GESIS Leibniz Institute for the Social Sciences

PoliCorp BETA

POLLUX

Qualiservice data sharing

Pollux Political Corpora (PoliCorp)

https://demo-pollux.gesis.org/

Leibniz Association

Nina Smirnova, Ahsan Shahid, Philipp Mayr

02.06.2025

DFG Deutsche Forschungsgemeinschaft

About the project

POLLUX

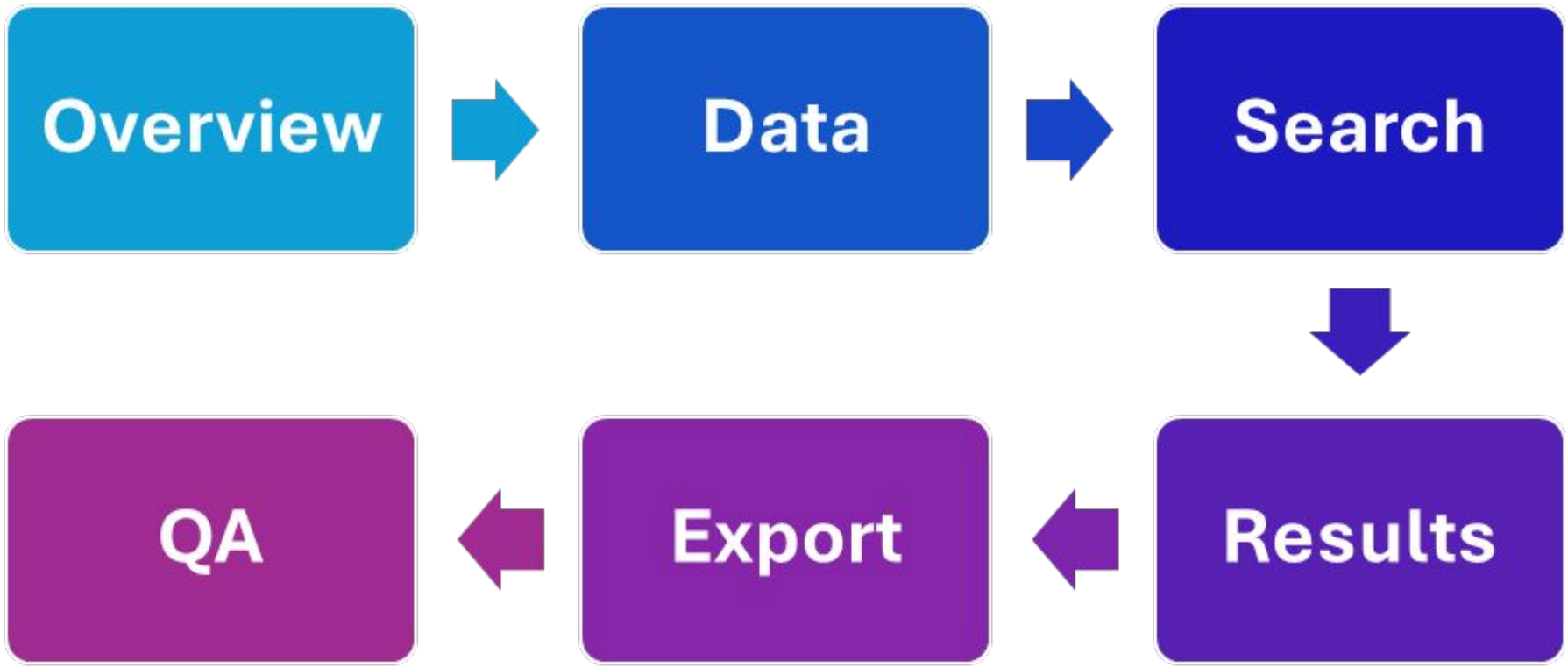
- Specialised Information Service (FID) for Political Science
- provide literature and the information infrastructure in the field of political science in Germany

GESIS Leibniz-Institut für Sozialwissenschaften

- provides fundamental research-based services for the social science
- part of the Leibniz Association and receives federal and state fundings

GESIS Leibniz Institute for the Social Sciences

Agenda



What is PoliCorp?

https://demo-pollux.gesis.org/

PoliCorp BETA

- an open resource for easy access to and analysis of processed political text data
- enables political scientists and multidisciplinary researchers to access **parliamentary speeches**
- no programming skills are required

What data are inside?

