

Coding the Sources

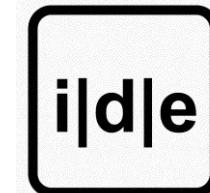
Digitales Edieren in den Geisteswissenschaften

Online Summerschool und Workshop

Lübeck/Online, 17.08. – 27.08.2020



Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums



Christian-Albrechts-Universität zu Kiel

Historisches Seminar

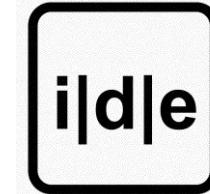
TEI Vertiefung

Transkription und Faksimile

Torsten Schaßan
schassan@hab.de
Lübeck/Online, 20.8.2020



Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums



TEI Grundstruktur(en)

- Ein TEI-Dokument enthält mindestens zwei Unterelemente:
 - <teiHeader> für Metadaten
 - Eines der folgenden Elemente für die Repräsentation des Textes:
 - <text> für „normale“ Transkriptionen
 - <facsimile> für bildhafte Repräsentation des Textes
 - <sourceDoc> für Embedded Transcription

<text>

- <text> ist in der TEI das älteste Element zur Repräsentation des Textes

```
1 ▼<TEI xmlns="http://www.tei-c.org/ns/1.0">
2 ▼  <teiHeader>
3 ▼    <fileDesc>
4 ▼      <titleStmt>
5        <title>Title</title>
6      </titleStmt>
7 ▼      <publicationStmt>
8        <p>Publication Information</p>
9      </publicationStmt>
10 ▼     <sourceDesc>
11       <p>Information about the source</p>
12     </sourceDesc>
13   </fileDesc>
14 </teiHeader>
15 ▼  <text>
16 ▼    <body>
17      <p>Some text here.</p>
18    </body>
19  </text>
20 </TEI>
```

<facsimile>

- Mit Aufkommen der Digitalisierung wurde <facsimile> hinzugefügt
- „<facsimile> contains a representation of some written source in the form of a set of images rather than as transcribed or encoded text.“
- Wichtigste Kindelemente sind
 - <graphic>
 - <surface>

```
1 ▾ <TEI xmlns="http://www.tei-c.org/ns/1.0">
2 ▾   <teiHeader>
3 ▾     <fileDesc>
4 ▾       <titleStmt>
5         <title>Title</title>
6       </titleStmt>
7 ▾       <publicationStmt>
8         <p>Publication Information</p>
9       </publicationStmt>
10 ▾      <sourceDesc>
11        <p>Information about the source</p>
12      </sourceDesc>
13    </fileDesc>
14  </teiHeader>
15 ▾  <facsimile>
16    <graphic url="page1.png"/>
17 ▾    <surface>
18      <graphic url="page2-highRes.png"/>
19      <graphic url="page2-lowRes.png"/>
20    </surface>
21    <graphic url="page3.png"/>
22    <graphic url="page4.png"/>
23  </facsimile>
24 </TEI>
```

<graphic>

- „<graphic> indicates the location of a graphic or illustration, either forming part of a text, or providing an image of it.”
- Im Faksimile
- In der Transkription

```
21 ▾ <text>
22 ▾   <body>
23 ▾     <div>
24       <pb facs="#bild_1"/>
25 ▾       <figure>
26         <graphic url="fig1.png"/>
27         <head>Figure One</head>
28         <figDesc><!-- Bildbeschreibung --></figDesc>
29       </figure>
30     </div>
31   </body>
32 </text>
```

Verbindung von <text> und <facsimile>

- Identifikation der Bilder per @xml:id

```
1 ▾ <TEI xmlns="http://www.tei-c.org/ns/1.0">
2 ▾   <teiHeader>
3 ▾     <fileDesc>
4 ▾       <titleStmt>
5         <title>Title</title>
6         Zeilen für Fold : 2
7           </titleStmt>
8         <publicationStmt>
9           <p>Publication Information</p>
10          </publicationStmt>
11        <sourceDesc>
12          <p>Information about the source</p>
13        </sourceDesc>
14      </fileDesc>
15    </teiHeader>
16    <facsimile>
17      <graphic url="page1.png" xml:id="bild_1"/>
18      <graphic url="page2.png" xml:id="bild_2"/>
19      <graphic url="page3.png" xml:id="bild_3"/>
20      <graphic url="page4.png" xml:id="bild_4"/>
21    </facsimile>
22  </TEI>
```

Verbindung von <text> und <facsimile>

- Verweis von Elementen der Transkription auf das Bild mit @facṣ

```
15 ▾   <facsimile>
16     <graphic url="page1.png" xml:id="bild_1"/>
17     <graphic url="page2.png" xml:id="bild_2"/>
18     <graphic url="page3.png" xml:id="bild_3"/>
19     <graphic url="page4.png" xml:id="bild_4"/>
20   </facsimile>
21 ▾   <text>
22 ▾     <body>
23 ▾       <div>
24         <pb facs="#bild_1"/>
25         <!-- Transkription hierher -->
26       </div>
27     </body>
28   </text>
```

Genetic Editing

- Text, Dokument, Prozess
- TEI Working Group “Genetic Editions” (Special Interest Group on “Manuscripts”)
- Proposal → 2011 revision of module 11: “Representation of Primary Sources”
- Aims:
 - Transcription, showing the evolution of the text on the document
 - Reconstruction of the genesis of a text from different documents
 - Editing, i.e. the representation of the genesis of a text in an edition

Embedded Transcription

- „An embedded transcription is one in which words and other written traces are encoded as subcomponents of elements representing the physical surfaces carrying them rather than independently of them.”
- The following elements are available for this purpose:
 - [sourceDoc](#) contains a transcription or other representation of a single source document potentially forming part of a dossier génétique or collection of sources.
 - [surface](#) defines a written surface as a two-dimensional coordinate space, optionally grouping one or more graphic representations of that space, zones of interest within that space, and transcriptions of the writing within them.
 - [zone](#) defines any two-dimensional area within a [surface](#) element.
 - [line](#) contains the transcription of a topographic line in the source document
 - [seg](#) (arbitrary segment) represents any segmentation of text below the ‘chunk’ level.

