

# Current Topics in Digital Philology

## The openLegislature project

Dan Häberlein, Peggy Lucke, J. Nathanael Philipp, Alexander  
Richter

Universität Leipzig

# Gliederung

- 1 Introduction
  - Korpus
  - Questions
- 2 Methoden
  - Vorgehensweise
  - Ergebnisse
- 3 Outlook
  - Anstehende Aufgaben
  - Ziele
  - Projektumfang
- 4 Give us your money!

# Informations

## Plenary Protocols from Bundestag

- stenographic reports in PDF
- open to the public
- siehe [**bundestag** ]
- size of corpus circa 10GB → more than 3900 PDF

# Questions to the information in the corpus

## Statistic:

- How many speakers are in one legislative period/total?
- How many speeches from one party/speaker?

## Keyword-search:

- Which speaker spoke to a special topic?

## Why this questions?

We want more transparency! The answers are there, but too difficult to reach for all other people. That will be changed!

# Gliederung

- 1 Introduction
  - Korpus
  - Questions
- 2 Methoden
  - Vorgehensweise
  - Ergebnisse
- 3 Outlook
  - Anstehende Aufgaben
  - Ziele
  - Projektumfang
- 4 Give us your money!

# Methoden 1

## Preprocessing:

- Lemmatizing
- ggfs. Part of Speech Analyse
- Vergleichen von N-Grammen zur Ähnlichkeitsüberprüfung
- Keine Stopwortentfernung, da dadurch Informationen verloren gehen!

# Methoden 2

## IR Methoden:

- Kookkurrenz auf verschiedenen Ebenen (Reden per Partei, Redner, Gesamt)
- Clustering

## IR Algorithmen

- Ähnlichkeitsmaße berechnen und vergleichen
- Topic Modell / LDA

*"Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.", siehe [latent ]*

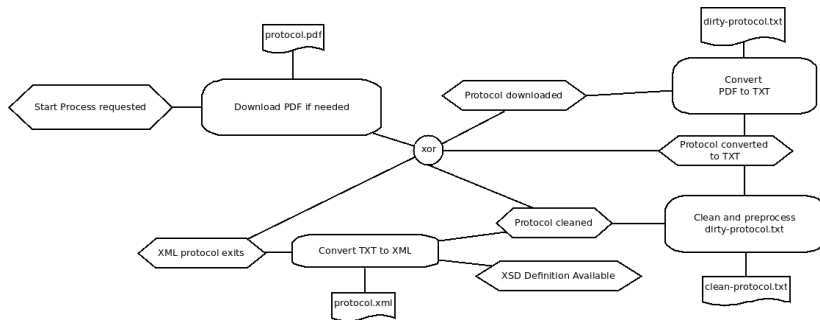
# Vorläufige Ergebnisse

- Unstrukturierte Textdateien liegen vor (PDF / TXT)
- Semistrukturierte XML Dateien aus TXT erzeugt
- Metadatenbank angelegt
- Erste einfache XPath Anfragen auf Dateien realisiert



# Architektur Prozess Datenextraktion und -aufbereitung

- Nutzung des Listener Patterns [javainse19]
- Verwendung der Github-Library Async [async] zur einfachen Erstellung Nebenläufiger Prozessketten



# Gliederung

- 1 Introduction
  - Korpus
  - Questions
- 2 Methoden
  - Vorgehensweise
  - Ergebnisse
- 3 Outlook
  - Anstehende Aufgaben
  - Ziele
  - Projektumfang
- 4 Give us your money!

# Ausblick I: nächste Schritte

## Nächste Schritte:

- erweitern der Metadaten-Datenbank mit Metadaten:
  - aus bestehenden XML Files
  - aus zusätzlichen Quellen (z.B. Sitzverteilungen)
- Überführen der Strukturierten Daten im XML-Format in einen Document Store
- Analyse der Daten:
  - Clustering (Top-Down, Bottom-Up)
  - LDA (Latend Dirichlet Allocation)

## Ausblick II: Ziele Vorlesungszeit

Ziele bis Ende Vorlesungszeit:

- XML-Daten in Document-Store ablegen
- Metadaten-Datenbank mit weiteren Metadaten erweitern
- analysieren der Daten mittels mind. zwei Clustering-Verfahren
- Cluster mit wahrscheinlich gleichen Schreiber (aber nicht Redner) finden und darstellen

# Ausblick III: Ziele Semester

Ziele bis Ende Semester:

- Analyse mittels LDA
- Visualisierung der Ergebnisse der LDA-Analysen
- weitere Cluster-Verfahren nutzen
- alle (sinnvollen) Ergebnisse vereinen und darstellen
- Untersuchung warum manche Analysen fehlerhafte/schlechte Ergebnisse lieferten

# Ausblick IV: Erweiterbarkeit

Erweiterbarkeit wenn uns Zeit bleibt:

- zusätzlichen Metadaten-Quellen auffinden und in die bestehende Metadaten-Datenbank überführen
- weitere Metadaten erzeugen (Bsp. POS-Tagging, N-Gramme und Kookkurrenzen)
- Analyse-Verfahren erweitern
  - andere Cluster-Algorithmen
  - LDA mit anderen Parametern
  - LDA mit anderen Features
- Visualisieren der Ergebnisse

# Ausblick V: Einschränkungen I

Einschränkungen wenn wir nicht alle Ziele schaffen:

- weniger Analyseverfahren nutzen (Bsp. nur ein Clusteringverfahren)
- Datenbereinigung verkürzen
- weniger Metadaten als Feature nutzen

# Ausblick VI: Einschränkungen II

## Wichtigste Ziele:

- Daten in Datenbank strukturiert ablegen
- Clustering-Verfahren auf unsere Daten anwenden
- Interpretation der Ergebnisse

## Neue Ziele:

- Einfaches Query Interface (ähnlich Google) um Nutzern Zugang zu Daten zu geben
- Query Interface als Webanwendung



# Gliederung

- ① Introduction
  - Korpus
  - Questions
- ② Methoden
  - Vorgehensweise
  - Ergebnisse
- ③ Outlook
  - Anstehende Aufgaben
  - Ziele
  - Projektumfang
- ④ Give us your money!

# Give us your money!

We try to achieve reproducible and professional results. Our project could be really interesting in the following sence:

- History / Political Science
- Educational Purposes
- Parties

We could also make this dataset that we just created more human accessible by developping an easy user interface (something like google). Our work would contribute to more transparent german politics, in which every citizen has the power to validate and measure politicians by there speeches.