Curation and Integration of Historical Text Resources of the 15th–19th Century

WG 1: German Philology – Curation Project 1 (Frederike Neuber, BBAW)

Quick Facts

Partner Institutions

- BBAW – Berlin-Brandenburg Academy of Sciences and Humanities, project coordination by the German Text Archive (Deutsches Textarchiv, DTA)
- JLU – Justus Liebig University Giessen
- HAB – Herzog August Library Wolfenbüttel
- IDS – Institute for the German Language Mannheim

Duration: 2012/09/01–2014/02/28

Project Description

The curation project aims to improve the current status of German corpora of the Early New High German and the historical New High German language (1400–1900) to facilitate further research, e.g. in historical linguistics.

To this end, digital full-text resources (and corresponding scans of the original prints) are scanned throughout the web or stored locally – which often contain project-specific mark-ups, lack sufficient metadata, and are stored in inflexible formats – are to be integrated into the repository-structures of the BBAW (DTAE) and HAB (WDB). Through a newly implemented sustainable infrastructure for the integration of full-texts – using already existing tools from the DTA and developing a new workflow – more than 50,000 pages have been curated so far.

As the final step, IDS and BBAW will integrate the resources into the CLARIN-D infrastructure making the data available through the Virtual Language Observatory (VLO) and the CLARIN-D Federated Content Search (FCS).

DTA Tools

DTA ›base format‹ (DTABf)

www.deutschestextarchiv.de/doku/software#cab

All texts are structurally annotated according to the DTA ›base format‹, a subset of TEI/PS containing about 50 <teileader> elements and 75 <text> elements. DTABf helps to gain coherence at the annotation level, given the heterogeneity of the texts regarding time and text types.

The DTA ›base format‹ is CLARIN-D best practice format for the encoding of historical printed texts (cf. CLARIN-D User Guide).

Workflow

In DTAQ texts may be proofread page-wise in comparison with their source images. The GUI is highly customizable, so that transcriptions and metadata are offered in different formats (TCF, XML/TEI, HTML, raw text, linguistic analyses, CMDI, DC). Users may categorize errors and report them for correction.

Linguistic Analysis with CAB

www.deutschestextarchiv.de/doku/software#cab

Once integrated into DTAQ, the texts are analyzed by the Cascaded Analysis Broker (CAB). CAB provides a normalization of the historical orthography in order to allow for spelling-tolerant, lemma based corpus searches.

Contact and Links

Project Coordinator

Christian Thomas (BBAW/DTA, CLARIN-D) thomas@bbaw.de

Project Website

www.clarin-d.de/en/discipline-specific-working-groups/wg-1-german-philology/curation-project-1.html

Integrated Texts (DTAE)

www.deutschestextarchiv.de/dte

Statistics (June 2013)

<table>
<thead>
<tr>
<th>Integration progress</th>
<th>volumes</th>
<th>pages</th>
<th>types</th>
<th>token</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other: 12,081</td>
<td>408</td>
<td>56,263</td>
<td>490,379</td>
<td>13,636,427</td>
</tr>
<tr>
<td>Gutenberg.org: 1,097</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WDB, Oberhofprediger: 23,526</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wikisource: 19,560</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sources

Primary sources for the selection of texts:

- large scale text collections with focus on Wikisource and Gutenberg.org (BBAW)
- research projects and scholarly editions, e.g. Oberhofprediger (HAB), DFG-project “Entstehung und Entwicklung der Zeitungssprache im 17. Jhd.” (JLU), sandart.net (BBAW)
- collections from individual scholars on different subjects, e.g. cookery, clothing, sport (JLU)