<sourceDoc>

- Das jüngste der Elemente für die Repräsentation von Text
- „<sourceDoc> contains a transcription or other representation of a single source document potentially forming part of a dossier génétique or collection of sources.”
- “This element may be used as an alternative to [facsimile](#) for TEI documents containing only page images, or for documents containing both images and transcriptions. Transcriptions may be provided within the [surface](#) elements making up a source document, in parallel with them as part of a [text](#) element, or in both places if the encoder wishes to distinguish these two modes of transcription.”

<sourceDoc>

- <sourceDoc> kann die <graphic>-Elemente selbst enthalten

```
21 ▾    <sourceDoc>
22 ▾        <surfaceGrp n="leaf1">
23 ▾            <surface facs="page1.png">
24                <zone>All the writing on page 1</zone>
25            </surface>
26 ▾            <surface>
27                <graphic url="page2-highRes.png"/>
28                <graphic url="page2-lowRes.png"/>
29 ▾                    <zone>
30                        <line>A line of writing on page 2</line>
31                        <line>Another line of writing on page 2</line>
32                    </zone>
33            </surface>
34        </surfaceGrp>
35    </sourceDoc>
```

<surface>

- „<surface> defines a written surface as a two-dimensional coordinate space, optionally grouping one or more graphic representations of that space, zones of interest within that space, and transcriptions of the writing within them.”
- Vgl. IIIF (International Image Interoperability Framework, <http://iiif.io>)
 - SharedCanvas Modell
 - IIIF wird als JSON serialisiert, daher z.Z. noch keine TEI-Integration
 - Bsp. Manuscripts from German-Speaking Lands – A Polonsky Foundation Digitization Project (<https://hab.bodleian.ox.ac.uk/en/>)

<zone>

- The position of every zone for a given surface is always defined by reference to the coordinate system defined for that surface.
- A zone may be of any shape. The attribute points may be used to define a polygonal zone, using the coordinate system defined by its parent surface.
- A zone is always a closed polygon. Repeating the initial coordinate at the end of the sequence is optional. To encode an unclosed path, use the [path](#) element.
- [att.coordinated](#) (@start, @ulx, @uly, @lrx, @lry, @points)

<zone>

```
21 ▾ <sourceDoc>
22 ▾   <surface ulx="0" uly="0" lrx="200" lry="300">
23 ▾     <zone ulx="0" uly="0" lrx="200" lry="300">
24       <graphic url="Bovelles-49r.png"/>
25     </zone>
26   <!-- ... -->
27 ▾     <zone ulx="28" uly="75" lrx="175" lry="178">
28 ▾       <line>LEs cloches ont quasi
29         fi</line>
30       <line>gures de rondes pyra</line>
31 ▾       <line>mides imperfaictes &amp;
32         </line>
33       <line> irregulieres: &amp; leur accord se</line>
34       <line> fait par reigle geometrique. Com</line>
35 ▾       <line>me si les deux cloches C
36         &amp; D </line>
37 ▾       <line> sont <zone ulx="45" uly="125" lrx="60"
38         lry="130">pendans</zone> à ung mesme axe</line>
39 ▾       <line> ou essieu A B: je dis que
40         leur ac</line>
41       <line>cord se fera en cōtraires parties</line>
```

<line>

- „<line> contains the transcription of a topographic line in the source document”
- „This element should be used only to mark up writing which is topographically organized as a series of lines, horizontal or vertical. It should not be used to mark lines of verse (for which use [l](#)) nor to mark linebreaks within text which has been encoded using structural elements such as [p](#) (for which use [lb](#)).”

Typen von Zeilen

Embedded Transcription	„normale“ Traskription	Gedichtzeilen
<pre><surface> <zone> <line>Poem</line> <line>As in Visions of — at</line> <line>night —</line> <line>All sorts of fancies running through</line> <line>the head</line> </zone> </surface></pre>	<pre><div> <pb/> <p> <lb/>Poem <lb/>As in Visions of — at <lb/>night — <lb/>All sorts of fancies running through <lb/>the head </p> </div></pre>	<pre><div> <lg> <head>Poem</head> <l>As in Visions of — at night — </l> <l>All sorts of fancies running through the head</l> </lg> </div></pre>

Tools

- TEI-Facsimile-Plugin für Oxygen (<https://github.com/oxygenxml/TEI-Facsimile-Plugin>)
- Image Markup Tool
(https://hcmc.uvic.ca/~mholmes/image_markup/xml.php)
- Zur Ermittlung der Zonen: IrfanView (<http://www.irfanview.com>)

Weiterführende Literatur:

An Encoding Modell for Genetic Editions (<https://tei-c.org/Vault/TC/tcw19.html>)

Martina Semlak: [Genetic Editing](#) (Talk, https://dixit.uni-koeln.de/wp-content/uploads/2015/04/Camp2-10-Martina_Semlak_-_Genetic_Editing__talk.pdf)

Fragen?

schassan@hab.de



Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums

