From MS Word, Transkribus & Co. to TEI Encoded Text

Lecture Session 6

Oslo Summer School
2025

Annika Rockenberger, PhD

XML TEI and How to Get There

1. Hard Coding

Pros

- Quickly learn and use TEI Guidelines and Unicode codes
- Familiarise oneself with plain text editors
- Getting used to "reading" code

Cons

- Steep learning curve
- Slow, time-consuming process
- Doesn't work well for long documents
- GUI fatigue

```xml
<pb facs="#f0044" n="B10b"/>
<fw type="header" place="top">
<hi rendition="#aq">Complements</hi>
</fw>
<lb/>
vnd dero Sachen <hi rendition="#aq">importans</hi>, ha-<lb/>
ben willig er&#x017F;cheinen wollen / mit<lb/>
Erbietung zu allen beha&#x0364;glichen<lb/>
Gegendien&#x017F;ten / &#xA75B;c. Darauff die<lb/>
Sachen an jhr &#x017F;elb&#x017F;t fein form- vnd<lb/>
vmb&#x017F;ta&#x0364;ndlich vorbringen / daru&#x0364;ber<lb/>
eines jeden <hi rendition="#aq">Votum</hi>, Stimm vnd<lb/>
Meinung / nach be&#x017F;ten Wi&#x017F;&#x017F;en vnd<lb/>
richtigen Gewi&#x017F;&#x017F;en begehren / vnd<lb/>
herumb gehen la&#x017F;&#x017F;en / auch &#x017F;chrifft-<lb/>
lich <hi rendition="#aq">protocolli</hi>ret vnd auffge&#x017F;etzet.</p>
<lb/>
<p>Die <hi rendition="#aq">Collegen</hi> aber einer nach<lb/>
dem andern / neg&#x017F;t <hi rendition="#aq">replici</hi>rung jres<lb/>
Ampts vnd anderer Vmb&#x017F;ta&#x0364;nden<lb/>
mehr / mu&#x0364;&#x017F;&#x017F;en die Sachen wol be-<lb/>
da&#x0364;chtlich <hi rendition="#aq">ponderi</hi>ren vnd erwe-<lb/>
gen / wozu gute <hi rendition="#aq">Logici</hi> geho&#x0364;ren /<lb/>
welche be&#x017F;ta&#x0364;ndige vnwiedertreibli-<lb/>
che ver&#x017F;ta&#x0364;ndige <hi rendition="#aq">Argumenta</hi> vnd<lb/>
Gru&#x0364;nde / &#x017F;o wol <hi rendition="#aq">ex Apodicticis</hi></hi>
<lb/>
als <hi rendition="#aq">ex Topicis polité &amp; politicé</hi>,<lb/>
<fw type="catch" place="bottom">anzu-</fw>
<lb/>
```

XML TEI and How to Get There

2. MS Word et al.

Pros

- Readily accessible (!)
- Familiarity with GUI and markup
- Encoding task easier to delegate

Cons

- TEI Guidelines and encoding remain obscure
- WYSIWYG sloppiness
- Doesn't work well for long documents (but for different reasons)
- [Proprietary format, expensive license]

XML TEI and How to Get There

https://teigarage.tei-c.org/#

https://teigarage.tei-c.org/#

XML TEI and How to Get There

https://www.transkribus.org/

https://www.transkribus.org/
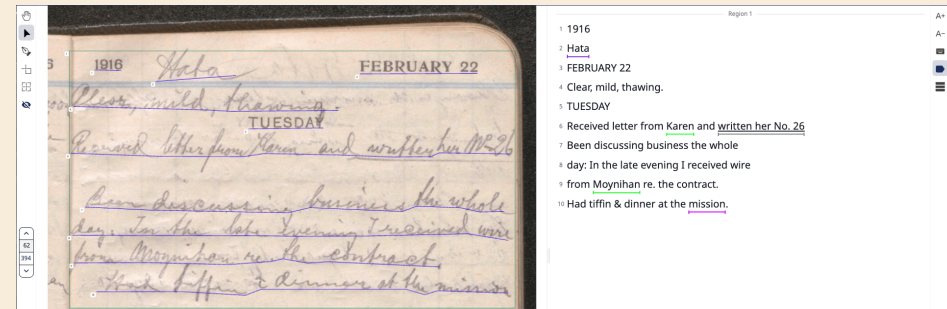
XML TEI and How to Get There

3. Transkribus

Pros

- Readily accessible (!)
- User-friendly GUI and intuitive markup
- Encoding task easier to delegate

Cons

- TEI Guidelines and encoding remain obscure
- Page unit vs. layout/semantic unit
- In "the cloud", not on a local server
- Only with a subscription or membership, can be expensive
- One code-maintainer

TEI XML: Export in TEI XML format, suitable for academic research and digital humanities projects. You have a choice between two stylesheet options, each catering to different project requirements:

- **Standard**:
  This option refers to Transkribus's default TEI XML export format. It is designed to provide broad compatibility with TEI guidelines, ensuring that your exported document adheres to widely accepted standards for encoding textual data.

- **Page2tei**:
  Choosing the page2tei option allows you to export your document using a specific XSL transformation, created by Dario Kampkaspar. This transformation is available on GitHub: page2tei.

  This option is particularly useful for those requiring a more customised or detailed approach to TEI XML encoding. With Page2TEI, you will download a TEI XML file that is specifically structured according to the guidelines and enhancements provided by the Page2TEI XSL transformation. This can include more detailed representations of page layouts, specialized tagging, and other features that are not covered by the standard TEI XML export.

Try it Out!

1. Take a small MS Word file (or Markdown, Open Office, LaTeX...) and convert it into a TEI XML file using the TEIGarage:
https://teigarage.tei-c.org/#

ATTENTION: Large documents take a lot of time, and the server might time out!

2. Go to **https://github.com/TEIC/TEIGarage** and read the About section. You can also look at some recent presentations about TEIGarage.
OR
Go to **https://zde-uzh.github.io./proto4diged-handbook/en/2_Editionsarbeit/04_converting/** and read the documentation for the Transkribus export.

Discuss with your neighbour the pros and cons of using a converter like TEIGarage or an export from Transkribus for your project.

Contact

annika.rockenberger@ub.uio.no

baerut-admin@ub.uio.no