Language Data: Curation Projects and Integration Scenarios

CLARIN-D

SPONSORED BY THE Federal Ministry of Education and Research

The aim of the curation projects was to identify key language data and resources and then to sustainably integrate those data into the CLARIN-D infrastructure. In the area of German studies three curation projects made it possible to add three types of data into the CLARIN-D centres: **historical texts** of the 15th-19th centuries, **spoken language data**, and **computer mediated communication (CMC)**.

In addition, we developed **integration scenarios** that can be used for further **integration use cases** in a sustainable way.

Curation Project 1 / UC 1 Integration of Historical Texts

Integration and Enhancement of Historical Textual Resources of the 15th-19th centuries
Integration und Aufwertung historischer Textressourcen des 15.-19. Jahrhunderts

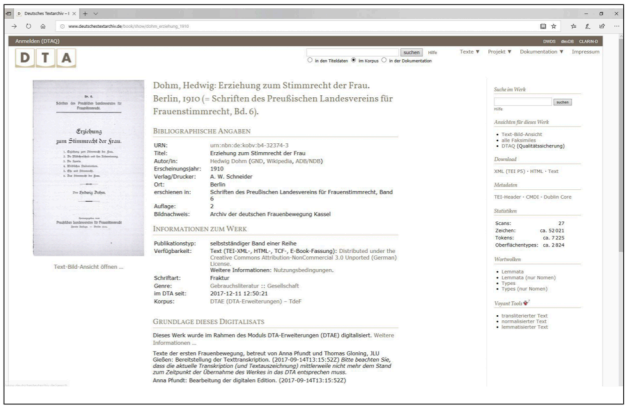
The curation project aimed to improve the current status of German corpora of the 15th to 19th century. The curated resources are free available through the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW). The curation project is coordinated by the Deutsches Textarchiv (DTA) at BBAW.

The curation project has also delivered an integration scheme for historical data that can be used in a sustainable way.

Examples of Curated Resources:
– "Texte der ersten Frauenbewegung (TdeF)"
The corpus "Texte der ersten Frauenbewegung (TdeF)" (Anna Pfundt/Thomas Gloning) consist of historical texts of early women's movement of the 19th century in Germany. It delivers booklets, brochures, and periodical publications.

– Dingler's "Polytechnisches Journal" - Dingler Online
The corpus makes all 375 volums of the „Polytechnisches Journal" (1820-1931) digital available.

– diGltexte
The corpus of diGltexte contains German speaking historical texts of the 15th-19th centuries. Different text formats are converted into the DTA basic format template.



Curation Project 2 / UC 2 Integration of Spoken Language Data

GeWiss: A Comparable Corpus of Spoken Scientific Language
Gesprochene Wissenschaftssprache (GeWiss-Korpus)

The curation project aimed to integrate all existing published and unpublished resources of the GeWiss project and to make them available to the academic community at large, using a format compatible with CLARIN-standards. GeWiss provides the academic community with the first freely available corpus resource for the comparative analysis of spoken academic language.

The curated resources are made available through the existing web portal of the GeWiss project (free registration). In addition, the Institut für Deutsche Sprache (IDS) will make the data available through its CLARIN service centre infrastructure in the near future.

The curation project has also delivered an integration scheme for spoken language data that can be used in a sustainable way.

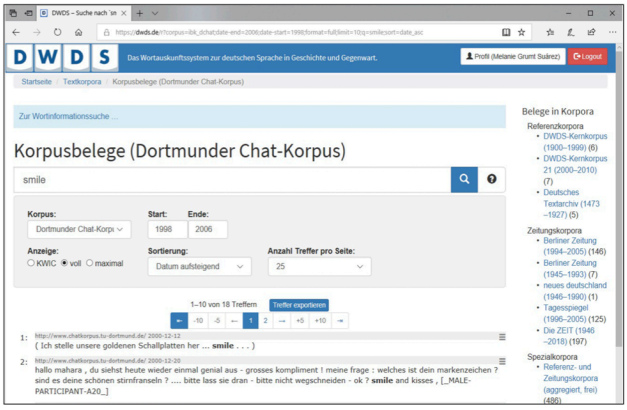
Curation Project 3 / UC 3 Integration of CMC Data

ChatCorpus2CLARIN: Integration of the Dortmund Chat Corpus into CLARIN-D
ChatCorpus2CLARIN: Integration des Dortmunder Chat-Korpus in CLARIN-D

Here, an existing corpus of computer-mediated communication (CMC), the Dortmund Chat Corpus, and samples of other CMC resources have been restructured to conform to current standards for the representation of corpora in the Digital Humanities context.