- Corpus of the protocols of the plenary debates published by the German Bundestag
- covers 76 years of debates
- 7 September 1949 to 11 February 2025
  - 7 September 1949 to 7 September 2021- Source GermaParl
    - Blaette, Andreas (2017): GermaParl. Corpus of Plenary Protocols of the German Bundestag.
  - 8 September 2021 to 11 Februar 2025 - Source Bundestag Open Data

source:https://www.stern.de/politik/deutschland/ordnungsrufe-im-bundestag-steigen---warum-das-mit-der-afd-zu-tun-hat-34853994.html

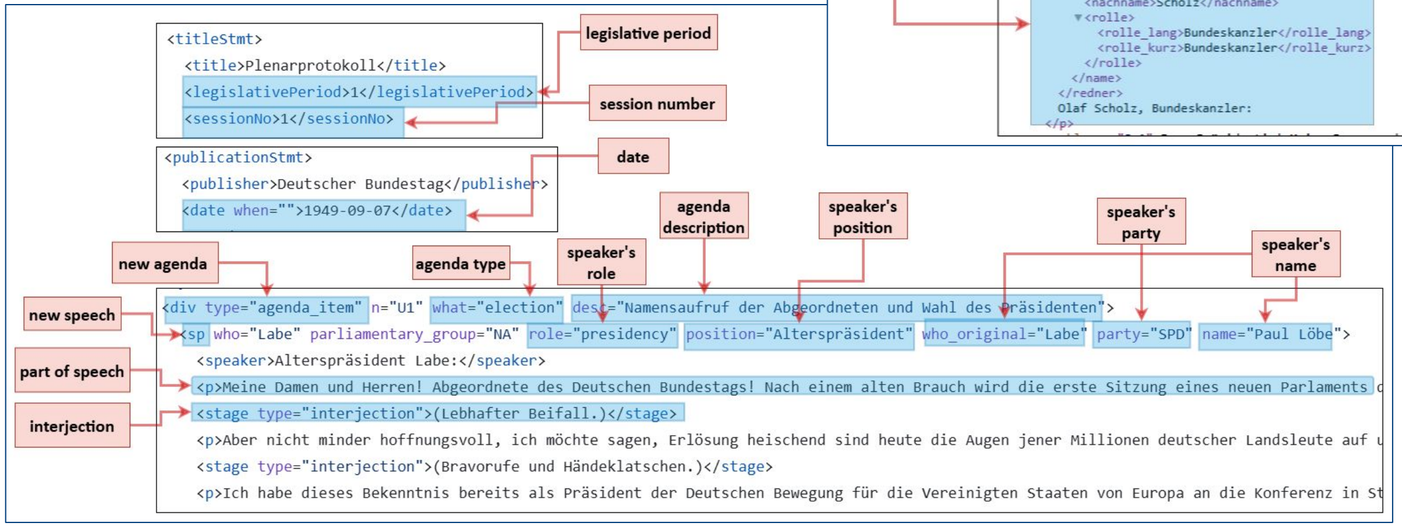
https://demo-pollux.gesis.org/germaparl_v3/details?id=8PhYaJMBWpaIjqYneewG

sources: https://www.uni-due.de/politik/germaparl.php
https://www.bundestag.de/resource/blob/710792/1412e3a264dedb70095c5662743aee3e/adler.png

