Corpus der Drucksachen des Deutschen Bundestages

(CDRS-BT)

Codebook

Version 2021-04-02

DOI: 10.5281/zenodo.4643066

Titel	Corpus der Drucksachen des Deutschen Bundestages
Abkürzung	CDRS-BT
Autor	Seán Fobbe
Version	2021-04-02
Download	https://doi.org/10.5281/zenodo.4643066
Lizenz	CC0 1.0 Universal

Zitiervorschlag

Seán Fobbe (2021). Corpus der Drucksachen des Deutschen Bundestages (CDRS-BT). Version 2021-04-02. Zenodo. DOI: 10.5281/zenodo.4643066.

Digital Object Identifier (DOI): Concept DOI und Version DOI

Soweit nicht anders angegeben ist die DOI immer eine »Version DOI« und bezieht sich nur auf eine bestimmte Version des Datensatzes. Sie verweist daher nur auf Version 2021-04-02. Für das Gesamtkonzept dieses Datensatzes steht eine »Concept DOI« zur Verfügung, die auf der Zenodo-Seite jeder Version unter »Cite all versions?« zu finden ist. Sie lautet 10.5281/zenodo.4643065. Die »Concept DOI« verlinkt immer die aktuellste Version.

Urheberrecht

Disclaimer

Dieser Datensatz ist eine private wissenschaftliche Initiative und steht in keiner Verbindung zum Deutschen Bundestag oder anderen amtlichen Stellen der Bundesrepublik Deutschland.

Inhaltsverzeichnis

1 Einführung

Der **Deutsche Bundestag** ist das Parlament der Bundesrepublik Deutschland. Der Bundestag und der Bundesrat bilden gemeinsam die Legislative auf Bundesebene und somit die primäre Quelle für das deutsche Bundesrecht. Beide fungieren auch als einige der wichtigsten öffentlichen und nicht-öffentlichen Foren für gesellschaftliche Debatten.

Der **Corpus der Drucksachen des Deutschen Bundestages (CDRS-BT)** ist eine digitale Zusammenstellung von allen Drucksachen des Deutschen Bundestages der 1. bis 18. Wahlperiode. *Drucksachen* sind schriftliche Dokumente, welche die Beratungen des Bundestages vor- und nachbereiten und als Verhandlungsgegenstand auf die Tagesordnung des Bundestages gesetzt werden können (§ 75 Geschäftsordnung des Bundestages). Die inhaltliche Bandbreite ist hierbei sehr weit und umfasst beispielsweise Gesetzentwürfe, Beschlussvorlagen, kleine Anfragen, Antworten der Bundesregierung, Berichte von Untersuchungsausschüssen und Wahlvorschläge.

Dem **Bundesrecht** kommt im Normengefüge der Bundesrepublik Deutschland herausragende Bedeutung zu. Zwar sind die Länder gemäß Art. 30, 70 GG primär für die Gesetzgebung zuständig, im Katalog der Art. 71 ff GG sind aber derart viele Kompetenzen dem Bund zugewiesen, dass das Bundesrecht praktisch jedes rechtliche Problem in der Bundesrepublik dominiert. Ausnahmen sind in der Regel nur die Bereiche innere Sicherheit, Bildung und Kultur, die weitgehend in der Hand der Bundesländer verblieben sind. Aber auch in diesen Bereichen finden sich Regelungen des Bundes. Beispiele dafür sind manche Regelungen des Bundespolizeigesetzes (BPolG) oder das Kulturgutschutzgesetz (KGSG).

Die quantitative Analyse von politischen Texten ist mittlerweile fester Bestandteil des Forschungsprogramms der Politikwissenschaften, steht in den Rechtswissenschaften aber noch ganz am Anfang. Drucksachen spielen für die Gesetzgebung eine zentrale Rolle, wurden aber von Rechtswissenschaftler:innen bisher kaum systematisch untersucht. Die einzige frei verfügbare Sammlung (»Every Single Word«-Korpus) wurde dementsprechend auch von Politikwissenschaftler:innen (Kroeber und Remschel 2020) veröffentlicht.[1]

Ein weiterer einschlägiger Korpus wurde möglicherweise von der Universität Siegen erstellt, war aber im April 2021 seit über einem Jahr offline.[2] In Anbetracht der flächendeckenden Verfügbarkeit von hochwertigen wissenschaftlichen Repositorien ist ein Datensatz mit einer solchen Erreichbarkeitslücke keine vertretbare Grundlage für reproduzierbare Forschung.

In einem funktionierenden Rechtsstaat muss die Gesetzgebung öffentlich, transparent und nachvollziehbar sein. Im 21. Jahrhundert bedeutet dies auch, dass sie quantitativen Analysen zugänglich sein muss. Der Erstellung und Aufbereitung des Datensatzes liegen daher die Prinzipien der allgemeinen Verfügbarkeit durch Urheberrechtsfreiheit, strenge Transparenz und vollständige wissenschaftliche Reproduzierbarkeit zugrunde. Die FAIR-Prinzipien (Findable, Accessible, Interoperable and Reusable) für freie wissenschaftliche Daten inspirieren sowohl die Konstruktion, als auch die Art der Publikation.[3]

[1] Remschel, T. and Kroeber, C. (2020). Every Single Word: A New Data Set Including All Parliamentary Materials Published in Germany. Government and Opposition, 1-20. https://www.doi.org/10.1017/gov.2020.29; Kroeber, C. and Remschel, T., 2020, Every Single Word - A New Dataset Including All Parliamentary Materials Published in Germany. https://doi.org/10.7910/DVN/7EJ1KI. Harvard Dataverse. V2.

[2] https://diskurslinguistik.net/korpus-repository/

[3] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

2 Nutzung

Die Daten sind in offenen, interoperablen und weit verbreiteten Formaten (CSV, TXT, XML) veröffentlicht. Sie lassen sich grundsätzlich mit allen modernen Programmiersprachen (z.B. R, Python), sowie mit grafischen Programmen nutzen.

Wichtig: Nicht vorhandene Werte sind sowohl in den Dateinamen als auch in der CSV-Datei mit »NA« codiert.

2.1 CSV-Dateien

Am einfachsten ist es die **CSV-Dateien** einzulesen. CSV[4] ist ein einfaches und maschinell gut lesbares Tabellen-Format. In diesem Datensatz sind die Werte Komma-separiert. Jede Spalte entspricht einer Variable, jede Zeile einer Drucksache. Die Variablen sind unter Punkt 5 genauer erläutert.

Zum Einlesen empfehle ich für **R** dringend das package **data.table** (via CRAN verfügbar). Dessen Funktion **fread()** ist etwa zehnmal so schnell wie die normale **read.csv()**-Funktion in Base-R. Sie erkennt auch den Datentyp von Variablen sicherer. Beispiel:

```
library(data.table)
dt <- fread("./filename.csv")
```

2.2 TXT-Dateien

Die **TXT-Dateien** inklusive Metadaten können zum Beispiel mit **R** und dem package **readtext** (via CRAN verfügbar) eingelesen werden. Beispiel:

```
library(readtext)
txt <- readtext("./*.txt",
                docvarsfrom = "filenames",
                docvarnames = c("datensatz",
                                "organ",
                                "dokumentart",
                                "wahlperiode",
                                "nummer_dok",
                                "datum",
                                "xml"),
                dvsep = "_",
                encoding = "UTF-8")
```

2.3 XML-Dateien

Das Einlesen der **XML-Rohdaten** ist technisch anspruchsvoller als das Einlesen der CSV- oder TXT-Varianten. Da die XML-Dateien bis zur 18. Wahlperiode keine besonders komplexe Datenstruktur aufweisen, wird sich in den meisten Fällen kein Mehrwert gegenüber dem CSV-Format ergeben. Falls Sie dennoch die XML-Dateien nutzen möchten, lesen Sie bitte die Document Type Definition (DTD) genau und greifen Sie ggf. auf den im Source Code zur Verfügung gestellten XML Parser zurück.

[4] Das CSV-Format ist in RFC 4180 definiert, siehe https://tools.ietf.org/html/rfc4180

3 Konstruktion

3.1 Beschreibung

Der **Corpus der Drucksachen des Deutschen Bundestages (CDRS-BT)** ist eine digitale Zusammenstellung von allen Drucksachen des Deutschen Bundestages, die auf dessen Open Data Portal in maschinenlesbaren XML-Dateien veröffentlicht wurden. Die derzeitige Sammlung beschränkt sich auf die Drucksachen der 1. bis 18. Wahlperiode.

Der Stichtag des Abrufs für jede Version entspricht exakt der Versionsnummer.

Zusätzlich zu den aufbereiteten maschinenlesbaren Formaten (CSV und TXT) sind die XML-Rohdaten enthalten, damit Analyst:innen gegebenenfalls ihre eigene Konvertierung vornehmen können. Die XML-Rohdaten wurden inhaltlich nicht verändert.

3.2 Datenquelle

Datenquelle	Vollzitat
Primäre Datenquelle	https://www.bundestag.de/services/opendata

3.3 Sammlung der Daten

Die Daten wurden vollautomatisiert gesammelt und mit Abschluss der Verarbeitung kryptographisch signiert. Das Open Data Portal des Bundestages ist ausdrücklich für die vollautomatisierte Datensammlung freigegeben (»offenen Daten können zur maschinellen Weiterverarbeitung genutzt werden«[5]). Der Abruf geschieht ausschließlich über TLS-verschlüsselte Verbindungen.

3.4 Source Code und Compilation Report

Der gesamte Source Code — sowohl für die Erstellung des Datensatzes, als auch für dieses Codebook — ist öffentlich einsehbar und dauerhaft erreichbar im wissenschaftlichen Archiv des CERN hier hinterlegt: https://doi.org/10.5281/zenodo.4643068

Mit jeder Kompilierung des vollständigen Datensatzes wird auch ein umfangreicher **Compilation Report** in einem attraktiv designten PDF-Format erstellt. Der Compilation Report enthält den vollständigen Source Code, dokumentiert relevante Rechenergebnisse, gibt sekundengenaue Zeitstempel an und ist mit einem klickbaren Inhaltsverzeichnis versehen. Er ist zusammen mit dem Source Code hinterlegt. Wenn Sie sich für Details der Herstellung interessieren, lesen Sie diesen bitte zuerst.

[5] https://www.bundestag.de/services/opendata

3.5 Einschränkungen

Nutzer sollten folgende wichtige Einschränkungen beachten:

1. Der Datensatz enthält nur das, was der Bundestag auch tatsächlich veröffentlicht (*publication bias*).
2. Es werden nur XML-Dateien abgerufen (*file type bias*).
3. Einige wenige XML-Dateien waren fehlerhaft und konnten nicht ausgewertet werden (*error bias*).
4. Die Sammlung beschränkt sich zunächst auf die 1. bis 18. Wahlperiode (*temporal bias*). Die Frequenztabellen geben hierzu genauer Auskunft. Weitere Wahlperioden werden in Zukunft berücksichtigt.

3.6 Urheberrechtsfreiheit von Rohdaten und Datensatz

An Drucksachen besteht gem. § 5 UrhG kein Urheberrecht, da sie amtliche Werke sind. § 5 UrhG ist auf amtliche Datenbanken analog anzuwenden (BGH, Beschluss vom 28.09.2006, I ZR 261/03, »Sächsischer Ausschreibungsdienst«).

Alle eigenen Beiträge (z.B. durch Zusammenstellung und Anpassung der Metadaten) und damit den gesamten Datensatz stelle ich gemäß einer *CC0 1.0 Universal Public Domain Lizenz* vollständig urheberrechtsfrei.

3.7 Metadaten

3.7.1 Allgemein

Die Metadaten wurden fast ausschließlich aus dem Inhalt der XML-Dateien extrahiert bzw. berechnet. Der volle Satz an Metadaten ist nur in den CSV-Dateien enthalten. Alle hinzugefügten Metadaten sind zusammen mit dem Source Code vollständig maschinenlesbar dokumentiert.

Die Dateinamen der TXT-Dateien enthalten den Namen des Datensatzes, den Namen des Organs, die Art des Dokuments, die Wahlperiode, die laufende Nummer des Dokuments, das Datum der Sitzung (Langform nach ISO-8601, d.h. YYYY-MM-DD) und den Namen der XML-Datei aus der sie stammen.

3.7.2 Schema für die TXT-Dateinamen

```
[datensatz]_[organ]_[dokumentart]_[wahlperiode]_[nummer_dok]_[datum]_[xml]
```

3.7.3 Beispiel eines Dateinamens

```
CDRS-BT_Bundestag_Drucksache_1_16_1949-09-14_0100016.txt
```

3.8 Qualitätsprüfung

Die möglichen Werte der jeweiligen Variablen wurden durch Frequenztabellen und Visualisierungen auf ihre Plausibilität geprüft. Insgesamt werden zusammen mit jeder Kompilierung eine Vielzahl Tests zur Qualitätsprüfung durchgeführt. Alle Ergebnisse der Qualitätsprüfungen sind aggregiert im Compilation Report zusammen mit dem Source Code und einzeln im Archiv »ANALYSE« zusammen mit dem Datensatz veröffentlicht.

4 Varianten und Zielgruppen

Dieser Datensatz ist in unterschiedlichen Varianten verfügbar, die sich jeweils an verschiedene Zielgruppen richten. Zielgruppe sind nicht nur quantitativ forschende Politik- und Rechtswissenschaftler:innen, sondern auch traditionell arbeitende Forscher:innen. Idealerweise müssen quantitative Methoden ohnehin immer durch qualitative Interpretation, Theoriebildung und kritische Auseinandersetzung verstärkt werden (*mixed methods*).

Lehrende werden zudem von den vorbereiteten Tabellen und Diagrammen besonders profitieren, die Zeit im universitären Alltag sparen und bei der Erläuterung der Charakteristika der Daten hilfreich sein werden. Alle Tabellen und Diagramme liegen auch als separate Dateien vor um sie einfach z.B. in Präsentations-Folien oder Handreichungen zu integrieren.

Variante	Zielgruppe und Beschreibung
CSV_Datensatz	**Legal Tech/Quantitative Forschung**. Diese CSV-Datei ist die für statistische Analysen empfohlene Variante des Datensatzes. Sie enthält den Volltext aller Dokumente, sowie alle in diesem Codebook beschriebenen Metadaten.
CSV_Metadaten	**Legal Tech/Quantitative Forschung**. Wie die andere CSV-Datei, nur ohne den Inhalt der Dokumente. Sinnvoll für Analyst:innen, die sich nur für die Metadaten interessieren und Speicherplatz sparen wollen.
TXT	**Traditionelle Forschung**. Die TXT-Dateien stellen einen Kompromiss zwischen den Anforderungen quantitativer und qualitativer Forschung dar und können sowohl als Lesefassung, als auch als Grundlage für quantitative Analysen benutzt werden. Die Dateinamen sind so konzipiert, dass sie auch für die traditionelle qualitative Arbeit einen erheblichen Mehrwert bieten. Im Vergleich zu den CSV-Dateien enthalten die Dateinamen nur einen reduzierten Umfang an Metadaten, um Kompatibilitätsprobleme unter Windows zu vermeiden und die Lesbarkeit zu verbessern.
XML	**Legal Tech/Quantitative Forschung**. Die XML-Rohdaten, so wie sie vom Bundestag veröffentlicht wurden. In der Regel nur für Replikationen von Interesse.
ANALYSE	**Alle Lehrenden und Forschenden**. Dieses Archiv enthält alle während dem Kompilierungs- und Prüfprozess erstellten Tabellen (CSV) und Diagramme (PDF, PNG) im Original. Sie sind inhaltsgleich mit den in diesem Codebook verwendeten Tabellen und Diagrammen. Das PDF-Format eignet sich besonders für die Verwendung in gedruckten Publikationen, das PNG-Format besonders für die Darstellung im Internet. Analyst:innen mit fortgeschrittenen Kenntnissen in R können auch auf den Source Code zurückgreifen. Empfohlen für Nutzer:innen die einzelne Inhalte aus dem Codebook für andere Zwecke (z.B. Präsentationen, eigene Publikationen) weiterverwenden möchten.

