezielten Zugriff auf alle explizit gekennzeichneten Textelemente
- ermöglicht die Anreicherung eines Textes mit zusätzlichen Informationen
- birgt durch die Internationalisierung ein hohes Vernetzungspotenzial
- ist Ausgangsplattform für die unterschiedlichen Publikationsmedien Buch, Internet und CD-ROM.[3]“

---

[2] Quelle: http://kompetenzzentrum.uni-trier.de/de/arbeitsfelder/standardisierte-auszeichnung/

[3] Weiterhin zu erwähnen: e-Print, Satz, Datenbanken, wissenschaftliche (quantitative) Analysen.

TEI – eine Grammatik für XML

TEI = Text Encoding Initiative

Das Ziel der TEI besteht in der Normierung von XML-Auszeichnungen zum Zwecke der größeren Vergleichbarkeit und langfristigen Verfügbarkeit von Daten.

TEI – eine Grammatik für XML

TEI = Text Encoding Initiative
Das Ziel der TEI besteht in der Normierung von XML-Auszeichnungen zum Zwecke der größeren Vergleichbarkeit und langfristigen Verfügbarkeit von Daten.

- Standard für Datenauszeichnungen
- globale Verbreitung
- Vielzahl von Regeln verfügbar (Korpora, Romane, Handschriften, Lexika, Dramen)
- Austauschbarkeit von Daten (Interdisziplinarität!)
- Gewährleistung von Verarbeitungsfähigkeit auch i.S. einer Langzeitarchivierung

TEI – Rahmendaten

- Entstanden 1987 als Initiative von Philologen
- Ziel: Standardisierung von SGML[4]
- Mittel: hardware- und softwareunabhängige Codierung von Texten
- Unabhängigkeit von Publikationsformen (ein(!) Quelltext, dutzende Publikationsarten)

[4] Standard Generalized Markup Language. ISO-Standard (8859) zur Definition von Textauszeichnungssprachen (=Metasprache) mit XML als Teilmenge.

Auf der nächsten Folie folgt ein Beispiel für einen TEI-Header einer Dissertation.

```
<?xml version="1.0" encoding="UTF−8" xml:lang="ger"?>
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>TEI und TUSTEP. Grammatik und Werkzeug fuer den digitalen Germanisten</title>
   <author>Karl−Heinz Mayer</author>
  <sourceDesc>
   <bibl>Erste, leicht korrigierte Fassung der Dissertation des Autors.</bibl>
   </sourceDesc>
 </fileDesc>
 <publicationStmt>
   <pubPlace>Trier</pubPlace>
   <date>2012</date>
 </publicationStmt>
 <noteStmt>
  <note>Die Arbeit wird eingereicht zur Erlangung eines Doktor phil.</note>
 <encodingDesc>
  <projectDesc>
   <p>Die Dissertation fokussiert die fruchtbare Verbindung von TUSTEP und TEI.</p>
  </projectDesc> </encodingDesc>
 <revisionDesc>
  <list><item><date when="2012−06−29">29. April 2012</date>
  <p>Letze Erweiterung hinzugefuegt.</p></item>
  </list>
 </revisionDesc>
</teiHeader>
```

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China
- Transkription in MS-Word (RTF)

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China
- Transkription in MS-Word (RTF)
- Korrektur von Fehlern/Unsicherheiten bei der Transkription

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China
- Transkription in MS-Word (RTF)
- Korrektur von Fehlern/Unsicherheiten bei der Transkription
- `#*IMPORT` in TUSTEP

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China
- Transkription in MS-Word (RTF)
- Korrektur von Fehlern/Unsicherheiten bei der Transkription
- `#*IMPORT` in TUSTEP
- Auszeichnung der Daten mittels `#KOPIERE`

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China
- Transkription in MS-Word (RTF)
- Korrektur von Fehlern/Unsicherheiten bei der Transkription
- `#*IMPORT` in TUSTEP
- Auszeichnung der Daten mittels `#KOPIERE`
- Umwandlung in XML

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China
- Transkription in MS-Word (RTF)
- Korrektur von Fehlern/Unsicherheiten bei der Transkription
- `#*IMPORT` in TUSTEP
- Auszeichnung der Daten mittels `#KOPIERE`
- Umwandlung in XML
- Prüfung der Daten auf Wohlgeformheit

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China
- Transkription in MS-Word (RTF)
- Korrektur von Fehlern/Unsicherheiten bei der Transkription
- `#*IMPORT` in TUSTEP
- Auszeichnung der Daten mittels `#KOPIERE`
- Umwandlung in XML
- Prüfung der Daten auf Wohlgeformheit
- Validierung der Daten gemäß TEI-Schema

Verbindung von TEI-XML und TUSTEP in der Praxis

Digitalisierung und Publikation

- Digitalisierung = Scannen
- Transkription in China
- Transkription in MS-Word (RTF)
- Korrektur von Fehlern/Unsicherheiten bei der Transkription
- `#*IMPORT` in TUSTEP
- Auszeichnung der Daten mittels `#KOPIERE`
- Umwandlung in XML
- Prüfung der Daten auf Wohlgeformheit
- Validierung der Daten gemäß TEI-Schema
- Publikation (Web, CD-ROM, Buch.....)

