Language Data: Curation Projects and Integration Scenarios

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, 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

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

Examples of Curated Resources:

Curation Project 2 / UC 2 Integration of Spoken Language Data

The curation project aimed to integrate all existing published and unpublished resources of the GfWiss project and to make them available to the academic community at large, using a format compatible with CLARIN standards.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

The curated resources are made available through the existing web portal of the GfWiss 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

The project aimed to integrate the existing corpus of computer-mediated communication (CMC), the Dortmund Chat Corpus, and samples of other CMC resources have been restructured to conform to 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 DigiZed Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDSS) (free registration).

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

Bibliography and Weblogiography

Publications CP1 / UC 1


Publications CP2 / UC 2


Publications CP3 / UC 3