5 Variablen

5.1 Datenstruktur

```
## Classes 'data.table' and 'data.frame':   131835 obs. of  20 variables:
##  $ doc_id         : chr  "0100001.xml" "0100002.xml" "0100003.xml" "0100004.
    xml" ...
##  $ wahlperiode    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dokumentart    : chr  "DRUCKSACHE" "DRUCKSACHE" "DRUCKSACHE" "DRUCKSACHE"
    ...
##  $ nummer_original: chr  "01/1" "01/2" "01/3" "01/4" ...
##  $ drs_typ        : chr  "Unterrichtung" "Antrag" "Antrag" "Antrag" ...
##  $ k_urheber      : chr  "Bundestag" "Fraktion der SPD" "Fraktion der SPD" "
    Fraktion der SPD" ...
##  $ p_urheber      : chr  "NA" "Ollenhauer, Erich" "Ollenhauer, Erich" "
    Ollenhauer, Erich" ...
##  $ datum          : IDate, format: "1949-10-07" "1949-09-07" ...
##  $ titel          : chr  "Alphabetisches Verzeichnis der Mitglieder des
    Bundestags" "Demontagen" "Groß-Berlin" "Vorläufiger Sitz der leitenden
    Bundesorgane" ...
##  $ text           : chr  "Drucksache Nr. i\nDeutscher Bundestag\n1.
    Wahlperiode\n1949\nAlphabetisches Verzeichnis\nder Mitglieder des Bun"| __
    truncated__ "Drucksache Nr. 2\nDeutscher Bundestag\n1. Wahlperiode\n1949\
    nAntrag\nder Fraktion der SPD\nbetr. Demontagen\nDe"| __truncated__ "
    Drucksache Nr. 3\nDeutscher Bundestag 1. Wahlperiode\n1949\nAntrag\nder
    Fraktion der SPD\nbetr. Groß-Berlin\nDe"| __truncated__ "Drucksache Nr. 4\
    nDeutscher Bundestag\nI. Wahlperiode\n1949\nAntrag\nder Fraktion der SPD\
    nbetr. Vorläufigen Si"| __truncated__ ...
##  $ jahr           : int  1949 1949 1949 1949 1949 1949 1949 1949 1949 1949 ...
##  $ nummer_zusatz  : chr  "NA" "NA" "NA" "NA" ...
##  $ nummer_dok     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ doi_concept    : chr  "10.5281/zenodo.4643065" "10.5281/zenodo.4643065"
    "10.5281/zenodo.4643065" "10.5281/zenodo.4643065" ...
##  $ doi_version    : chr  "10.5281/zenodo.4643066" "10.5281/zenodo.4643066"
    "10.5281/zenodo.4643066" "10.5281/zenodo.4643066" ...
##  $ version        : IDate, format: "2021-04-02" "2021-04-02" ...
##  $ zeichen        : int  40684 1830 955 350 982 1003 629 1066 1181 1836 ...
##  $ tokens         : int  8895 266 147 59 139 147 93 161 170 277 ...
##  $ typen          : int  1765 170 92 39 90 110 70 102 125 162 ...
##  $ saetze         : int  657 15 10 8 10 12 9 8 10 15 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

5.2 Hinweise

- Fehlende Werte sind immer mit »NA« codiert
- Strings können grundsätzlich alle in UTF-8 definierten Zeichen (insbesondere Buchstaben, Zahlen und Sonderzeichen) enthalten.

5.3 Erläuterungen zu den einzelnen Variablen

Variable	Typ	Erläuterung
doc_id	String	(Nur CSV-Datei) Der Name der extrahierten XML-Datei.
text	String	(Nur CSV-Datei) Der vollständige Text der Drucksache. Aus der XML-Datei extrahiert.
wahlperiode	Natürliche Zahl	Die Wahlperiode aus der die Drucksache stammt. Aus der XML-Datei extrahiert.
datum	Datum (ISO)	Das Datum der Drucksache im Format YYYY-MM-DD (Langform nach ISO-8601). Die Langform ist für Menschen einfacher lesbar und wird maschinell auch öfter automatisch als Datumsformat erkannt. Aus der XML-Datei extrahiert. In den XML-Rohdaten ist das Datum im Format DD.MM.YYYY dokumentiert und wurde vom Autor des Datensatzes in das ISO-Format transformiert.
jahr	Natürliche Zahl	(Nur CSV-Datei) Das Jahr der Drucksache im Format YYYY (Langform nach ISO-8601). Vom Autor des Datensatzes aus der Variable »datum« berechnet.
dokumentart	Alphabetisch	Es ist nur der Wert »DRUCKSACHE« vergeben. Wird vor allem dann relevant, wenn dieser Korpus mit einem Korpus der Plenarprotokolle verbunden wird. Aus der XML-Datei extrahiert.
drs_typ	String	(Nur CSV-Datei) Der Typ der Drucksache, beispielsweise »Gesetzentwurf« oder »Große Anfrage«. Alle möglichen Typen finden Sie unter Punkt 7.4. Aus der XML-Datei extrahiert.
p_urheber	String	(Nur CSV-Datei) Der oder die persönliche Urheber:in der Drucksache. Viele Drucksachen haben mehr als eine/n persönliche/n Urheber:in, diese sind dann in einem einzelnen String dokumentiert, jeweils getrennt durch einen vertikalen Strich (»|«). Bei schriftlichen Fragen und Fragen für die Fragestunde nicht erfasst. Aus der XML-Datei extrahiert.
k_urheber	String	(Nur CSV-Datei) Der körperschaftliche Urheber der Drucksache. Mehrere körperschaftliche Urheber sind in einem einzelnen String dokumentiert, jeweils getrennt durch einen vertikalen Strich (»|«). Aus der XML-Datei extrahiert.
titel	String	(Nur CSV-Datei) Der Titel der Drucksache. Aus der XML-Datei extrahiert.

