Richtlinien für die Volltextdigitalisierung

05/04/2024 21:36:07

FAQ Article Print

Category:	FAQ Clarin::Dienste	**Votes:**	1
State:	public (all)	**Result:**	0.00 %
Language:	de	**Last update:**	11:28:17 - 07/25/2016 (Europe/Berlin)

Keywords

Digitalisierung, Konvertierung

Symptom (public)

Ich verfüge über einen interessanten Text in Druckform, den ich als digitalen Volltext aufbereiten möchte. Gibt es bestimmte Richtlinien, die ich bei der Volltextdigitalisierung befolgen muss, damit dieser Text für eine bestimmte Software aus der Clarin-Umgebung „lesbar“ ist?

Problem (public)

Solution (public)

Einen Leitfaden für die Transkription von Texten der geschriebenen Sprache bietet das [1]Deutsche Textarchiv an. Der von CLARIN-D gewählte Annotationsstandard ist im Allgemeinen XML/TEI-P5. Speziell werden zwei spezifische Untermengen von TEI-P5 empfohlen:

1. das Basisformat des Deutschen Textarchivs ([2]DTABf) an der[3] Berlin-Brandenburgischen Akademie der Wissenschaften

2. das[4] IDS-XCES-Format des Instituts für Deutsche Sprache.

Beide Formate sind im [5]CLARIN-D-Benutzerhandbuch beschrieben (Abschnitt „Geschriebene Korpora“).

Für historische gedruckte Texte empfiehlt CLARIN-D das DTABf. Texte, die im DTABf vorliegen, können von verschiedenen CLARIN-D Zentren (Uni Leipzig, BBAW, IDS) verarbeitet und in die Repositorien von CLARIN-D aufgenommen werden. Dies beinhaltet die Konvertierung des DTABf-konformen TEI-Headers in das CLARIN-eigene Metadatenformat CMDI sowie die Konvertierung der annotierten Textdaten in das Text Corpus Format TCF, welches das Zugangsformat für die linguistischen Tools in WebLicht ist.

[1] http://www.deutschestextarchiv.de/doku/basisformat
[2] http://www.deutschestextarchiv.de/doku/basisformat
[3] http://www.bbaw.de/
[4] http://www1.ids-mannheim.de/start/
[5] http://www.clarin-d.de/de/hilfe/benutzerhandbuch