Corpus der Entscheidungen des Bundesverfassungsgerichts

(CE-BVerfG)

Codebook

Version 2022-02-01

DOI: 10.5281/zenodo.5910152

Titel	Corpus der Entscheidungen des Bundesverfassungsgerichts
Abkürzung	CE-BVerfG
Autor	Seán Fobbe
Version	2022-02-01
Download	https://doi.org/10.5281/zenodo.5910152
Lizenz	CC0 1.0 Universal

Zitiervorschlag

Seán Fobbe (2022). Corpus der Entscheidungen des Bundesverfassungsgerichts (CE-BVerfG). Version 2022-02-01. Zenodo. DOI: 10.5281/zenodo.5910152.

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 2022-02-01. 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.3902658. Die »Concept DOI« verlinkt immer die aktuellste Version.

Urheberrecht

Disclaimer

Dieser Datensatz ist eine private wissenschaftliche Initiative und steht in keiner Verbindung zu Behörden, Gerichten oder anderen amtlichen Stellen der Bundesrepublik Deutschland.

Inhaltsverzeichnis

1 Einführung

Das **Bundesverfassungsgericht (BVerfG)** ist das höchste Gericht der Bundesrepublik Deutschland und ein Verfassungsorgan. Als »Hüter der Verfassung« ist es seit seiner Gründung im Jahr 1951 mit der Auslegung und Durchsetzung des Grundgesetzes betraut.

Seine Bedeutung im Verfassungsgefüge der Bundesrepublik Deutschland ist kaum zu überschätzen. So richtet es nicht nur über Streitigkeiten zwischen Verfassungsorganen und über Normenkontrollanträge, welche die Nichtigkeit von Gesetzen zur Folge haben können, sondern auch über »Verfassungsbeschwerden, die von jedermann mit der Behauptung erhoben werden können, durch die öffentliche Gewalt in einem seiner Grundrechte oder in einem seiner [grundrechtsgleichen Rechte] verletzt zu sein« (Art. 93 Abs. 4b GG). Das Instrument der Verfassungsbeschwerde ist in seiner Beliebtheit und Effektivität in der Geschichte Deutschlands beispiellos und von hoher wissenschaftlicher und praktischer Bedeutung. In nicht wenigen Verfahrensarten haben die Entscheidungen des BVerfG zudem Gesetzeskraft (§ 31 Abs. 2 BVerfGG).

Die quantitative Analyse von juristischen Texten, insbesondere denen des Bundesverfassungsgerichts, ist in den deutschen Rechtswissenschaften ein noch junges und kaum bearbeitetes Feld.[1] Zu einem nicht unerheblichen Teil liegt dies auch daran, dass die Anzahl an frei nutzbaren Datensätzen außerordentlich gering ist.

Die meisten hochwertigen Datensätze lagern (fast) unerreichbar in kommerziellen Datenbanken und sind wissenschaftlich gar nicht oder nur gegen Entgelt zu nutzen. Frei verfügbare Datenbanken wie *Opinio Iuris*[2] und *openJur*[3] verbieten ausdrücklich das maschinelle Auslesen der Rohdaten. Wissenschaftliche Initiativen wie der Juristische Referenzkorpus (JuReKo) sind nach jahrelanger Arbeit hinter verschlossenen Türen verschwunden.

In einem funktionierenden Rechtsstaat muss die Rechtsprechung öffentlich, transparent und nachvollziehbar sein. Im 21. Jahrhundert bedeutet dies auch, dass sie systematischer Überprüfung mittels 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.[4]

[1] Besonders positive Ausnahmen finden sich unter: https://www.quantitative-rechtswissenschaft.de/
[2] https://opinioiuris.de/
[3] https://openjur.de/
[4] 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, PDF) veröffentlicht. Sie lassen sich grundsätzlich mit allen modernen Programmiersprachen (z.B. Python oder R), 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[5] 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 Entscheidung. Die Variablen sind unter Punkt 5 genauer erläutert.