Variable	Typ	Erläuterung
nummer_original	String	(Nur CSV-Datei) Eine Kombination von Wahlperiode und laufender Nummer der Drucksache. Beispielsweise steht »16/121« für die 121. Drucksache der 16. Wahlperiode. Manche Drucksachen sind Zusätze zu anderen Drucksachen und in diesem Fall mit einem »zu« am Ende der Nummer gekennzeichnet. Aus der XML-Datei extrahiert.
nummer_dok	Natürliche Zahl	Die laufende Nummer der Drucksache. Die Nummerierung beginnt in jeder Wahlperiode bei 1 und steigt bis zur maximalen Anzahl der Drucksachen einer Wahlperiode an. Durch den Autor des Datensatzes aus der Variable »nummer_original« berechnet.
nummer_zusatz	Alphabetisch	(Nur CSV-Datei) Ob es sich um einen Zusatz zu einem Dokument handelt. Mögliche Werte sind »ZUSATZ« oder »NA«. Zusatz-Dokumente haben den gleichen Wert der Variable »nummer_dok« wie das Dokument, auf das sie sich beziehen. Durch den Autor des Datensatzes aus der Variable »nummer_original« berechnet.
zeichen	Natürliche Zahl	(Nur CSV-Datei) Die Anzahl Zeichen eines Dokumentes. Berechnung durch den Autor des Datensatzes.
tokens	Natürliche Zahl	(Nur CSV-Datei) Die Anzahl Tokens (beliebige Zeichenfolge getrennt durch whitespace) eines Dokumentes. Diese Zahl kann je nach Tokenizer und verwendeten Einstellungen erheblich schwanken. Für diese Berechnung wurde eine reine Tokenisierung ohne Entfernung von Inhalten durchgeführt. Benutzen Sie diesen Wert eher als Anhaltspunkt für die Größenordnung denn als exakte Aussage und führen sie ggf. mit ihrer eigenen Software eine Kontroll-Rechnung durch. Berechnung durch den Autor des Datensatzes.
typen	Natürliche Zahl	(Nur CSV-Datei) Die Anzahl einzigartiger Tokens (beliebige Zeichenfolge getrennt durch whitespace) eines Dokumentes. Diese Zahl kann je nach Tokenizer und verwendeten Einstellungen erheblich schwanken. Für diese Berechnung wurde eine reine Tokenisierung und Typenzählung ohne Entfernung von Inhalten durchgeführt. Benutzen Sie diesen Wert eher als Anhaltspunkt für die Größenordnung denn als exakte Aussage und führen sie ggf. mit ihrer eigenen Software eine Kontroll-Rechnung durch. Berechnung durch den Autor des Datensatzes.

Variable	Typ	Erläuterung
saetze	Natürliche Zahl	(Nur CSV-Datei) Die Anzahl Sätze. Entsprechen in etwa dem üblichen Verständnis eines Satzes. Die Regeln für die Bestimmung von Satzanfang und Satzende sind im Detail sehr komplex und in »Unicode Standard: Annex No 29« beschrieben. Diese Zahl kann je nach Software und verwendeten Einstellungen erheblich schwanken. Für diese Berechnung wurde eine reine Zählung ohne Entfernung von Inhalten durchgeführt. Benutzen Sie diesen Wert eher als Anhaltspunkt für die Größenordnung denn als exakte Aussage und führen sie ggf. mit ihrer eigenen Software eine Kontroll-Rechnung durch. Berechnung durch den Autor des Datensatzes.
version	Datum (ISO)	(Nur CSV-Datei) Die Versionsnummer des Datensatzes im Format YYYY-MM-DD (Langform nach ISO-8601). Die Versionsnummer entspricht immer dem Datum an dem der Datensatz erstellt und die Daten von der Webseite des Bundestages abgerufen wurden. Eingefügt durch den Autor des Datensatzes.
doi_concept	String	(Nur CSV-Datei) Der Digital Object Identifier (DOI) des Gesamtkonzeptes des Datensatzes. Dieser ist langzeit-stabil (persistent). Über diese DOI kann via www.doi.org immer die **aktuellste Version** des Datensatzes abgerufen werden. Prinzip F1 der FAIR-Data Prinzipien (»data are assigned globally unique and persistent identifiers«) empfiehlt die Dokumentation jeder Messung mit einem persistenten Identifikator. Selbst wenn die CSV-Dateien ohne Kontext weitergegeben werden kann ihre Herkunft so immer zweifelsfrei und maschinenlesbar bestimmt werden. Eingefügt durch den Autor des Datensatzes.

Variable	Typ	Erläuterung
doi_version	String	(Nur CSV-Datei) Der Digital Object Identifier (DOI) der **konkreten Version** des Datensatzes. Dieser ist langzeit-stabil (persistent). Über diese DOI kann via www.doi.org immer diese konkrete Version des Datensatzes abgerufen werden. Prinzip F1 der FAIR-Data Prinzipien (»data are assigned globally unique and persistent identifiers«) empfiehlt die Dokumentation jeder Messung mit einem persistenten Identifikator. Selbst wenn die CSV-Dateien ohne Kontext weitergegeben werden kann ihre Herkunft so immer zweifelsfrei und maschinenlesbar bestimmt werden. Eingefügt durch den Autor des Datensatzes.

5.4 Konkordanztabelle: XML-Struktur und CSV-Variablen

CSV-Variable	XPath	Attribut
titel	/TITEL	-
wahlperiode	/WAHLPERIODE	-
datum	/DATUM	-
dokumentart	/DOKUMENTART	-