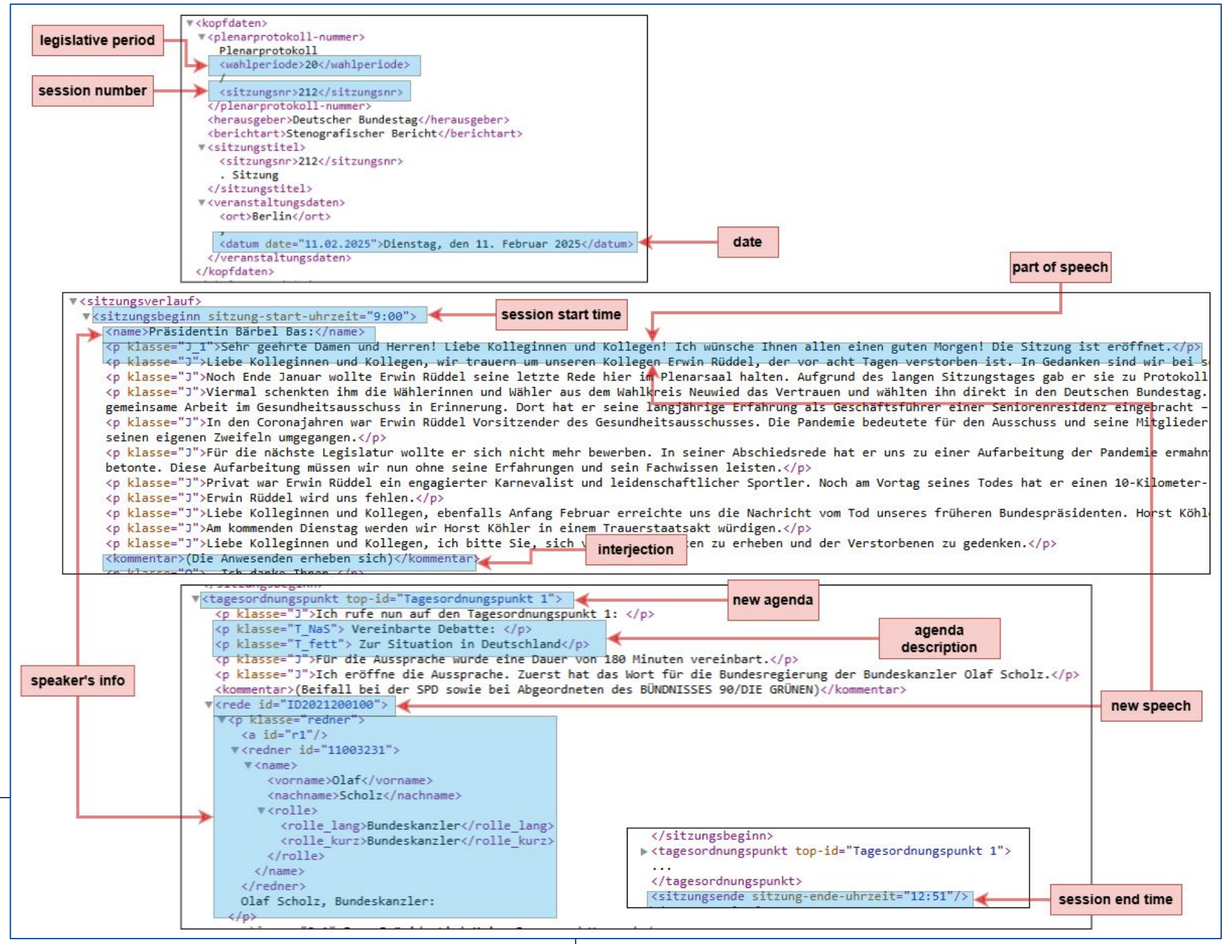
PoliCorp data processing

- converting XML to PostgreSQL
- splitting text into sentences

Raw GermaParl data (XML)



Raw Bundestag data (XML)