The main goal of this work was to pave the way for the inclusion of linguistically annotated CMC resources into CLARIN-D corpus infrastructures and to create the prerequisites for investigating linguistic peculiarities of CMC with state-of-the art corpus technology. The curated resources are available through the Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS) (free registration).

The curation project has also delivered an integration scheme for CMC data that can be used in a sustainable way.



Bibliography and Webliography

Publications CP1 / UC 1

Geyken, Alexander/Thomas **Gloning** (2015): "A living text archive of 15th-19th century German. Corpus strategies, technology, organization." In: Jost Gippert (eds.): Historical corpora: challenges and perspectives. Tübingen: Narr, 165-180.

Pfundt, Anna (2017): "Frauenwahlrecht? Oder Damenwahlrecht? Oder doch ein allgemeines Wahlrecht? – Zum Wortgebrauch in der Diskussion um das Frauenwahlrecht um 1900." In: Im Zentrum Sprache, 2.11.2017, https://sprache.hypotheses.org/542 (last retrieved May 28, 2018).

Thomas, Christian/Frank **Wiegand** (2015): "Making great work even better. Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th-19th centuries) in CLARIN-D." In: Jost Gippert (eds.): Historical corpora: challenges and perspectives. Tübingen: Narr, 181-196. [Online Version 31.10.2012 https://edoc.bbaw.de/opus4-bbaw/frontdoor/index/index/docId/2005 (last retrieved November 30, 2017)].

Publications CP2 / UC 2

Fandrych, Christian/Cordula **Meißner**/Adriana **Slavcheva** (2012): "The GeWiss Corpus: Comparing Spoken Academic German, English and Polish." In: Thomas Schmidt/Kai Wörner (eds.): Multilingual corpora and multilingual corpus analysis. Amsterdam: Benjamins, 319-337.

Fandrych, Christian/Cordula **Meißner**/Franziska **Wallner** (eds.) (2017): Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora. Tübingen: Stauffenburg.

Jettka, Daniel (2017): "Zur Integration des GeWiss-Korpus in die CLARIN-Infrastruktur." In: Christian Fandrych/Cordula Meißner/Franziska Wallner (eds.): Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora. Tübingen: Stauffenburg, 31-47.

Publications CP3 / UC 3

Beißwenger, Michael/Harald **Lüngen**/Jan **Schallaböck**/John H. **Weitzmann**/Axel **Herold**/Pawel **Kamocki**/Angelika **Storrer**/Julia **Wildgans** (2017): "Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens." In: Michael Beißwenger (eds.): Empirische Erforschung internetbasierter Kommunikation. Berlin/Boston: De Gruyter, 7-46. (doi: https://doi.org/10.1515/9783110567786-002) (last retrieved November 30, 2017).

Beißwenger, Michael/Eric **Ehrhardt**/Andrea **Horbach**/Harald **Lüngen**/Diana **Steffen**/Angelika **Storrer** (2015): Adding Value to CMC Corpora: CLARINification and Part-of-Speech Annotation of the Dortmund Chat Corpus. In: Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015). Essen, 12-16. https://sites.google.com/site/nlp4cmc2015/proceedings (last retrieved November 30, 2017).

Erhard.Hinrichs@uni-tuebingen.de (http://www.sfs.uni-tuebingen.de/~eh/) | Scientific Coordinator CLARIN-D ||| Melanie.Grumt-Suarez@germanistik.uni-giessen.de (@mela_g_s) | Working Group 1 "German Philology"