drs_typ	/DRS_TYP	-
k_urheber	/K_URHEBER	-
p_urheber	/P_URHEBER	-
nummer_original	/NR	-
text	/TEXT	-

Diese Konkordanztabelle bezieht sich auf die von der 1. bis zur 18. Wahlperiode gültige Document Type Definition (DTD) des Bundestages. Die DTD ist im Datensatz als separate Datei dokumentiert.

6 Computerlinguistische Kennzahlen

Zur besseren Einschätzung des inhaltlichen Umfangs des Korpus dokumentiere ich an dieser Stelle die Verteilung der Werte für verschiedene klassische computerlinguistische Kennzahlen:

Kennzahl	Definition
Zeichen	Zeichen entsprechen grob den *Graphemen*, den kleinsten funktionalen Einheiten in einem Schriftsystem. Beispiel: das Wort »Richterin« besteht aus 9 Zeichen.
Tokens	Eine beliebige Zeichenfolge, getrennt durch whitespace-Zeichen, d.h. ein Token entspricht in der Regel einem »Wort«, kann aber auch Zahlen, Sonderzeichen oder sinnlose Zeichenfolgen enthalten, weil es rein syntaktisch berechnet wird.
Typen	Einzigartige Tokens. Beispiel: wenn das Token »Verfassungsrecht« zehnmal in einem Dokument vorhanden ist, wird es als ein Typ gezählt.
Sätze	Entsprechen in etwa dem üblichen Verständnis eines Satzes. Die Regeln für die Bestimmung von Satzanfang und Satzende sind im Detail aber sehr komplex und in »Unicode Standard: Annex No 29« beschrieben.

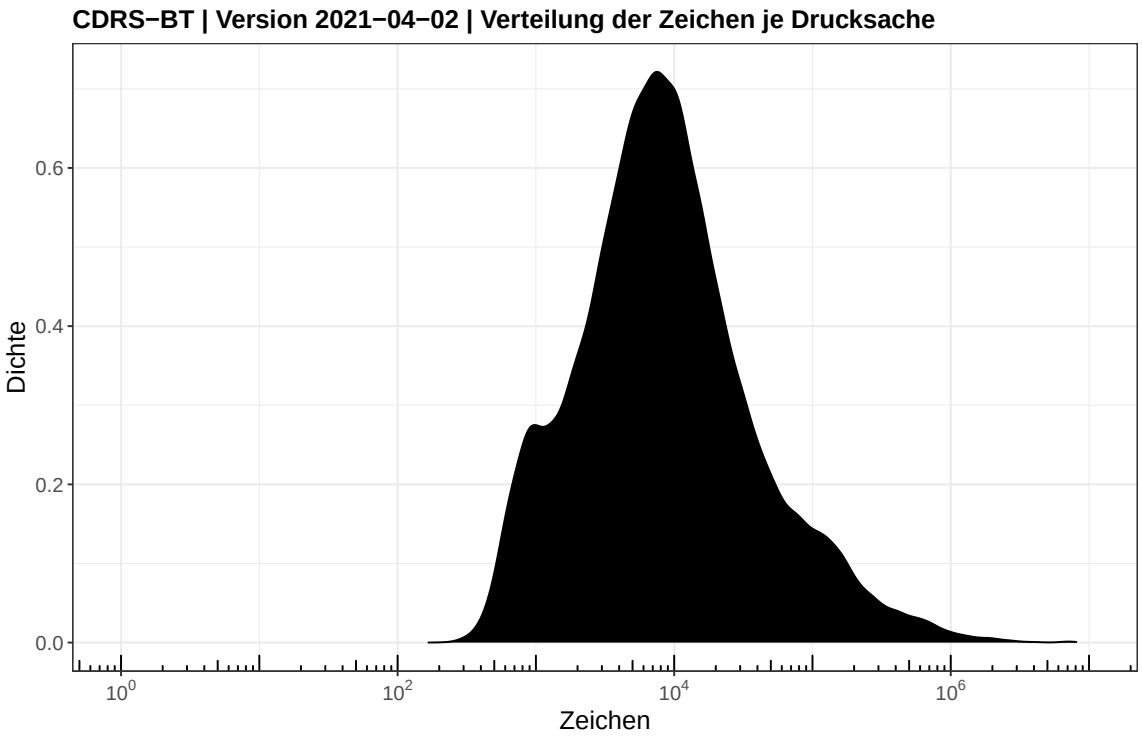
Es handelt sich bei den Diagrammen jeweils um »Density Charts«, die sich besonders dafür eignen die Schwerpunkte von Variablen mit stark schwankenden numerischen Werten zu visualisieren. Die Interpretation ist denkbar einfach: je höher die Kurve, desto dichter sind in diesem Bereich die Werte der Variable. Der Wert der y-Achse kann außer Acht gelassen werden, wichtig sind nur die relativen Flächenverhältnisse und die x-Achse.

Vorsicht bei der Interpretation: Die x-Achse it logarithmisch skaliert, d.h. in 10er-Potenzen und damit nicht-linear. Die kleinen Achsen-Markierungen zwischen den Schritten der Exponenten sind eine visuelle Hilfestellung um diese nicht-Linearität zu verstehen.

6.1 Werte der Kennzahlen

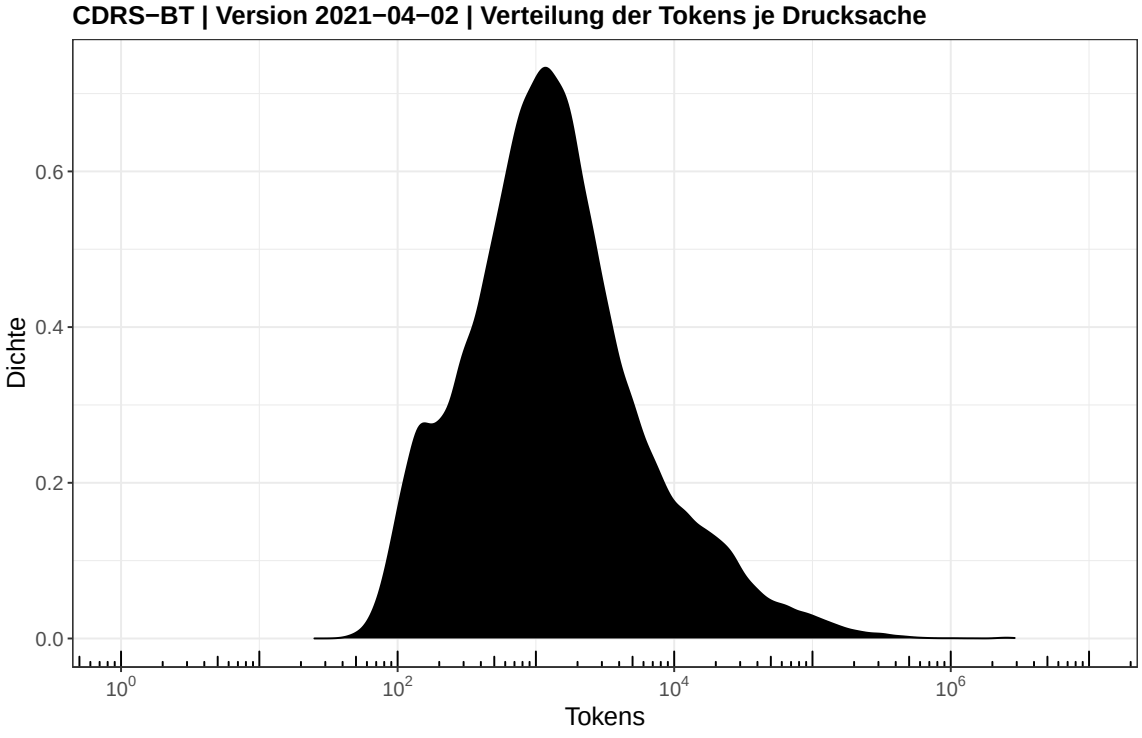
Variable	Summe	Min	Quart1	Median	Mittel	Quart3	Max
zeichen	4,726,077,216	0	3,241	7,815	35,848.43	19,688	8,088,489
tokens	805,465,703	0	510	1,214	6,109.65	3,041	2,899,377
saetze	33,381,383	0	29	63	253.21	146	134,121
typen	NA	0	238	478	1,067.37	970	79,666

6.2 Verteilung Zeichen



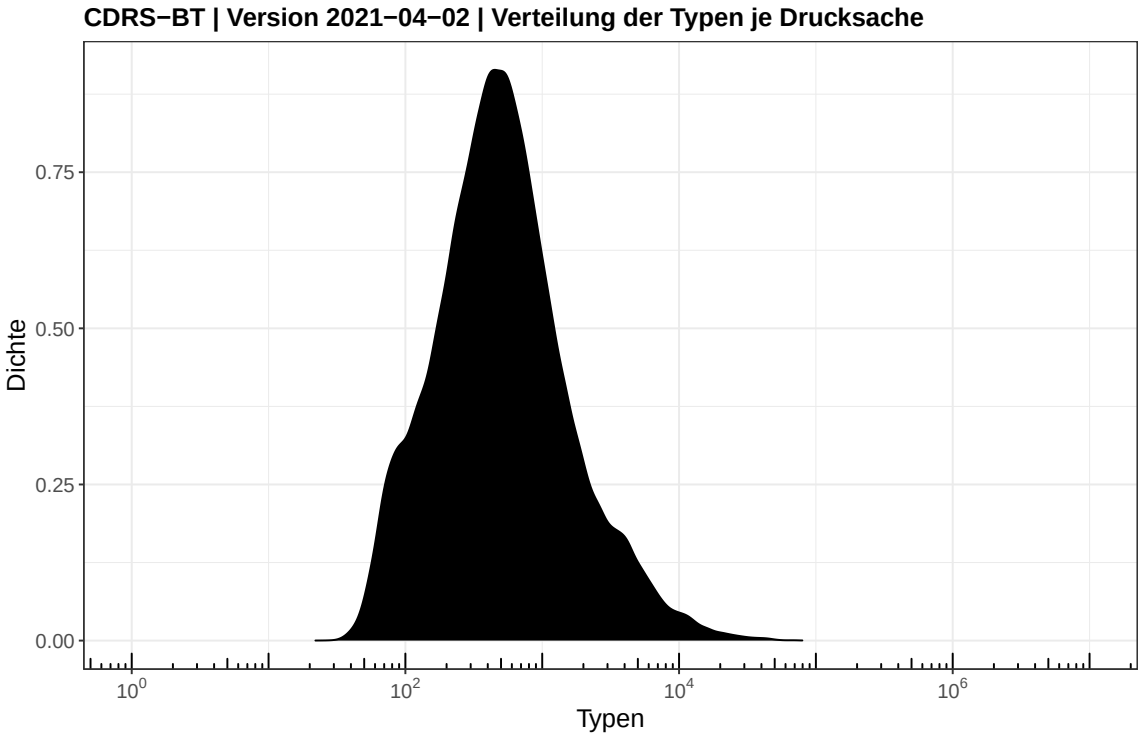
DOI: 10.5281/zenodo.4643066

6.3 Verteilung Tokens



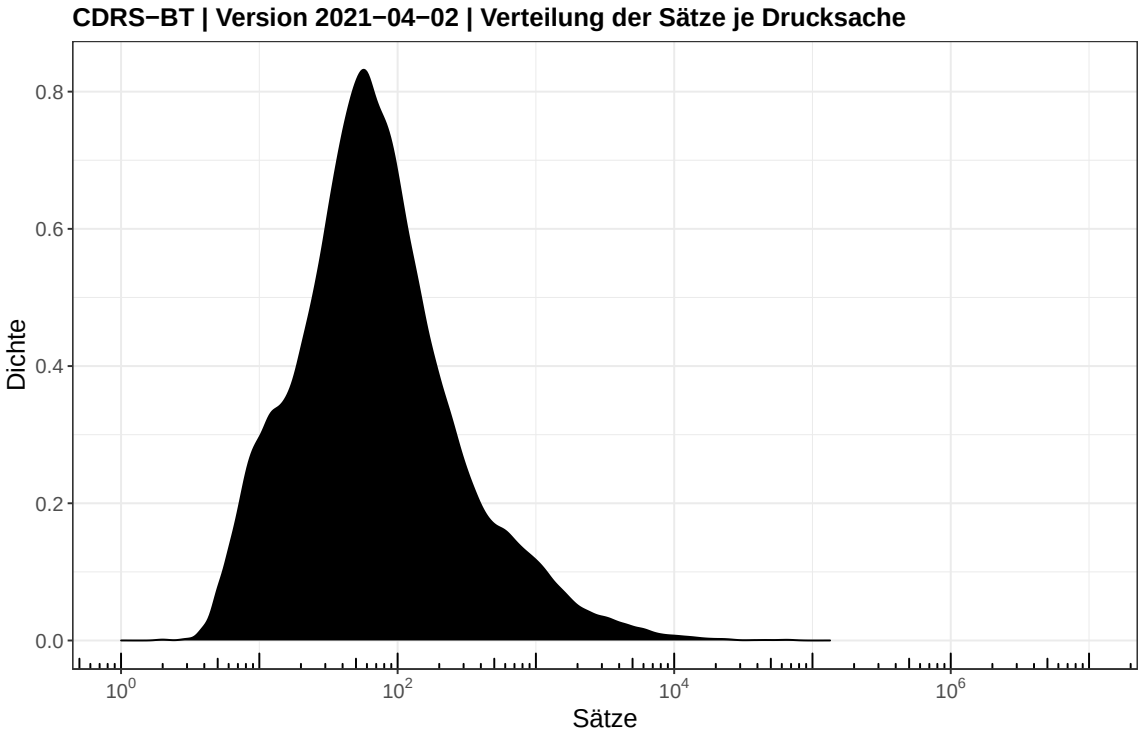
DOI: 10.5281/zenodo.4643066

6.4 Verteilung Typen



DOI: 10.5281/zenodo.4643066

6.5 Verteilung Sätze



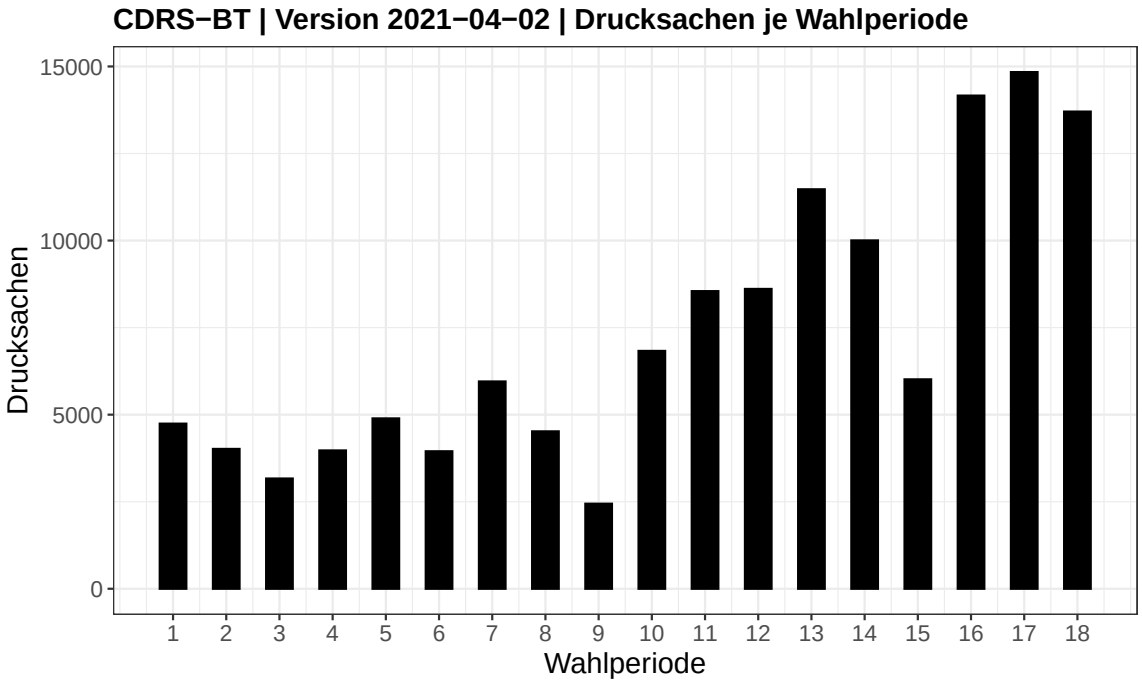
DOI: 10.5281/zenodo.4643066

7 Inhalt

7.1 Zusammenfassung

Variable	Anzahl	Min	Quart1	Median	Mittel	Quart3	Max
wahlperiode	18	1	8	13	11.81	16	18
jahr	69	1949	1978	1996	1992.02	2008	2017
nummer_dok	14838	1	1811	3726	4675.87	6912	14838

7.2 Nach Wahlperiode



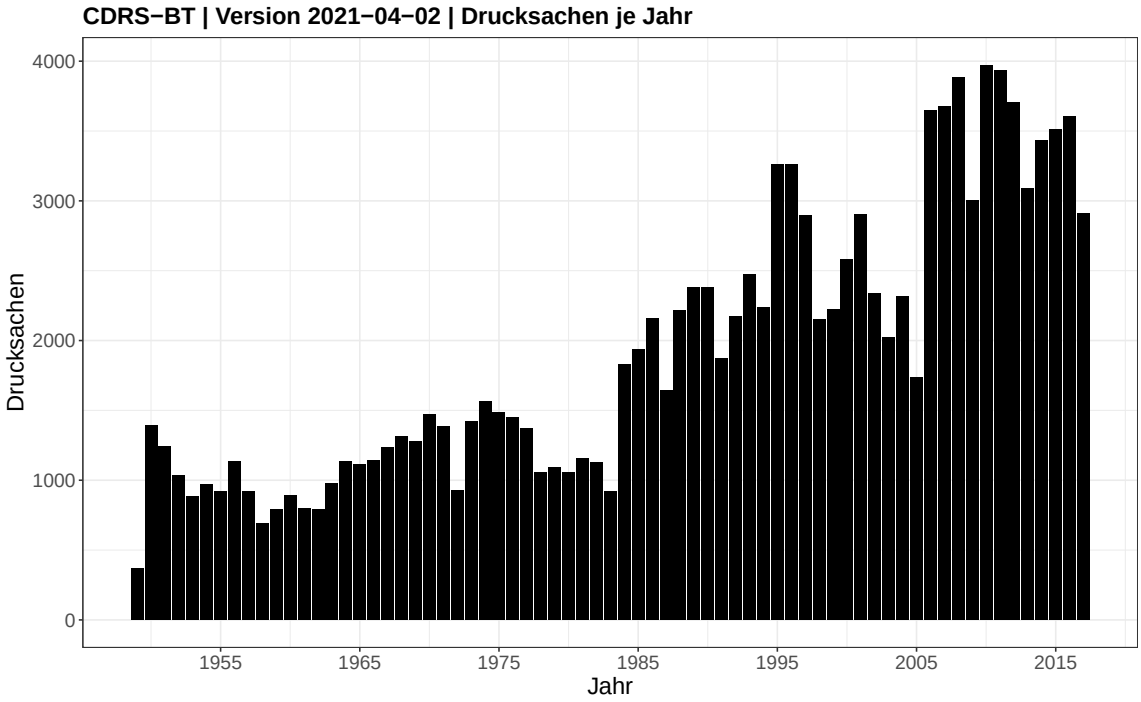
DOI: 10.5281/zenodo.4643066

Wahlperiode	Drucksachen	% Gesamt	% Kumulativ
1	4742	3.60	3.60
2	4016	3.05	6.64
3	3166	2.40	9.04
4	3972	3.01	12.06
5	4892	3.71	15.77
6	3948	2.99	18.76

(continued)

Wahlperiode	Drucksachen	% Gesamt	% Kumulativ
7	5953	4.52	23.28
8	4520	3.43	26.71
9	2443	1.85	28.56
10	6830	5.18	33.74
11	8547	6.48	40.22
12	8611	6.53	46.76
13	11472	8.70	55.46
14	10004	7.59	63.05
15	6014	4.56	67.61
16	14163	10.74	78.35
17	14838	11.25	89.61
18	13704	10.39	100.00
Total	131835	100.00	100.00

7.3 Nach Jahr



DOI: 10.5281/zenodo.4643066

Jahr	Drucksachen	% Gesamt	% Kumulativ
1949	367	0.28	0.28
1950	1391	1.06	1.33
1951	1239	0.94	2.27
1952	1036	0.79	3.06
1953	882	0.67	3.73
1954	972	0.74	4.47
1955	921	0.70	5.16
1956	1137	0.86	6.03
1957	921	0.70	6.73
1958	691	0.52	7.25
1959	791	0.60	7.85
1960	889	0.67	8.52
1961	795	0.60	9.13
1962	790	0.60	9.73
1963	978	0.74	10.47

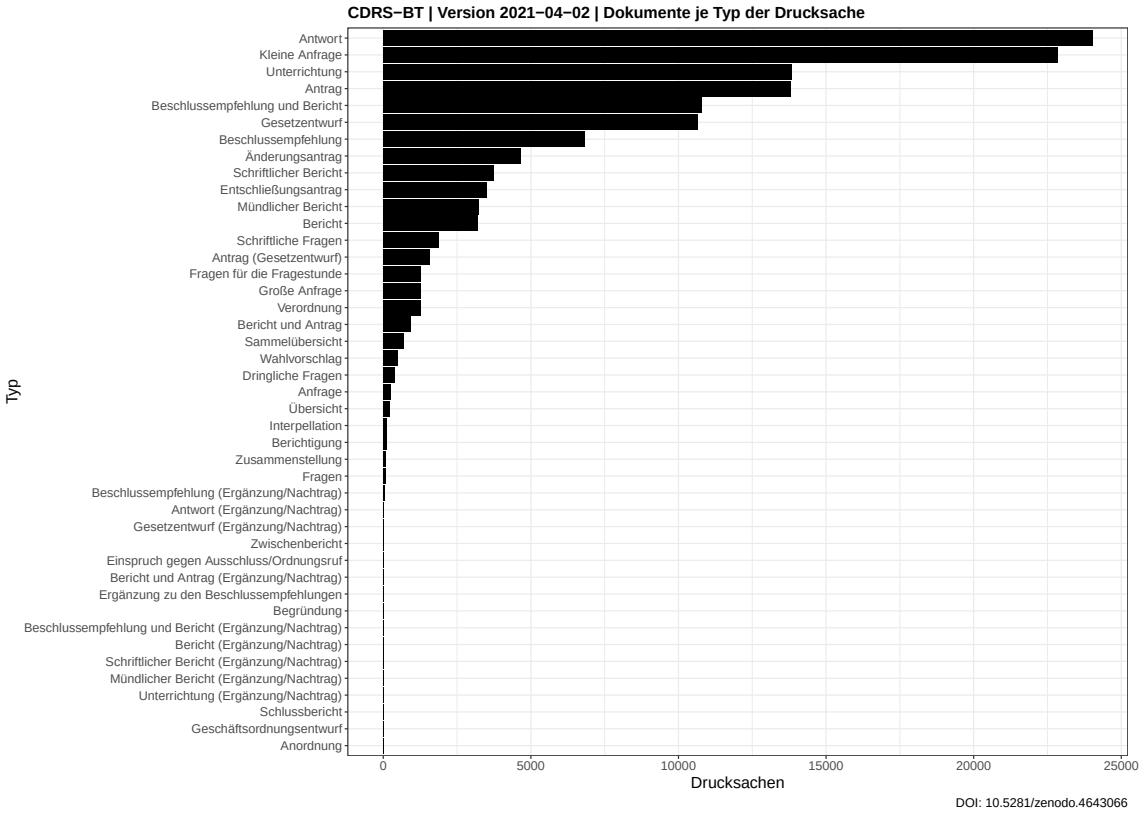
(continued)

Jahr	Drucksachen	% Gesamt	% Kumulativ
1964	1133	0.86	11.33
1965	1116	0.85	12.17
1966	1145	0.87	13.04
1967	1234	0.94	13.98
1968	1313	1.00	14.97
1969	1277	0.97	15.94
1970	1472	1.12	17.06
1971	1383	1.05	18.11
1972	924	0.70	18.81
1973	1418	1.08	19.88
1974	1567	1.19	21.07
1975	1485	1.13	22.20
1976	1452	1.10	23.30
1977	1370	1.04	24.34
1978	1056	0.80	25.14
1979	1095	0.83	25.97
1980	1055	0.80	26.77
1981	1159	0.88	27.65
1982	1128	0.86	28.51
1983	922	0.70	29.21
1984	1827	1.39	30.59
1985	1936	1.47	32.06
1986	2157	1.64	33.70
1987	1641	1.24	34.94
1988	2219	1.68	36.62
1989	2379	1.80	38.43
1990	2382	1.81	40.24
1991	1873	1.42	41.66
1992	2170	1.65	43.30
1993	2472	1.88	45.18