Demo

- additional data processing
  - parsing calls to order
  - cleaning and disambiguation of names of politicians and parties
- additional annotations
  - NER
    - https://huggingface.co/flair/ner-german
    - https://huggingface.co/flair/ner-german-legal
  - topic classification
    - https://huggingface.co/chkla/parlbert-topic-german
- technical background:
  - Elasticsearch, Node.js, Express

SPEAKER'S NAME	Christian Lindner
SPEAKER'S PARTY	FDP
SPEAKER'S ROLE	mp
URL TO THE SOURCE FILE	https://dserver.bundestag.de/btp/20/20212.xml
TOPIC	Macroeconomics
GERMAN NER	SPEECH: Nein, danke schön, ich will im Zusammenhang vortragen. Ich bin nämlich schon einen Gedanken weiter. Der Bundeskanzler hatte ja Steuererhöhungen angekündigt bei der Einkommensteuer, und den Wirtschaftsminister muss man in gleicher Weise verstehen. Nun, **Donald Trump**[PER] in den **USA**[LOC] will die effektive Belastung seiner Wirtschaft, so kündigte er an, auf 15 Prozent reduzieren. In **Deutschland**[LOC] ist die Belastung der Wirtschaft bei 30 Prozent. INTERJECTION: (Zuruf des Abg. Dr. **Ralf Stegner**[PER] [**SPD**[PARTY]]) SPEECH: Die Zölle werden gefürchtet. Tatsächlich aber haben wir auch einen Steuerwettbewerb mit den **USA**[LOC]. Weil unser Land nicht mehr doppelt so gut ist wie die **Vereinigten Staaten von Amerika**[LOC], können wir auch nicht doppelt so teuer sein. INTERJECTION: (Beifall bei der **FDP**[PARTY] sowie bei Abgeordneten der **CDU**[PARTY]/**CSU**[PARTY] – **Britta Haßelmann**[PER] [**BÜNDNIS 90/DIE GRÜNEN**[PARTY]]: Was will denn die **FDP**[PARTY] eigentlich?) SPEECH: Statt wie die **SPD**[PARTY] – ich glaube, die **Grünen**[ORG] wollen das auch – Subventionen und Investitionsprämien für alle Unternehmen zu zahlen, und damit auch an die nicht erfolgreichen Unternehmen, muss die Steuerlast für alle gesenkt werden, damit an der Stelle, wo Erfolg sichtbar ist, zusätzliches Kapital weitere Schritte finanzieren kann.