Hier 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. Ein Vorschlag:

```
library(data.table)
dt.bverfg <- 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. Ein Vorschlag:

```
library(readtext)
df.bverfg <- readtext("*.txt",
                      docvarsfrom = "filenames",
                      docvarnames = c("gericht",
                                      "datum",
                                      "spruchkoerper_typ",
                                      "spruchkoerper_az",
                                      "registerzeichen",
                                      "eingangsnummer",
                                      "eingangsjahr_az",
                                      "kollision",
                                      "name",
                                      "band",
                                      "seite"),
                      dvsep = "_",
                      encoding = "UTF-8")
```

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

3 Konstruktion

3.1 Beschreibung des Datensatzes

Dieser Datensatz ist eine digitale Zusammenstellung von möglichst allen Entscheidungen, die auf der amtlichen Internetpräsenz des Bundesverfassungsgerichts (BVerfG) veröffentlicht sind. Er enthält alle Entscheidungen, die auf der offiziellen Webseite des Bundesverfassungsgerichts am jeweiligen Stichtag veröffentlicht waren. Die Stichtage für jede Version entsprechen exakt der Versionsnummer.

Zusätzlich zu den einfach maschinenlesbaren Formaten (TXT und CSV) sind die PDF-Rohdaten enthalten, damit Analyst:innen gegebenenfalls ihre eigene Konvertierung vornehmen können. Die PDF-Rohdaten wurden inhaltlich nicht verändert und nur die Dateinamen angepasst um die Lesbarkeit für Mensch und Maschine zu verbessern.

3.2 Datenquellen

Datenquelle	Fundstelle
Primäre Datenquelle	https://www.bundesverfassungsgericht.de
Source Code	https://doi.org/10.5281/zenodo.5910155
Entscheidungsnamen	https://doi.org/10.5281/zenodo.5910155
BVerfGE-Fundstellen	https://doi.org/10.5281/zenodo.5910155
Personendaten	https://doi.org/10.5281/zenodo.4568682
Registerzeichen	https://doi.org/10.5281/zenodo.4569564

Die Personendaten stammen aus folgendem Datensatz: »Seán Fobbe and Tilko Swalve (2021). Presidents and Vice-Presidents of the Federal Courts of Germany (PVP-FCG). Version 2021-04-08. Zenodo. DOI: 10.5281/zenodo.4568682«.

Die Tabelle der Registerzeichen und der ihnen zugeordneten Verfahrensarten stammt aus dem folgenden Datensatz: “Seán Fobbe (2021). Aktenzeichen der Bundesrepublik Deutschland (AZ-BRD). Version 1.0.1. Zenodo. DOI: 10.5281/zenodo.4569564.”

3.3 Sammlung der Daten

Die Daten wurden unter Beachtung des Robot Exclusion Standard (RES) gesammelt. Der Abruf geschieht ausschließlich über TLS-verschlüsselte Verbindungen. Die Entscheidungen sind laut dem Gericht anonymisiert, aber ungekürzt.

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 unter dieser Addresse hinterlegt: https://doi.org/10.5281/zenodo.5910155