TUSTEP

Tübinger System von Textverarbeitungsprogrammen

zur quantitativen Textanalyse mit TUSTEP:
„Kaum ein Frageansatz ließ sich *nicht* innerhalb kürzester Frist programmtechnisch realisieren, wo es mit anderen Mitteln – aufgrund der meist schwierigen Formalisierung von Sprache – wochen- und monatelanger Entwicklungsarbeit bedurft hätte.“

Trauth 2002: S. 317.

TUSTEP-Anwendungsmöglichkeiten: eine Auswahl

Fokus: Historische Germanistik

- automatische XML-Annotation von Texten (s. Bsp.)

[5] Formate u.a.: RTF, DOCX, XML, PS, TXT.

TUSTEP-Anwendungsmöglichkeiten: eine Auswahl

Fokus: Historische Germanistik

- automatische XML-Annotation von Texten (s. Bsp.)
- quantitative Textanalyse

[5] Formate u.a.: RTF, DOCX, XML, PS, TXT.

TUSTEP-Anwendungsmöglichkeiten: eine Auswahl

Fokus: Historische Germanistik

- automatische XML-Annotation von Texten (s. Bsp.)
- quantitative Textanalyse
- Transkription von Texten mit exakter Sonderzeichencodierung (z.B. alt-/mittelhochdeutsch)

[5] Formate u.a.: RTF, DOCX, XML, PS, TXT.

TUSTEP-Anwendungsmöglichkeiten: eine Auswahl

Fokus: Historische Germanistik

- automatische XML-Annotation von Texten (s. Bsp.)
- quantitative Textanalyse
- Transkription von Texten mit exakter Sonderzeichencodierung (z.B. alt-/mittelhochdeutsch)
- Satz von Hausarbeiten, B.A.-/M.A.-Arbeiten, Dissertationen, Editionen...

[5] Formate u.a.: RTF, DOCX, XML, PS, TXT.

TUSTEP-Anwendungsmöglichkeiten: eine Auswahl

Fokus: Historische Germanistik

- automatische XML-Annotation von Texten (s. Bsp.)
- quantitative Textanalyse
- Transkription von Texten mit exakter Sonderzeichencodierung (z.B. alt-/mittelhochdeutsch)
- Satz von Hausarbeiten, B.A.-/M.A.-Arbeiten, Dissertationen, Editionen...
- Konvertierung[5] und Verarbeitung von Texten,
  Bsp.: Register erstellen, Indizierung

[5] Formate u.a.: RTF, DOCX, XML, PS, TXT.

TUSTEP-Anwendungsmöglichkeiten: eine Auswahl

Fokus: Historische Germanistik

- automatische XML-Annotation von Texten (s. Bsp.)
- quantitative Textanalyse
- Transkription von Texten mit exakter Sonderzeichencodierung (z.B. alt-/mittelhochdeutsch)
- Satz von Hausarbeiten, B.A.-/M.A.-Arbeiten, Dissertationen, Editionen...
- Konvertierung[5] und Verarbeitung von Texten,
  Bsp.: Register erstellen, Indizierung
- Erstellung von Synopsen (Zeilen-/Worts.)

[5] Formate u.a.: RTF, DOCX, XML, PS, TXT.

#KOPIERE

Kernprogramm zur Textmanipulation

Funktionsweise

Quelldatei $\rightarrow$ $Zwischendatei_1$ $\rightarrow$ $Zwischendatei_n$ $\rightarrow$ Zieldatei

Während des Kopiervorgangs können die Daten manipuliert werden:

#KOPIERE

Kernprogramm zur Textmanipulation

Funktionsweise

Quelldatei $\rightarrow$ *Zwischendatei*$_1$ $\rightarrow$ *Zwischendatei*$_n$ $\rightarrow$ Zieldatei

Während des Kopiervorgangs können die Daten manipuliert werden:

Austauschen (`#/+` in `<hi rend = "italics" >`)
Durchnummerieren (Zeilen, Seiten, Sätze, Wörter)
Vergleich von Daten
Ergänzen von Textteilen nach Markierungen
Bearbeitung von Textteilen in bestimmten Abschnitten (Absätze, Zeilen, Seiten, Kapitel)

#KOPIERE

Kernprogramm zur Textmanipulation

Funktionsweise

Quelldatei $\rightarrow$ $Zwischendatei_1$ $\rightarrow$ $Zwischendatei_n$ $\rightarrow$ Zieldatei

Während des Kopiervorgangs können die Daten manipuliert werden:

Austauschen (`#/+` in `<hi rend = "italics" >`)
Durchnummerieren (Zeilen, Seiten, Sätze, Wörter)
Vergleich von Daten
Ergänzen von Textteilen nach Markierungen
Bearbeitung von Textteilen in bestimmten Abschnitten (Absätze, Zeilen, Seiten, Kapitel)

Wichtig (!): **Die Ursprungsdaten werden nicht verändert und bleiben damit frei von Verarbeitungsfehlern u.ä.**
Wichtiger Vorteil gegenüber anderen Verarbeitungsarten (z.B. Suche-Ersetze im Editor.)