1994	2240	1.70	46.88

(continued)

Jahr	Drucksachen	% Gesamt	% Kumulativ
1995	3258	2.47	49.35
1996	3259	2.47	51.82
1997	2893	2.19	54.01
1998	2150	1.63	55.65
1999	2220	1.68	57.33
2000	2578	1.96	59.28
2001	2903	2.20	61.49
2002	2335	1.77	63.26
2003	2022	1.53	64.79
2004	2316	1.76	66.55
2005	1735	1.32	67.86
2006	3645	2.76	70.63
2007	3678	2.79	73.42
2008	3882	2.94	76.36
2009	3004	2.28	78.64
2010	3972	3.01	81.66
2011	3937	2.99	84.64
2012	3702	2.81	87.45
2013	3091	2.34	89.79
2014	3432	2.60	92.40
2015	3510	2.66	95.06
2016	3602	2.73	97.79
2017	2911	2.21	100.00
Total	131835	100.00	100.00

7.4 Nach Typ der Drucksache



Typ	Drucksachen	% Gesamt	% Kumulativ
Anfrage	241	0.18	0.18
Anordnung	1	0.00	0.18
Antrag	13809	10.47	10.66
Antrag (Gesetzentwurf)	1559	1.18	11.84
Antwort	24019	18.22	30.06
Antwort (Ergänzung/Nachtrag)	24	0.02	30.08
Begründung	4	0.00	30.08
Bericht	3192	2.42	32.50
Bericht (Ergänzung/Nachtrag)	3	0.00	32.50
Bericht und Antrag	929	0.70	33.21
Bericht und Antrag (Ergänzung/Nachtrag)	5	0.00	33.21
Berichtigung	108	0.08	33.29
Beschlussempfehlung	6836	5.19	38.48

(continued)

Typ	Drucksachen	% Gesamt	% Kumulativ
Beschlussempfehlung (Ergänzung/Nachtrag)	34	0.03	38.51
Beschlussempfehlung und Bericht	10802	8.19	46.70
Beschlussempfehlung und Bericht (Ergänzung/Nachtrag)	3	0.00	46.70
Dringliche Fragen	404	0.31	47.01
Einspruch gegen Ausschluss/Ordnungsruf	11	0.01	47.02
Entschließungsantrag	3506	2.66	49.68
Ergänzung zu den Beschlussempfehlungen	4	0.00	49.68
Fragen	68	0.05	49.73
Fragen für die Fragestunde	1284	0.97	50.70
Geschäftsordnungsentwurf	1	0.00	50.71
Gesetzentwurf	10653	8.08	58.79
Gesetzentwurf (Ergänzung/Nachtrag)	16	0.01	58.80
Große Anfrage	1273	0.97	59.76
Interpellation	120	0.09	59.85
Kleine Anfrage	22847	17.33	77.18
Mündlicher Bericht	3227	2.45	79.63
Mündlicher Bericht (Ergänzung/Nachtrag)	2	0.00	79.63
Sammelübersicht	695	0.53	80.16
Schlussbericht	1	0.00	80.16
Schriftliche Fragen	1873	1.42	81.58
Schriftlicher Bericht	3757	2.85	84.43
Schriftlicher Bericht (Ergänzung/Nachtrag)	2	0.00	84.43
Unterrichtung	13829	10.49	94.92
Unterrichtung (Ergänzung/Nachtrag)	1	0.00	94.92
Verordnung	1256	0.95	95.88
Wahlvorschlag	476	0.36	96.24
Zusammenstellung	83	0.06	96.30
Zwischenbericht	12	0.01	96.31
Änderungsantrag	4648	3.53	99.84

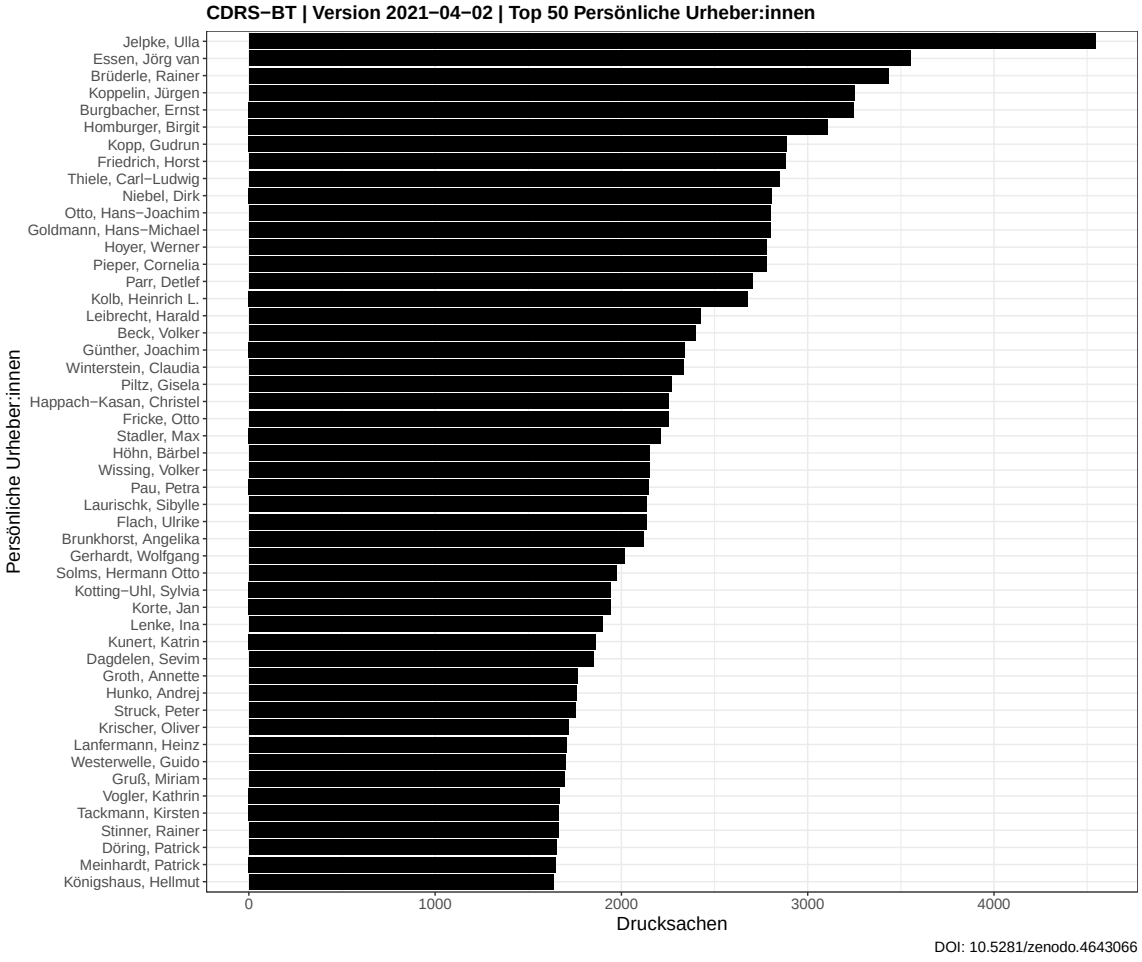
(continued)

Typ	Drucksachen	% Gesamt	% Kumulativ
Übersicht	217	0.16	100.00
Total	131835	100.00	100.00

7.5 Top 50 Persönliche Urheber:innen

Hinweis: Aus Gründen der Lesbarkeit werden an dieser Stelle nur die 50 meistgenannten persönlichen Urheber:innen dargestellt. Die vollständige Frequenztabelle finden Sie als Lesefassung im Compilation Report abgedruckt und in einer maschinenlesbaren Fassung als CSV-Datei im ZIP-Archiv »ANALYSE« dokumentiert.

Berücksichtigt sind alle Drucksachen in denen die Person mindestens einmal als persönliche/r (Mit-)Urheber:in genannt ist.



Persönliche Urheber:in	Drucksachen	% Gesamt
Jelpke, Ulla	4545	0.67
Essen, Jörg van	3552	0.52
Brüderle, Rainer	3434	0.51
Koppelin, Jürgen	3250	0.48
Burgbacher, Ernst	3249	0.48
Homburger, Birgit	3109	0.46

(continued)

Persönliche Urheber:in	Drucksachen	% Gesamt
Kopp, Gudrun	2888	0.43
Friedrich, Horst	2882	0.42
Thiele, Carl-Ludwig	2850	0.42
Niebel, Dirk	2809	0.41
Goldmann, Hans-Michael	2801	0.41
Otto, Hans-Joachim	2801	0.41
Hoyer, Werner	2778	0.41
Pieper, Cornelia	2777	0.41
Parr, Detlef	2704	0.40
Kolb, Heinrich L.	2679	0.40
Leibrecht, Harald	2426	0.36
Beck, Volker	2397	0.35
Günther, Joachim	2342	0.35
Winterstein, Claudia	2334	0.34
Piltz, Gisela	2269	0.33
Happach-Kasan, Christel	2254	0.33
Fricke, Otto	2252	0.33
Stadler, Max	2213	0.33
Höhn, Bärbel	2150	0.32
Wissing, Volker	2149	0.32
Pau, Petra	2148	0.32
Laurischk, Sibylle	2135	0.31
Flach, Ulrike	2134	0.31
Brunkhorst, Angelika	2118	0.31
Gerhardt, Wolfgang	2017	0.30
Solms, Hermann Otto	1975	0.29
Korte, Jan	1944	0.29
Kotting-Uhl, Sylvia	1944	0.29
Lenke, Ina	1897	0.28
Kunert, Katrin	1863	0.27
Dagdelen, Sevim	1850	0.27

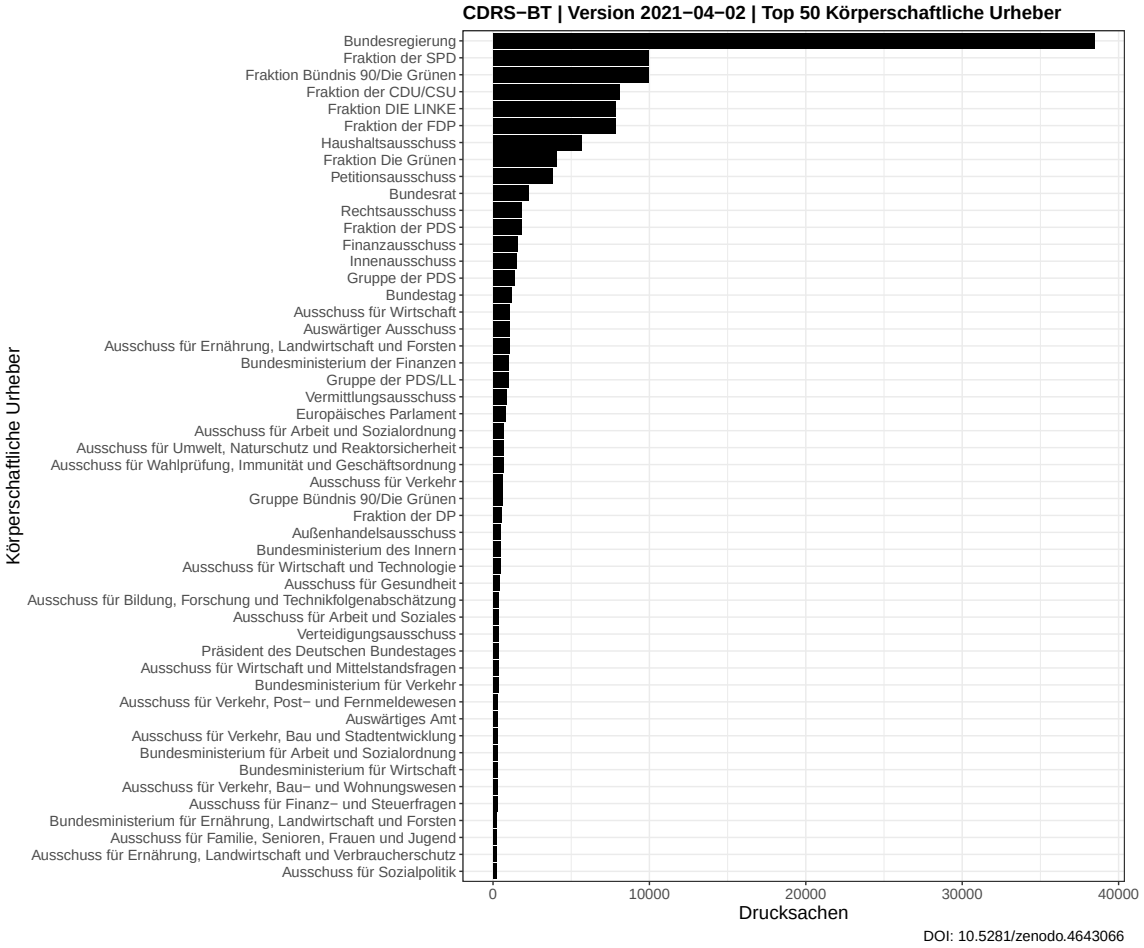
(continued)

Persönliche Urheber:in	Drucksachen	% Gesamt
Groth, Annette	1765	0.26
Hunko, Andrej	1758	0.26
Struck, Peter	1754	0.26
Krischer, Oliver	1715	0.25
Lanfermann, Heinz	1704	0.25
Westerwelle, Guido	1701	0.25
Gruß, Miriam	1695	0.25
Vogler, Kathrin	1670	0.25
Tackmann, Kirsten	1665	0.25
Stinner, Rainer	1661	0.24
Döring, Patrick	1653	0.24
Meinhardt, Patrick	1648	0.24
Königshaus, Hellmut	1635	0.24

7.6 Top 50 Körperschaftliche Urheber