Mit jeder Kompilierung des vollständigen Datensatzes wird auch ein umfangreicher **Compilation Report** in einem attraktiv designten PDF-Format erstellt (ähnlich diesem Codebook). 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 des Erstellungs-Prozesses interessieren, lesen Sie diesen bitte zuerst.

3.5 Grenzen des Datensatzes

Nutzer sollten folgende wichtige Grenzen beachten:

1. Der Datensatz enthält nur das, was das Gericht auch tatsächlich veröffentlicht, nämlich begründete Entscheidungen (*publication bias*). Laut dem BVerfG sind dies »alle Senatsentscheidungen (nicht aber Nebenentscheidungen wie z.B. Streitwertbeschlüsse und Kostenfestsetzungen) und viele begründete Kammerbeschlüsse«.[6] Bei mehreren gleichgelagerten Sachverhalten wird nur das Pilotverfahren veröffentlicht.

2. Es kann aufgrund technischer Grenzen bzw. Fehler sein, dass manche — im Grunde verfügbare — Entscheidungen nicht oder nicht korrekt abgerufen werden (*automation bias*).

3. Es werden nur PDF- und HTML-Dateien abgerufen (*file type bias*). Manche Entscheidungen sind nur als HTML verfügbar. Die Metadaten der Entscheidungen ohne PDF-Datei werden explizit im Compilation Report dokumentiert.

4. Erst ab dem Jahr 1998 sind begründete Entscheidungen des BVerfG einigermaßen vollständig veröffentlicht, auch wenn frühere Entscheidungen vereinzelt auf der Webseite verfügbar sind (*temporal bias*). Die Frequenztabellen geben hierzu genauer Auskunft.

3.6 Urheberrechtsfreiheit von Rohdaten und Datensatz

An den Entscheidungstexten und amtlichen Leitsätzen besteht gem. § 5 Abs. 1 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«).

Der HTML-Quelltext wurde — wie in jeder HTML-Datei selbst dokumentiert ist — mit dem »Government Site Builder« der Bundesverwaltung erstellt, d.h. computergeneriert. Durch Maschinen generierte Texte sind keine »persönliche geistige Schöpfung« iSv § 2 Abs. 2 UrhG und daher urheberrechtlich nicht geschützt. Den verbleibenden Text-Bestandteilen (z.B. Buttons) fehlt es mindestens an der Schöpfungshöhe. Bilder oder andere Texte als Entscheidungstexte werden nicht abgerufen.

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.

[6] https://www.bundesverfassungsgericht.de/DE/Entscheidungen/Entscheidungen/Entscheidungen.html

3.7 Metadaten

3.7.1 Allgemein

Die Metadaten wurden weitgehend aus den Hyperlinks zur jeweiligen Datei und dem HTML-Quelltext extrahiert. Hinzugefügt wurden von mir eine Reihe weitere Variablen, sowie Unter- und Trennstriche um die Maschinenlesbarkeit zu erleichten. Der volle Satz an Metadaten ist nur in den CSV-Dateien enthalten. Alle hinzugefügten Metadaten sind zusammen mit dem Source Code vollständig maschinenlesbar dokumentiert und liegen entweder im CSV-Format vor oder sind direkt im Source Code enthalten.

Die Dateinamen der PDF- und TXT-Dateien enthalten Gerichtsname, Datum (Langform nach ISO-8601, d.h. YYYY-MM-DD), den Typ des Spruchkörpers, das offizielle Aktenzeichen, eine Kollisions-ID, den Namen der Entscheidung, sowie die BVerfGE-Fundstelle (Band und Seite).

3.7.2 Schema für die Dateinamen

