CLARIN-D Showcase: Integration of Wikisource data sets into the CLARIN-D Infrastructure

CLARIN Centres: Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)
Justus Liebig University Giessen
Frank Wiegand (German Text Archive, Deutsches Textarchiv)

URL: http://www.deutschestextarchiv.de/dtae/import/wikisource

status: in production since 2012/03

Focus Group

- historical linguistics
- scholarly editions
- historians

Use Case

Numerous high-quality primary text sources (full-text transcriptions, and corresponding image scans) of German works originating from the 15th to the 19th century are scattered among the web or stored remotely. Additionally, idiosyncratic, project-specific markup conventions and uncommon, out-of-date or inflexible storage formats often hinder further usage and analysis of the data. Often, textual resources are accompanied by scarce, insufficient or inaccurate bibliographic information, which is only one further reason why valuable resources, even if available on the web, remain undiscovered by and are of little use to the wider research community. The integration of these dispersed primary text sources into the sustainable, web- and centres-based research infrastructure of CLARIN-D will be an important step to solve this problem. Unfortunately, most textual resources can not be converted fully automatically into the preferred encoding format (DTA base format, a TEI P5 compliant XML subset) to qualify for subsequent use.

In this show case, we illustrate a CLARIN-D toolchain for the import of legacy data into the CLARIN-D infrastructure. As an example, we use text transcriptions of the German Wikisource project. Our toolchain has been in use since March 2012. Since then 146 works (19,418 pages of text) have been integrated into the German Text Archive, thus making it available to the CLARIN-D infrastructure.

Workflow

Step 1: Collecting metadata, text and images sources

Wikisource provides basic metadata information for each text. Our conversion tool tries to extract and restructure as much metadata information as possible. An extensive metadata form is provided to allow users to provide additional information about the primary text source, the source of the images, and information about the transcribed and annotated version of the text.

Step 2: Conversion to DTABf using DTAoX

After fetching the Wikisource HTML, and applying scripts to partially convert the data into DTABf, the DTA oXygen framework DTAoX helps the user to manually correct the text in order to get a fully DTABf compliant XML document.

Step 3: Quality assurance using DTAQ

Using DTAQ, several possible error sources can be addressed (metadata errors, erroneous parts of the transcription, alignment of images and text etc.).

Step 4: Integration into the CLARIN-D infrastructure

After quality assurance, the document and its metadata is available in several data formats. It is also searchable via Federated Content Search.

CLARIN-D Integration

internal data format:	DTABf (TEI P5 compliant XML subset)
exportable formats:	DTABf, HTML, TCF, plain text
online formats:	aligned image/text view, normalized orthography, POS
metadata:	CMDI, TEI header, DC, xepicur, export via OAI-PMH into the VLO
access:	during QA: free registration required after QA: free access (Creative Commons or compliant license)

References

- Alexander Geyken, Susanne Haaf, Frank Wiegand: The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora. In: Proceedings of the 11th Conference on Natural Language Processing (KONVENS). Vienna, 2012. http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf
- Frederike Neuber, Christian Thomas: Vorstellung des Kurationsprojekts 1 der Clarin-D-FAG 1 „Deutsche Philologie". In: Clarin-D-Newsletter 3, November 2012, pp. 11–13. http://www.clarin-d.de/images/newsletter/CLARIN-D-Newsletter-2012-3.pdf
- Christian Thomas, Frank Wiegand: Making great work even better. Appraisal and Digital Curation of widely dispersed Electronic Textual Resources (c. 15th–19th cent.) in CLARIN-D. Full Paper for the International Conference "Historical Corpora 2012", December 6–9, 2012; Goethe University, Frankfurt, Germany. urn:nbn:de:kobv:b4-opus-23081
- German Wikisource project: http://de.wikisource.org/
- DTA base format (DTABf): http://www.deutschestextarchiv.de/doku/basisformat
- DTA quality assurance (DTAQ): http://www.deutschestextarchiv.de/dtaq/about