Hinweis: Aus Gründen der Lesbarkeit werden an dieser Stelle nur die 50 meistgenannten körperschaftlichen Urheber dargestellt. Die vollständige Frequenztabelle finden Sie als Lesefassung im Compilation Report abgedruckt und in einer maschinenlesbaren Fassung als CSV-Datei im ZIP-Archiv »ANALYSE« dokumentiert.

Berücksichtigt sind alle Drucksachen in denen die Entität mindestens einmal als körperschaftlicher (Mit-)Urheber genannt ist.



Körperschaftlicher Urheber	Drucksachen	% Gesamt
Bundesregierung	38454	29.02
Fraktion der SPD	9967	7.52
Fraktion Bündnis 90/Die Grünen	9918	7.49
Fraktion der CDU/CSU	8091	6.11
Fraktion DIE LINKE	7836	5.91
Fraktion der FDP	7836	5.91
Haushaltsausschuss	5687	4.29

(continued)

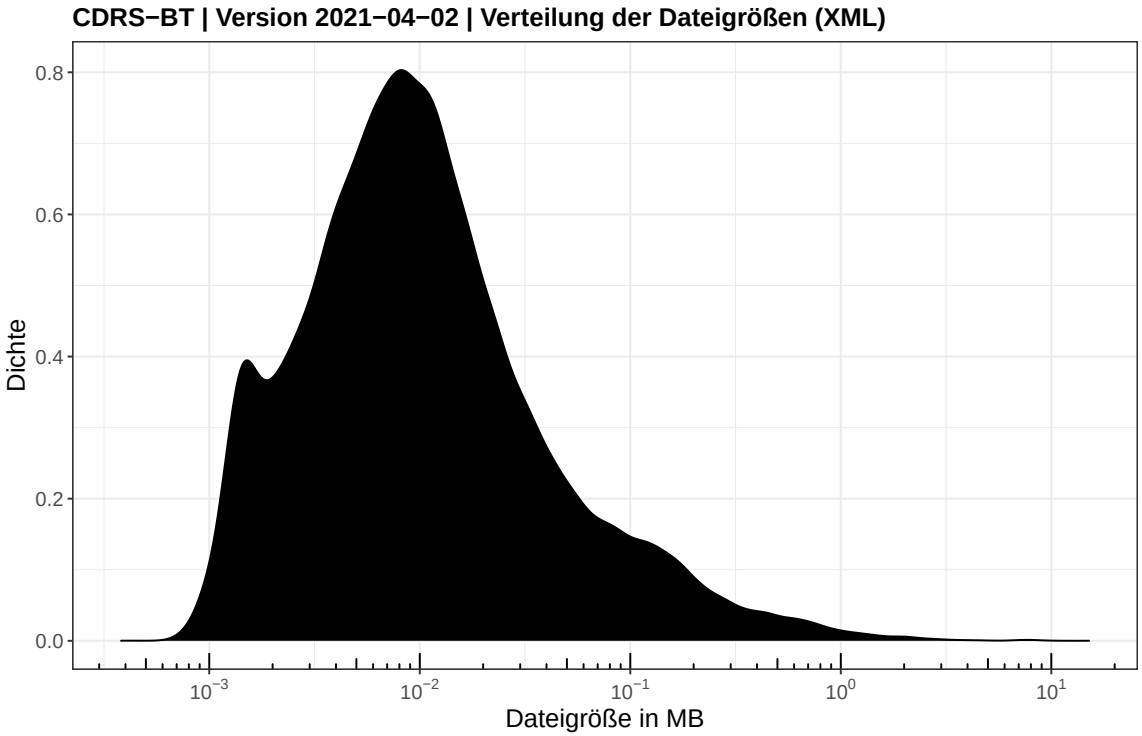
Körperschaftlicher Urheber	Drucksachen	% Gesamt
Fraktion Die Grünen	4040	3.05
Petitionsausschuss	3796	2.87
Bundesrat	2260	1.71
Rechtsausschuss	1853	1.40
Fraktion der PDS	1805	1.36
Finanzausschuss	1542	1.16
Innenausschuss	1521	1.15
Gruppe der PDS	1376	1.04
Bundestag	1157	0.87
Ausschuss für Wirtschaft	1068	0.81
Auswärtiger Ausschuss	1034	0.78
Ausschuss für Ernährung, Landwirtschaft und Forsten	1031	0.78
Bundesministerium der Finanzen	998	0.75
Gruppe der PDS/LL	994	0.75
Vermittlungsausschuss	843	0.64
Europäisches Parlament	807	0.61
Ausschuss für Arbeit und Sozialordnung	699	0.53
Ausschuss für Umwelt, Naturschutz und Reaktorsicherheit	698	0.53
Ausschuss für Wahlprüfung, Immunität und Geschäftsordnung	648	0.49
Ausschuss für Verkehr	603	0.46
Gruppe Bündnis 90/Die Grünen	581	0.44
Fraktion der DP	575	0.43
Außenhandelsausschuss	485	0.37
Bundesministerium des Innern	464	0.35
Ausschuss für Wirtschaft und Technologie	458	0.35
Ausschuss für Gesundheit	433	0.33
Ausschuss für Bildung, Forschung und Technikfolgenabschätzung	379	0.29
Ausschuss für Arbeit und Soziales	374	0.28

(continued)

Körperschaftlicher Urheber	Drucksachen	% Gesamt
Verteidigungsausschuss	361	0.27
Präsident des Deutschen Bundestages	352	0.27
Ausschuss für Wirtschaft und Mittelstandsfragen	349	0.26
Bundesministerium für Verkehr	340	0.26
Ausschuss für Verkehr, Post- und Fernmeldewesen	317	0.24
Auswärtiges Amt	296	0.22
Ausschuss für Verkehr, Bau und Stadtentwicklung	295	0.22
Bundesministerium für Arbeit und Sozialordnung	290	0.22
Bundesministerium für Wirtschaft	286	0.22
Ausschuss für Verkehr, Bau- und Wohnungswesen	281	0.21
Ausschuss für Finanz- und Steuerfragen	261	0.20
Bundesministerium für Ernährung, Landwirtschaft und Forsten	251	0.19
Ausschuss für Familie, Senioren, Frauen und Jugend	243	0.18
Ausschuss für Ernährung, Landwirtschaft und Verbraucherschutz	222	0.17
Ausschuss für Sozialpolitik	221	0.17

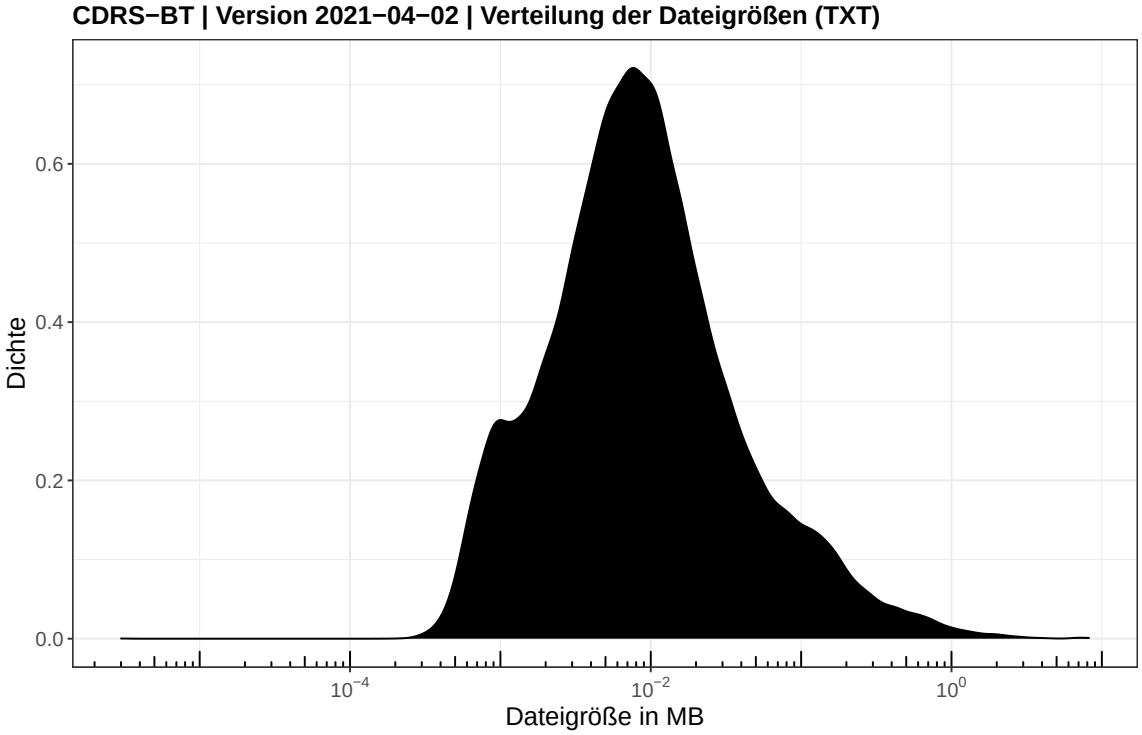
8 Dateigrößen

8.1 Verteilung XML-Dateigrößen



DOI: 10.5281/zenodo.4643066

8.2 Verteilung TXT-Dateigrößen



DOI: 10.5281/zenodo.4643066

8.3 Gesamtgröße je ZIP-Archiv

Datei	Größe in MB
CDRS-BT_2021-04-02_DE_ANALYSE.zip	1.17
CDRS-BT_2021-04-02_DE_CSV_Datensatz.zip	1,408.01
CDRS-BT_2021-04-02_DE_CSV_Metadaten.zip	10.92
CDRS-BT_2021-04-02_DE_TXT_Datensatz.zip	1,505.87
CDRS-BT_2021-04-02_DE_XML_Datensatz.zip	1,547.08
CDRS-BT_2021-04-02_Source_Files.zip	0.05

9 Prüfung kryptographischer Signaturen

9.1 Allgemeines

Die Integrität und Echtheit der einzelnen Archive des Datensatzes sind durch eine Zwei-Phasen-Signatur sichergestellt.

In **Phase I** werden während der Kompilierung für jedes ZIP-Archiv Hash-Werte in zwei verschiedenen Verfahren (SHA2-256 und SHA3-512) berechnet und in einer CSV-Datei dokumentiert.

In **Phase II** wird diese CSV-Datei mit meinem persönlichen geheimen GPG-Schlüssel signiert. Dieses Verfahren stellt sicher, dass die Kompilierung von jedermann durchgeführt werden kann, insbesondere im Rahmen von Replikationen, die persönliche Gewähr für Ergebnisse aber dennoch vorhanden bleibt.

Dieses Codebook ist vollautomatisch erstellt und prüft die kryptographisch sicheren SHA3-512 Signaturen (»hashes«) aller ZIP-Archive, sowie die GPG-Signatur der CSV-Datei, welche die SHA3-512 Signaturen enthält. SHA3-512 Signaturen werden durch einen system call zur OpenSSL library auf Linux-Systemen berechnet. Eine erfolgreiche Prüfung meldet »Verifiziert!«. Eine gescheiterte Prüfung meldet »FEHLER!«

9.2 Persönliche GPG-Signatur

Die während der Kompilierung des Datensatzes erstellte CSV-Datei mit den Hash-Prüfsummen ist mit meiner persönlichen GPG-Signatur versehen. Der mit dieser Version korrespondierende Public Key ist sowohl mit dem Datensatz als auch mit dem Source Code hinterlegt. Er hat folgende Kenndaten:

Name: Sean Fobbe (fobbe-data@posteo.de)

Fingerabdruck: FE6F B888 F0E5 656C 1D25 3B9A 50C4 1384 F44A 4E42

9.3 Import: Public Key

