

**Titel: Verfahrensorientierte Datensätze aus historischen Schriftsätzen erstellen. Automatische Translation von Informationen aus Parlamentsprotokollen der britischen Parlamente.**

*Autor:* Sebastian Fath

*Abstract:*

Ein jedes Parlament dokumentiert irgendwo was es wann wie tut. Normalerweise geschieht dies in Niederschriften, den sogenannten Protokollen. Heutzutage stellen die meisten Parlamente ihre Unterlagen in mehr oder weniger gut verarbeitbarer Struktur digital zur Verfügung. Als goldener Standard lässt sich hier schnell das “Dokumentations- und Informationssystem für Parlamentsmaterialien” (DIP) des deutschen Bundestags anbringen<sup>1</sup>, dessen Ansatz der Dokumentation eines jedem “Verfahren” vor dem Bundestag eine effiziente Analyse der Geschehnisse im Bundestag ermöglicht und dem Suchenden eine geraume Menge an Kontext vermag direkt mit darzustellen.

Dem gegenüber stehen eine Unzahl an Parlamentsprotokollen vor dem 21. Jahrhundert, die nirgends derartig abrufbar sind. In diesen ist es oftmals ein aufwändiger Prozess, den Entstehungsprozess eines einzelnen Gesetz zu rekonstruieren. Der Vorteil eines Datensatzes, wie er dem DIP zugrunde liegt, ist offensichtlich: Die Dokumentation der Verfahren vor dem Parlament erlauben es Forschern einfacher die tatsächliche Entstehungsgeschichte *eines* Gesetzes nachzuvollziehen. Dies ist nicht selten unwichtig, unter anderem bei der Bestimmung des Telos. Um den Prozess der Analyse und Durchsuchung der Parlamentsprotokolle zu vereinfachen wäre es also von Vorteil, wenn auch aus älteren (oder nicht gut vorbereiteten) Protokollen ein Datensatz der auftretenden Verfahren erstellt wird.

Um sich diesem Ziel zu nähern will dieses Projekt versuchen sich der Erzeugung der nötigen Daten für einen derartigen Datensatz sukzessive mit einer Reihe an gängigen Methoden anzunähern. Zunächst soll versucht werden mithilfe von Kategorisierungsalgorithmen, mglw. OCR, etc. einzelne Abschnitte des Protokoll zu Handlungen zugeordnet werden. Diese sollen dann gespeichert werden und mit lesbaren Metadaten (wie Sitzungsdatum etc.) angereichert werden. In einem zweiten Schnitt sollen dann einzelne Abschnitte (Bzw. Handlungen) mithilfe von NLP idealerweise Auskunft geben können über ihren Inhalt, zentrale Akteure, etc. und somit Daten hervorbringen, die in einem weiteren Schritt geeignete Verknüpfungen zwischen den einzelnen Abschnitten erzeugen können um aus den “Handlungen” einen Datensatz an Verfahren zu erstellen. Zuletzt stehen Versuche offen mithilfe von Mitteln der Diskursanalyse<sup>2</sup> für das finden historisch

---

<sup>1</sup>Für einen Eindruck empfiehlt sich schlicht die Betrachtung der API-Spec unter <https://search.dip.bundestag.de/api/v1/swagger-ui/>

<sup>2</sup>Wie vorgestellt in Wächter, Christian, 2023. Capturing Discourse through the Digital Lens: Towards a Framework for the Analysis of Pro-democratic Discourse in the Weimar Republic. In: Florentina ARMASELU und Andreas FICKERS (Hrsg.), Zoomland [online]. De Gruyter. S. 43–76. [Zugriff am: 28 November 2024]. ISBN 978-3-11-131777-9. Verfügbar unter: <https://www.degruyter.com/document/doi/10.1515/9783111317779-003/html>

interessanter Parameter und Entwicklungen.

Dieses Projekt beruht auf der Verfügbarkeit der Parlamentsprotokolle. Während sich alte Parlamentsprotokolle eines jeglichen Parlaments anbieten, muss sich in diesem Rahmen auf eine “Art” von Protokoll beschränkt werden. Die Unterschiede in Sprache, Layout, etc. unterschiedlicher Parlamente würde eine zeitlich zu große Herausforderung stellen. Aus persönlichem Interesse sowie Verfügbarkeit der Daten habe ich mich dazu entschlossen, in diesem Projekt die Protokolle der Britischen Parlamente zu bearbeiten. Diese sind zwar noch nicht verfügbar als Verfahrensbasierter Datensatz, doch sind sie gut erforscht und bilden so eine gute Grundlage um ohne zu viel Aufwand den Datensatz auf eine Mindestqualität prüfen zu können. Außerdem sind sie einfach zugänglich. Materialien und Daten finden sich auf dem Webauftritt des Parlaments sowie dem unter DFG-Lizenz zugänglichen Corpus “U.K. Parliamentary Papers”

Eine Grundlegende Analyse der generierten Daten soll dann anhand von Performance Indikatoren vorgenommen werden. Es wird im Rahmen dieses Projekt nicht möglich sein, eine umfassende Akkuranz der generierten Daten zu bestimmen. Stattdessen sollen Abgleiche mit externen Datensätzen (beispielsweise mit [legislation.gov.uk](http://legislation.gov.uk)) und Punktweise Kontrolle der Daten einen ersten Eindruck der Nutzbarkeit und Zuverlässigkeit vermitteln.

*Anzahl Wörter:* 490