Konvertierung von bestehenden Dokumenten in das DTA-Basisformat

Frank Wiegand, Deutsches Textarchiv

4. DTA-Workshop – Berlin, 7. Juli 2014

Szenarien

- Text
- Textverarbeitung (MS Word, …)
  - mit Templates
  - ohne Templates
- XML
  - TEI
  - anderes Format (SGML, HTML, …)
- PDF, Wikisource, …

Strategie

- alles ist konvertierbar
- Abwägung zwischen Automatismen und Handarbeit
- kenne die Tools (Recherche, Experten fragen)
- Workflow dokumentieren, Backup nicht vergessen
- nicht immer lohnt sich die Konvertierung

Fallbeispiel 1: Text

- einfachstes Dokumentenformat
- benötigt keine Expertentools
- Haken: Dokumentation und Verlässlichkeit

Ein erbärmlich,
Tyrannisch, wehemüthig, grausam, und barmhertziges, traurig,
erschröcklich, lustig, entsetzlich, und seltsames, ungeheuer, mörderisch,
lamentabel, kümmerlich und unvergnügtes,

Jammer-Elend-Noth-Freuden-und Todes-Lied,

In einer Historisch, und ergötzlich-betrübten Mord-Geschichte, sehr hertzbrechend und lächerlich vorgestellet:

Was sich bey dem

berühmt-und herrlichen Campement in Sachsen,

nicht weit von Mühlberg,
zwischen einem Koch und einer schwartzen Henne
Verwunderungs-würdig, furchtsam und anmuthig zugetragen.

Fallbeispiel 1: Text

Ein erbärmlich,⏎
Tyrannis#ch, wehemüthig, graus#am, und barmhertziges, traurig,⏎
ers#chröcklich, lus#tig, ents#etzlich, und s#elts#ames, ungeheuer, mörderis#ch,⏎
<lamentabel>, kümmerlich und unvergnügtes,⏎
+Jammer- Elend- Noth- Freuden- und⏎
++{T}odes-{L}ied,⏎
Jn einer his#toris#ch, und ergötzlich-betrübten Mord-Ges#chichte, s#ehr⏎
hertzbrechend und lächerlich vorges#tellet:⏎
Was s#ich bey dem⏎
+berühmt- und herrlichen <Campement> in Sachs#en,⏎
nicht weit von Mühlberg,⏎
zwis#chen einem Koch und einer s#chwartzen Henne⏎
Verwunderungs-würdig, furchts#am und anmuthig zugetragen.⏎

Fallbeispiel 1: Text

- Regeln definieren:
  - ä → `a&#x0364;`
  - ö → `o&#x0364;`
  - ü → `u&#x0364;`
  - s# → `&#x017F;`
  - ⏎ → `<lb/>`
  - <...> → `<hi rendition="#aq">...</hi>`
  - {...} → `<hi rendition="#in">...</hi>`

Fallbeispiel 1: Text

- Umsetzung:
  - Suchen+Ersetzen
  - reguläre Ausdrücke
  - im Editor oder mit Skript
- Vorsicht! Es gibt keine Standards für diese Formate (Dokumentation und Verlässlichkeit)

Reguläre Ausdrücke

- „variables“ Suchen+Ersetzen
  - nützlich, wenn man mehr als nur feste Zeichenketten finden und ersetzen möchte
- eigene Sprache
  - viele Tutorials im Netz
  - Friedl: Reguläre Ausdrücke, 3. Aufl., O'Reilly.

Reguläre Ausdrücke

.	beliebiges Zeichen
X?	0 oder 1 von X
X*	0 oder mehr von X
X+	1 oder mehr von X
X{n}	n Vorkommen von X
X{n,}	mindestens n Vorkommen von X
X{n,m}	n bis m Vorkommen von X
^ und $	Zeilenanfang, Zeilenende
[XYZ]	X oder Y oder Z
\d und \D	Ziffer: [012345678] → [0-9]
\w und \W	Wortzeichen: [a-zA-Z0-9_äöü]
\s und \S	Leerraum: [\t\n\r\f\v]
\. \{ \[\\	. { [\

Reguläre Ausdrücke

- Suchmuster
  - regulärer Ausdruck mit Matches
  - mit Optionen (Groß-/Kleinschreibung, . passt auf alles, …)
- Ersetzung
  - feste Zeichenkette[1]
  - mit Optionen (global, nur im Bereich, …)

Reguläre Ausdrücke

- Gefundenes in die Ersetzung bringen:
  - <(.*)> → <hi rendition="#aq">$1</hi>
- $n ist der Platzhalter für den Inhalt im n-ten matchenden Klammerpaar
- für Berechnungen im Ersetzungsteil: Skriptsprache benutzen

Reguläre Ausdrücke

- Hinweise:
  - Optionen beachten.
  - Jede Reguläre-Ausdrücke-Implementierung ist anders.
  - Vorsicht vor der Gierigkeit!
  - Backup machen.

Textverarbeitungen

- Hoffen, dass Vorlagen (Templates) benutzt wurden.
- OxGarage Conversion:
  - http://www.tei-c.org/oxgarage/
- DTA bietet Unterstützung bei der Konvertierung von MS Word nach DTABf.

Wikisource

- http://de.wikisource.org/
- 32316 Werke (deutschsprachig)
- MediaWiki-Markup (sehr auf Präsentation orientiert)
- DTA-Konverter:
  - http://www.deutschestextarchiv.de/dtæ/submit/wikisource

PDF

- präsentationsorientiertes Format
  - keine Semantik
  - drucktechnische Fallstricke (z. B. Ligaturen)
- zuerst immer Urheber fragen, ob Ausgangsdaten vorliegen
- pdftotext, pdf2html …

XML (TEI, HTML oder andere)

- Vorsicht mit regulären Ausdrücken:

```
<span class="italic">(.*)</span>
→
<hi rendition="#i">$1</hi>
```

```
<span class="bold">ein <span class="italic">kursives</span>
Wort</span>

→

<span class="bold">ein <hi rendition="#i">kursives</span>
Wort</hi>
```

XML (TEI, HTML oder andere)

- Transformationstool benutzen (XSLT)
- XSLT ist eine Programmiersprache in XML
- Ergebnis ist immer wohlgeformt
- oXygen bietet einen XSLT-Editor

```
<xsl:template match="h1">
  <head><xsl:value-of select="."/></head>
</xsl:template>
```

Fazit

- händische Wiederholungen vermeiden
- Experten fragen
- Aufwand abschätzen (lassen)
- kenne und benutze die Tools