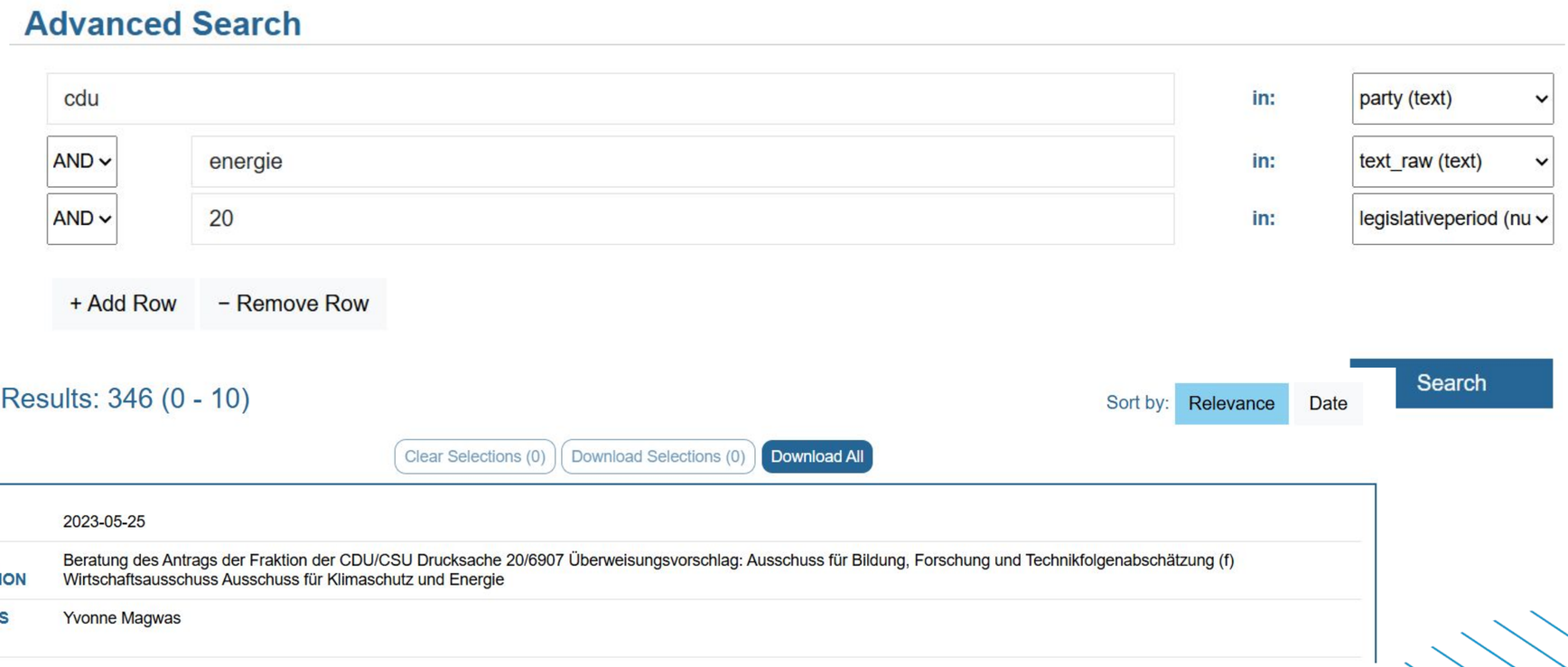
Search & Advanced Search

Fields:

- Date
- Legislative Period
- Politician Name
- Keyword in speech text
- Topic
- Party

Boolean logic:

- AND, OR, and NOT



❖ Generation of subcorpora

Data description:

https://github.com/kalawinka/policorp

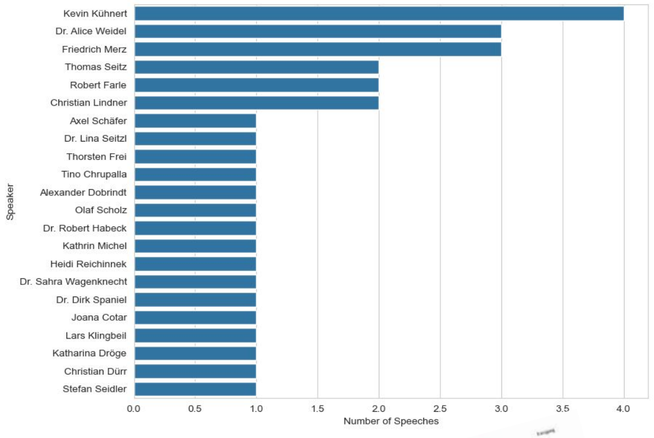
Use cases

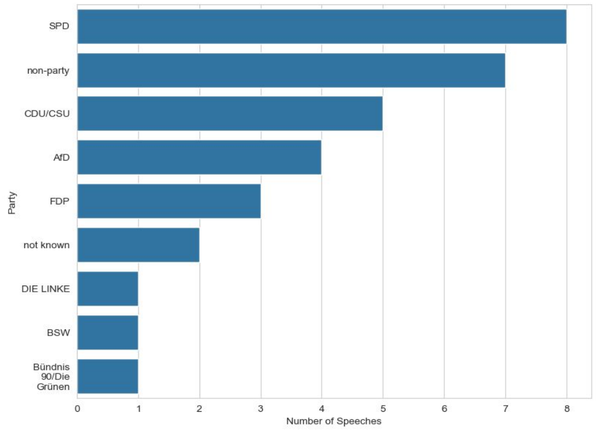
https://demo-pollux.gesis.org/

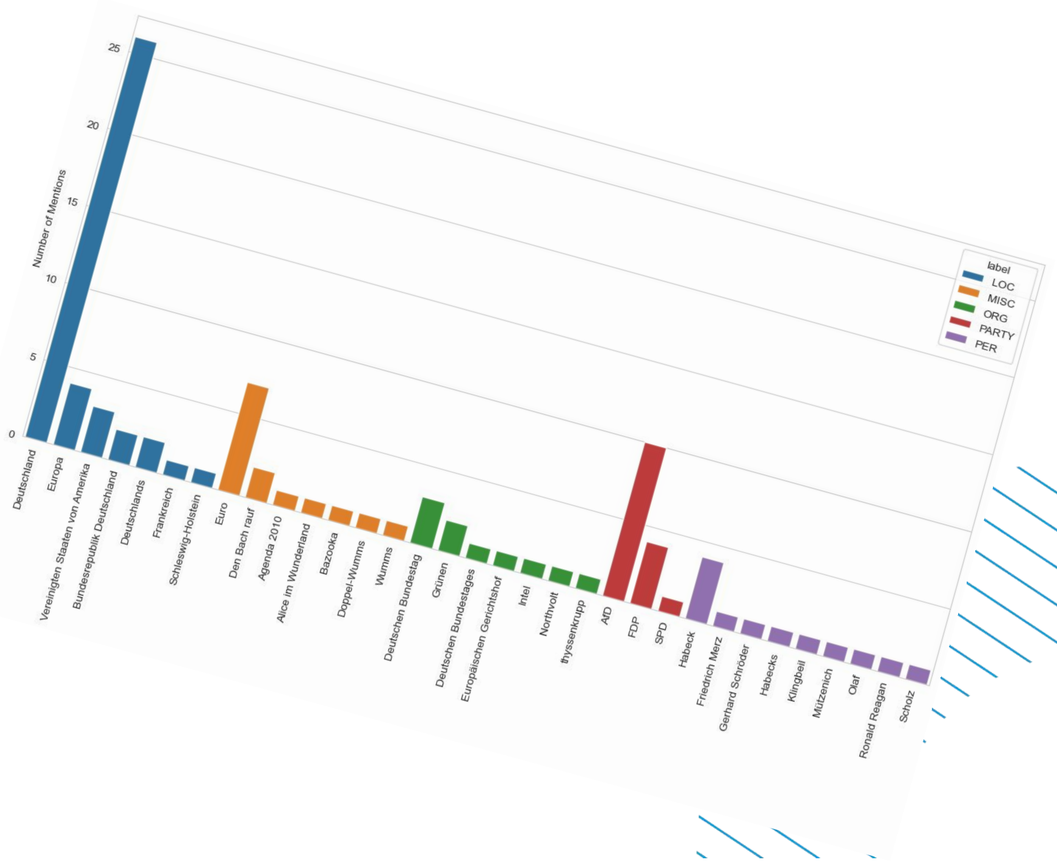
- Last Bundestag debate before the election
  - 2025-02-11
- Calls to Order issued in LP 20
- All speeches of Robert Habeck concerning Environment
- All speeches of Olaf Scholz or Christian Lindner
- All speeches with keyword Migration, but not from speeches of party DIE LINKE

Example of data analysis with python

https://github.com/kalawinka/policorp/blob/main/examples/example_data_analysis_policorp.ipynb







Scheduled:

- Regular updates of PoliCorp with new plenary debates
- Annotation of toxic/ negative speeches
- Annotation of calls to order
- Integrated data analysis
- Integration of other corpora
  - i.e., StateParl

Image is generated with GPT-4o

GESIS Leibniz Institute for the Social Sciences

Takeaway

- → advanced search over whole corpus
- → regular updates
- → generation of subcorpora for further analysis
- → free download options
- → experimental features

Thank you!

Leibniz
Association