

PolMine Project

Spezialisierung	parlamentarische Sprachdaten <ul style="list-style-type: none">• Aufbereitung• Bereitstellung• Analyse• Verknüpfung (Data Linkage)
Modalität	schriftlich / konzeptionelle Schriftlichkeit
Akzeptierte Datenformate	PDF, TXT, XML, CWB
Ansprechpersonen	Andreas Blätte Stine Ziegler

Mission

Die digitale Verfügbarkeit politisch relevanter Texte eröffnet der Forschung neue Horizonte. Das PolMine Project der Universität Duisburg-Essen erschließt diese Texte und macht sie als Korpora wissenschaftlich nutzbar. Als Kompetenzzentrum für parlamentarische Sprachdaten stellt das Projekt Expertise und Werkzeuge zur Erstellung von Korpora auf Basis von Protokollen des politischen Diskurses bereit und leistet Unterstützung bei entsprechenden Vorhaben.

GermaParl

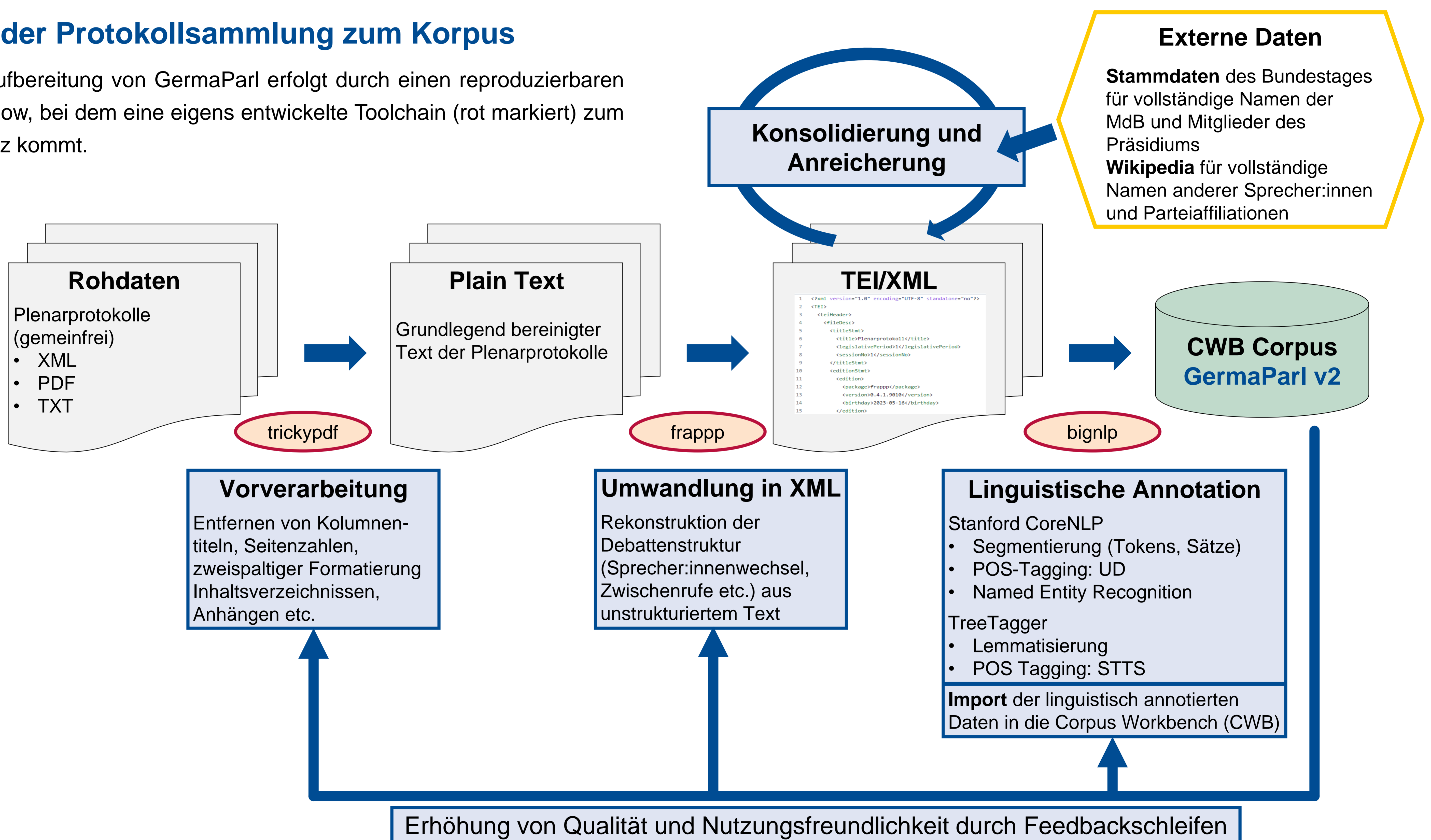
Eine Schlüsselressource des PolMine Project ist GermaParl als Korpus der Plenarprotokolle des Deutschen Bundestags. Am 23. Mai 2023 wurde GermaParl v2.0.0 veröffentlicht.

