Abstract

Numerous high-quality primary text sources—in the context of the curation project described here, this means full-text transcriptions (and corresponding image scans) of German works originating from the 15th to the 19th centuries—are scattered among the web or stored remotely. E.g., transcriptions of historical sources are stored locally on degrading recording media and cannot be found, let alone accessed by third parties. Additionally, idiosyncratic, project-specific markup conventions and uncommon, out-of-date or inflexible storage formats often hinder further usage and analysis of the data.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.

The Full Paper illustrates an exemplary approach taken by the »Deutsches Textarchiv« (DTA; www.deutschestextarchiv.de) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) to integrate dispersed textual resources and corresponding image scans from various sources into a large historical text corpus of its own and to insert these into the infrastructure of CLARIN-D.