

Interpret the code: Towards a generic processing and presentation of TEI encoded digital editions

Torsten Schaßan

Herzog August Bibliothek Wolfenbüttel, Germany; schassan@hab.de

The set of XSL stylesheets provided and maintained by the TEI is relatively cautious about processing and presentation of transcriptions and content of digital editions. Only very basic functions are implemented, such as to surround abbreviations with brackets or to process from the element choice in plain mode only those children that represent the "critical" reading. Processing in plain mode means that the elements will be treated as in-line elements.[1]

On the other hand the encoding must have been done with a special purpose. A general rule of text encoding is that the editor may encode only those structures and semantic features that he wants to process in the end. The processing might include elaborated examination and analysis of the encoded text or more complex queries as well as a reproduction of visual properties of the original document or provide a (simplified) reading text. Thus the encoding will tell something about the functionalities of the text in processing and presentation.

According to Patrick Sahle's "Textrad"[2], "the" text does not exist in a transcription but the encoded text usually represents multiple properties and serves multiple purposes. Whatever the editor might state in some introductory notes and the documentation of the edition which should contain some statements about the encoding used, in the end the encoded text will speak on its own, can be interpreted and will be processed as is.

In succession of the modelling of the TEI, realised in the modules, the grouping of elements and of attributes, the semantics of certain elements might let the processor estimate about the foreseen presentation, processing, and use:

- The elements `<pb>`, `<lb>`, `<l>`, `<lg>`, etc. as well as attributes `@rend`, `@rendition` or `@style` represent visual aspects of the text, therefore these might have to be reproduced; users may be given a choice to either see a document-centred view which visualises these aspects or switch to an editorial view on the text which eliminates these properties.
- The same applies to the element `<choice>`: If this is used the editor must have had in mind the opportunity to change the views on the document respectively encoded text.
- Entities encoded as `<rs>`, `<name>`, `<persName>`, `<placeName>`, etc might be referenced, especially if they are accompanied by the related list elements such as `<listPerson>`, `<listPlace>`, etc. Additionally, one might assume that there will be norm data available which allows for links into the open.
- Bibliographic records (`<bibl>`, `<msDesc>`) will serve a similar purpose and will have to be referenced.
- Quotations like `<cit>`, `<foreign>`, `<q>`, `<quote>` etc. will have to be distinguished from the surrounding text.

Concerning the overall structure of an (critical) edition one might expect up to three apparatuses: The critical apparatus, the commentary and maybe a bibliographical apparatus. How many of these are present in a given edition is up to the editor but the presence of certain elements and especially of the amount of certain elements will give anybody an idea of how many apparatuses are "appropriate": If the encoding contains

editorial elements like <choice>, <abbr>/<expan>, <add>/, etc. the representation of this information in an apparatus will be inevitable. If a certain amount of bibliographic references point to biblical or classical texts the tradition of the publication of editions has provided a separate apparatus as well. Last, editorial notes will have to be distinguished from the former two categories.

This paper will examine existing editions with statistical methods and by clustering the elements used it might be possible to assign the encoded text to one or more text types of Sahle's typology. Additionally, the paper shall foster the discussion about the presentation of an edited text according to the intended purpose of the encoding. On the basis of the typology and purpose of the edition it will be more likely that a generic presentation of any edited text is possible. Finally, with some statistical data about the editions some remarks about the interoperability of the TEI-encoded texts shall be possible.

[1] e.g. https://github.com/TEIC/Stylesheets/blob/master/html/html_core.xsl

[2] Patrick Sahle: *Digitale Editionsformen*, 3 vols. 2013, esp. vol. 3, p. 9ff.