TUSTEP: Vorteile

- keine Korrution der Ursprungsdaten während der Verarbeitung
- keine Korruption von Daten durch Versionsänderungen der Software (s. Seiten-, Zeilenumbruch bei MS-Word)
- einsteigerfreundliche Programmteile (`*SATZ`, `*IMPORT`, `*EXPORT`)
- sehr leistungsfähiges *pattern matching* (besser als *regular expressions*?!)

TUSTEP: Vorteile II

- ständige Weiterentwicklung (s. `#KOPIERE`, `TUSCRIPT`, `TXSTEP`)
- Performanz durch Hardware-nahe Programmierung (C, Fortran)
- Verarbeitung von Datenmengen im Gigabyte-Bereich problemlos möglich
- kleine, aber sehr engagierte Community
- reaktionsfreudige Entwicklergruppe (!)
- OpenSource, kostenlos

TUSTEP: Nachteile

- ungewohnte GUI (1990er-Jahre DOS-Charme)
- kein „Zusammenklicken“ von Funktionen, sondern kommandobasiertes Arbeiten
- sperriges Handbuch
- teilweise ungewohnte Funktionalität der Programmiersprachen
- steile Lernkurve

Fazit

Mit der TEI steht eine wichtige Standardisierungsinstanz für XML-Dokumente zur Verfügung, die genutzt werden sollte; insbesondere wenn Daten ausgetauscht und langfristige (sinnvoll) nutzbar sein sollen.

TUSTEP kann als mächtiges Werkzeug zur Textverarbeitung in den sehr unterschiedlichen Bereichen genutzt werden.
Es bietet eine out of the box-Lösung für die meisten Probleme und Herausforderungen für Geisteswissenschaftler. Spezielle Anforderungen kann mit begrenztem Aufwand begegnet werden, ohne dass kostenintensive und proprietäre Software zum Einsatz kommen muss.

Informationsquellen: TUSTEP

- TUSTEP-Wiki
- International TUSTEP User Group (ITUG)
- TUSTEP-Homepage: Download, Informationen
- Trier Center for Digital Humanities: Projekte, Anwendungsbeispiele, Publikationen zu TEI/TUSTEP

Zur quantitativen Textanalyse mit TUSTEP s.

- Michael TRAUTH, Quantifizierende Textanalyse. Mit der Hilfe des Computers auf der Suche nach dem anonymen Autor, in: Historische Sozialforschung Jg. 17 Heft 1, 1992, S. 133-141.
  sowie
- DERS., *Caesar incertus auctor*. Ein quantifizierendes Wort zur Kritik von Verfasserfragen in lateinischen Texten, in: Jürgen Jaehrling, Uwe Mewes und Erika Timm, Rollwägenbüchlein. Festschrift für Walter Röll zum 65. Geburtstag, Tübingen 2002, S. 313-335.

Beispielprojekte TUSTEP:

- Cusanus-Portal des Cusanus-Instituts Trier
- Digitale Gesamtausgabe: „Jeremias Gotthelf: Historisch-kritische Edition“
- Wörterbuchnetz, Kompetenzzentrum
- Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm, Kompetenzzentrum
- Digitalisierung der Ökonomischen Enzyklopädie von Krünitz an der UB Trier
- Goethe-Wörterbuch, Kompetenzzentrum
- „Dat Nuwe Boych.“ Digitale Neuedition, elektronische Publikation und Informationsnetzwerk einer Kölner Stadtchronik
- Satz der kritischen Editio Coloniensis von Albertus Magnus, Albertus-Magnus-Institut Köln

Informationsquellen: TEI

- TEI
- TEI P5-Guidelines
- TEI-Tutorials
- TEI-Bibliographie
- TEI-Lernhilfen
- Einführung in XML von der TEI

Informationsquellen: XML et al.

- W3C Consortium
- Youtube XML-Tutorial
- Unicode Character Code Charts
- SELF HTML – Gute Übersichtswebsite auf deutsch für HTML, CSS, XML, Javascript, PHP, Perl
- CSS Zengarden – Beispiele für gelungene Zusammenstellung und XML/XHTML und CSS
- Daniel Cohen/Roy Rosenzweig: Digital History
- Roy Rosenzweig – Center for History and New Media

Rückfragen zur Präsentation, Projekten ??

Matthias Schneider

s3msschn@uni-trier.de
mail@m-schneider.eu http://www.m-schneider.eu
skype: matz.tru