Christian Thomas | www.deutschestextarchiv.de | thomas@bbaw.de

DTAE: Enlarging the reference corpus of the Deutsches Textarchiv (DTA)
Production, conversion and interchange of XML/TEI encoded full text

Deutsches Textarchiv (DTA) – Short description

The Deutsches Textarchiv (henceforth: DTA) provides a broad selection of significant German-language works of various disciplines, ranging from the 17th to 19th century. These text sources are published via the internet as digital facsimiles and as XML-annotated transcriptions along with bibliographic metadata. The electronic full-texts are enriched with linguistic information gained through tokenization, lemmatization and part-of-speech-analysis.

DTAE: Enlarging the reference corpus…

In the course of the project (which runs until 2014), the DTA aims to publish around 1,300 volumes on its own. To even enhance this ›core collection‹, the software module DTAE (“E” stands for Enlargement or Extension) was developed. With the help of DTAE, external projects can integrate their historical text collections into the DTA reference corpus. They can present their data in a larger context and benefit from the elaborate linguistic search engine and text processing routines of the DTA. In addition, external contributors can integrate resp. re-import the processed text and metadata into their own web site via <iframe>.

DTAE provides routines for uploading metadata, text and images, as well as semiautomatic conversion tools from different source formats (plain text, MS Word, TUSTEP, HTML, TEI-XML, …) into the XML/TEI conformant ›base format‹ of the DTA.[1] DTAE thus demonstrates how interchange and interoperability among projects can work on a large scale.

The presentation illustrates the described approach by different examples of text interchange resp. text production partnerships between the DTA and its external partners, i.e. the MPI, the HAB Wolfenbüttel and the Göttingen Academy of Sciences and Humanities. Possibilities and challenges of the exchange of XML/TEI documents will be discussed.

Further Reading

Geyken, Alexander et al. (2011): „Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv“; in: Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, 20./21. September 2010. Beiträge der Tagung. Hrsg. von Silke Schomburg, Claus Leggewie, Henning Lobin und Cornelius Puschmann. 2., erg. Fassung. hbz, 2011, S. 157–161. (http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf#page=159)

Geyken, Alexander et al. (2011): „TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv“; in: Jahrbuch für Computerphilologie (forthcoming paper).

Haaf, Susanne et al. (2011): Measuring the correctness of double keying: Error classification and quality control in a large corpus of TEI annotated historical text (2011, forthcoming paper).

[1] Frequently updated information on the DTA "base format" can be found here: http://kaskade.dwds.de/dtaq/help/basisformat (description), http://kaskade.dwds.de/dtaq/help/basisformat_table (overview/table: elements within <text>)

Jurish, Bryan (2010): More than Words: Using Token Context to Improve Canonicalization of Historical German. In: Journal for Language Technology and Computational Linguistics (JLCL), vol. 25/1, 2010.

Unsworth, John (2011): Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI. In: Journal of the Text Encoding Initiative 1 (<http://jtei.revues.org/215>, 29. 8. 2011).

Bauman, Syd. “Interchange vs. Interoperability.” Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2 - 5, 2011. In *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies, vol. 7 (2011). doi:10.4242/BalisageVol7.Bauman01.