```
system2("gpg2", "--import GPG-Public-Key_Fobbe-Data.asc",
        stdout = TRUE,
        stderr = TRUE)
```

```
## [1] "gpg: key 50C41384F44A4E42: \"Sean Fobbe <fobbe-data@posteo.de>\" not
    changed"
## [2] "gpg: Total number processed: 1"
## [3] "gpg:              unchanged: 1"
```

9.4 Prüfung: GPG-Signatur der Hash-Datei

```
# CSV-Datei mit Hashes
print(hashfile)
```

```
## [1] "CDRS-BT_2021-04-02_KryptographischeHashes.csv"
```

```
# GPG-Signatur
print(signaturefile)
```

```
## [1] "CDRS-BT_2021-04-02_FobbeSignaturGPG_Hashes.gpg"
```

```
# GPG-Signatur prüfen
testresult <- system2("gpg2",
                      paste("--verify", signaturefile, hashfile),
                      stdout = TRUE,
                      stderr = TRUE)

# Anführungsstriche entfernen um Anzeigefehler zu vermeiden
testresult <- gsub('"', '', testresult)
```

```
kable(testresult, format = "latex", booktabs = TRUE,
      longtable = TRUE, col.names = c("Ergebnis"))
```

Ergebnis
gpg: Signature made Fri 02 Apr 2021 04:08:54 PM CEST
gpg: using RSA key FE6FB888F0E5656C1D253B9A50C41384F44A4E42
gpg: Good signature from Sean Fobbe <fobbe-data@posteo.de> [full]

9.5 Prüfung: SHA3-512 Hashes der ZIP-Archive

```
# Prüf-Funktion definieren
sha3test <- function(filename, sig){
    sig.new <- system2("openssl",
                       paste("sha3-512", filename),
                       stdout = TRUE)
    sig.new <- gsub("^.*\\= ", "", sig.new)
    if (sig == sig.new){
        return("Signatur verifiziert!")
    }else{
        return("FEHLER!")
    }
}



# Ursprüngliche Signaturen importieren
table.hashes <- fread(hashfile)
filename <- table.hashes$filename
sha3.512 <- table.hashes$sha3.512

# Signaturprüfung durchführen
sha3.512.result <- mcmapply(sha3test, filename, sha3.512, USE.NAMES = FALSE)

# Ergebnis anzeigen
testresult <- data.table(filename, sha3.512.result)
```

```
kable(testresult, format = "latex", booktabs = TRUE,
      longtable = TRUE, col.names = c("Datei", "Ergebnis"))
```

Datei	Ergebnis
CDRS-BT_2021-04-02_DE_ANALYSE.zip	Signatur verifiziert!
CDRS-BT_2021-04-02_DE_CSV_Datensatz.zip	Signatur verifiziert!
CDRS-BT_2021-04-02_DE_CSV_Metadaten.zip	Signatur verifiziert!
CDRS-BT_2021-04-02_DE_TXT_Datensatz.zip	Signatur verifiziert!
CDRS-BT_2021-04-02_DE_XML_Datensatz.zip	Signatur verifiziert!
CDRS-BT_2021-04-02_Source_Files.zip	Signatur verifiziert!

10 Changelog

Version	Details
2021-04-02	• Erstveröffentlichung

11 Parameter für strenge Replikationen

```
## [1] "OpenSSL 1.1.1i FIPS  8 Dec 2020"
```

```
## R version 4.0.4 (2021-02-15)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora 32 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libopenblas-r0.3.12.so
##
## locale:
##  [1] LC_CTYPE=en_US.utf8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.utf8        LC_COLLATE=en_US.utf8
##  [5] LC_MONETARY=en_US.utf8    LC_MESSAGES=en_US.utf8
##  [7] LC_PAPER=en_US.utf8       LC_NAME=C
##  [9] LC_ADDRESS=C              LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.utf8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] magick_2.7.1      quanteda_2.1.2    data.table_1.14.0 scales_1.1.1
##  [5] ggplot2_3.3.3     kableExtra_1.3.4  knitr_1.31        doParallel_1.0.16
##  [9] iterators_1.0.13  foreach_1.5.1     xml2_1.3.2        rvest_1.0.0
##
## loaded via a namespace (and not attached):
##  [1] tinytex_0.30       tidyselect_1.1.0   xfun_0.22          purrr_0.3.4
##  [5] lattice_0.20-41    colorspace_2.0-0   vctrs_0.3.6        generics_0.1.0
##  [9] htmltools_0.5.1.1  viridisLite_0.3.0  yaml_2.2.1         utf8_1.2.1
## [13] rlang_0.4.10       pillar_1.5.1       glue_1.4.2         withr_2.4.1
## [17] bit64_4.0.5        lifecycle_1.0.0    stringr_1.4.0      munsell_0.5.0
## [21] gtable_0.3.0       codetools_0.2-18   evaluate_0.14      labeling_0.4.2
## [25] fansi_0.4.2        highr_0.8          Rcpp_1.0.6         RcppParallel
    _5.0.3
## [29] webshot_0.5.2      bit_4.0.4          farver_2.1.0       systemfonts
    _1.0.1
## [33] fastmatch_1.1-0    stopwords_2.2      digest_0.6.27      stringi_1.5.3
## [37] dplyr_1.0.5        grid_4.0.4         tools_4.0.4        magrittr_2.0.1
## [41] tibble_3.1.0       crayon_1.4.1       pkgconfig_2.0.3    ellipsis_0.3.1
## [45] Matrix_1.3-2       rmarkdown_2.7      svglite_2.0.0      httr_1.4.2
## [49] rstudioapi_0.13    R6_2.5.0           compiler_4.0.4
```

Literaturverzeichnis

Analytics, Revolution, and Steve Weston. 2020. *Iterators: Provides Iterator Construct*. https://github.com/RevolutionAnalytics/iterators.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, Jiong Wei Lua, Jouni Kuha, and William Lowe. 2020. *Quanteda: Quantitative Analysis of Textual Data*. https://quanteda.io.

Corporation, Microsoft, and Steve Weston. 2020. *DoParallel: Foreach Parallel Adaptor for the Parallel Package*. https://CRAN.R-project.org/package=doParallel.

Dowle, Matt, and Arun Srinivasan. 2021. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Revolution Analytics, and Steve Weston. n.d. *Foreach: Provides Foreach Looping Construct*.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2021. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Jim Hester, and Jeroen Ooms. 2020. *Xml2: Parse Xml*. https://CRAN.R-project.org/package=xml2.

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with Kable and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.