GermaParl v2.0.0 umfasst:

- ➔ 19 Legislaturperioden
- ➔ 72 Jahre (1949-2021)
- ➔ 4341 Protokolle
- ➔ 273 Millionen Token

Von der Protokollsammlung zum Korpus

Die Aufbereitung von GermaParl erfolgt durch einen reproduzierbaren Workflow, bei dem eine eigens entwickelte Toolchain (rot markiert) zum Einsatz kommt.



Zugang und Dissemination

GermaParl steht als XML und linguistisch annotiert und indiziert im CWB-Datenformat frei unter der Lizenz CC-BY-SA über Zenodo zur Verfügung.

Die Toolchain für die Reproduktion der Daten und für andere Aufbereitungsprojekte wird über GitHub-Repositoryen bereitgestellt.

Die wachsende Nutzungsgemeinschaft wird aktiv in die Weiterentwicklung der Ressourcen eingebunden, um sie nutzungsfreundlicher zu gestalten und die Qualität der Daten zu erhöhen.

Analyseumgebung und Tutorials

polmineR

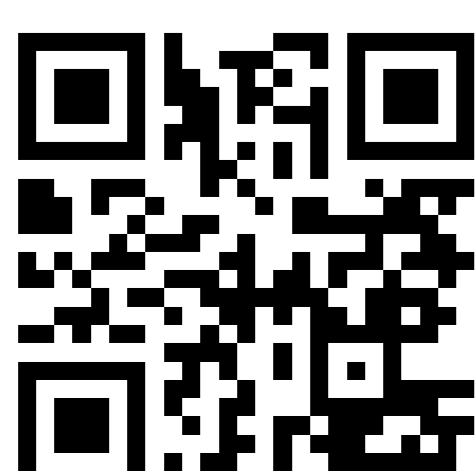
Mit *polmineR* wird für Korpora im CWB-Datenformat eine R-Analyseumgebung bereitgestellt, welche die interaktive Kombination qualitativer und quantitativer Analyseschritte ermöglicht.

Erste Schritte

Einen Einstieg in die sozialwissenschaftliche Arbeit mit Textdaten bieten die Online-Foliensätze „Using Corpora in Social Science Research“ sowie die Webinar-Reihe „Cookin' with GermaParl“.

Ausblick

- Kontinuierliche Pflege der Ressource (Qualitätsverbesserung, Aktualisierung)
- GermaParl v3: Verbesserung der Interoperabilität des Korpus durch die Adaption des ParlaMint-TEI-XML-Standards.
- Erweiterung der Annotationsschichten durch Evaluation und Nutzung alternativer NLP-Pipelines (insbesondere MONAPipe).
- Nutzung des Korpus als Beispiel- und Referenzdatensatz für die im KonsortSWD-Measure "Linking Textual Data" entwickelten Tools und Workflows.



andreas.blaette@uni-due.de
Twitter: @ablaette
stine.ziegler@uni-due.de
Twitter: @st_zie