```
[gericht]_[datum]_[spruchkoerper_typ]_[spruchkoerper_az]_[registerzeichen]_
[eingangsnummer]_[eingangsjahr_az]_[kollision]_[name]_[band]_[seite]
```

3.7.3 Beispiel eines Dateinamens

```
BVerfG_1997-07-08_S_1_BvR_1243_95_NA_Parteilehrer_96_152.txt
```

3.8 Qualitätsprüfung

Die Typen der Variablen wurden mit *regular expressions* strikt validiert. Die möglichen Werte der jeweiligen Variablen wurden zudem durch Frequenztabellen und Visualisierungen auf ihre Plausibilität geprüft. Insgesamt werden zusammen mit jeder Kompilierung Dutzende Tests zur Qualitätsprüfung durchgeführt. Alle Ergebnisse der Qualitätsprüfungen sind aggregiert im Compilation Report und einzeln im Archiv »ANALYSE« zusammen mit dem Datensatz veröffentlicht.

3.9 Grafische Darstellung

Die Robenfarbe der Bundesverfassungsrichter ist »scharlachrot«. Der Hex-Wert hierfür ist #ca2129. Das ist besonders bei der Erstellung thematisch passender Graphen hilfreich. Alle im Compilation Report und diesem Codebook präsentierten Graphen sind in diesem scharlachrot gehalten.

4 Varianten und Zielgruppen

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

Lehrende werden zudem von den vorbereiteten Tabellen und Diagrammen besonders profitieren, die bei der Erläuterung der Charakteristika der Daten hilfreich sein können und Zeit im universitären Alltag sparen. 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
PDF	**Traditionelle juristische Forschung.** Die PDF-Dokumente wie sie vom Bundesverfassungsgericht auf der amtlichen Webseite bereitgestellt werden, jedoch verbessert durch semantisch hochwertige Dateinamen, die der leichteren Auffindbarkeit von Entscheidungen dienen. Die Dateinamen sind so konzipiert, dass sie auch für die traditionelle qualitative juristische Arbeit einen erheblichen Mehrwert bieten. Im Vergleich zu den CSV-Dateien enthalten die Dateinamen nur einen reduzierten Umfang an Metadaten, um Kompatibilitätsprobleme zu vermeiden und die Lesbarkeit zu verbessern.
CSV_Datensatz	**Legal Tech/Quantitative Forschung.** Diese CSV-Datei ist die für statistische Analysen empfohlene Variante des Datensatzes. Sie enthält den Volltext aller Entscheidungen, sowie alle in diesem Codebook beschriebenen Metadaten. Jede Spalte entspricht einer Variable, jede Zeile einer Entscheidung.
CSV_Metadaten	**Legal Tech/Quantitative Forschung.** Wie die vorige CSV-Variante, nur ohne die Entscheidungstexte. Sinnvoll für Analyst:innen, die sich nur für die Metadaten interessieren und Speicherplatz sparen wollen. Jede Spalte entspricht einer Variable, jede Zeile einer Entscheidung.
CSV_Annotiert	**Legal Tech/Quantitative Forschung.** Alle Entscheidungen in tokenisierter Form mit linguistischen Annotationen. Beachten Sie bitte die besondere Variablen-Struktur unter Punkt 6. Jede Spalte entspricht einer Variable, jede Zeile einem Token.

Variante	Zielgruppe und Beschreibung
CSV_Segmentiert	**Legal Tech/Quantitative Forschung. Experimentell!** Alle Entscheidungen in segmentierter Form, d.h. sie sind in einzelne Text-Abschnitte unterteilt (z.B. Leitsätze, Entscheidungsformel, Begründung, Unterschriften). Manche Teile einer Entscheidung sind bewusst nicht enthalten (z.B. lange Zitate aus Gesetzen), weil diese nicht die eigentliche Aktivität des Gerichts wiedergeben. Die Nummerierung der Leitsätze und Absätze der Begründung sollte in der Regel (aber nicht immer!) der originalen Nummerierung in der PDF-Fassung entsprechen. Diese Fassung wurde aus den HTML-Dateien gewonnen und ist noch experimentell.
HTML	**Legal Tech/Quantitative Forschung.** Die HTML-Dokumente wie sie vom Bundesverfassungsgericht auf der amtlichen Webseite bereitgestellt werden, mit originalen Dateinamen.
TXT	**Subsidiär für alle Zielgruppen.** Diese Variante enthält die vollständigen aus den PDF-Dateien extrahierten Entscheidungstexte, aber nur einen reduzierten Umfang an Metadaten, der dem der PDF-Dateien entspricht. Die TXT-Dateien sind optisch an das Layout der PDF-Dateien angelehnt. Geeignet für qualitative Forscher:innen, die nur wenig Speicherplatz oder eine langsame Internetverbindung zur Verfügung haben oder für quantitative Forscher:innen, die beim Einlesen der CSV-Dateien Probleme haben.
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 (Allgemein)

5.1 Hinweise

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

5.2 Erläuterungen der einzelnen Variablen

Variable	Typ	Erläuterung
doc_id	String	(Nur CSV-Datei) Der Name der extrahierten TXT-Datei.
text	String	(Nur CSV-Datei) Der vollständige Inhalt der Entscheidung, so wie er in der von www.bundesverfassungsgericht.de heruntergeladenen PDF-Datei dokumentiert ist. In der segmentierten Variante stammt der Text aus der HTML-Dateien.
segment	String	(Nur segmentierte Variante) Das Segment der Entscheidung. Bezieht sich auf die Variable »text«. Segmentarten sind »leitsatz« (Leitsätze)), »gegenstand« (Entscheidungsgegenstand), »formel« (Entscheidungsformel), »tenor« (Tenor), »gruende« (Entscheidungsgründe, ggf. mit Anmerkung ob Sondervotum) und »unterschriften« (Unterschriften der Richter:innen). Die Erkennung von Sondervoten ist noch fehleranfällig. Einzelne Segmente sind mit einer Kombination aus Art und Ordinalzahl definiert, z.B. »gruende-133-sondervotum«.
gericht	String	In diesem Datensatz ist nur der Wert »BVerfG« vergeben. Dies ist der ECLI-Gerichtscode für »Bundesverfassungsgericht«. Diese Variable dient vor allem zur einfachen und transparenten Verbindung der Daten mit anderen Datensätzen.
datum	Datum (ISO)	Das Datum der Entscheidung 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.
entscheidung_typ	String	(Nur CSV-Datei) Der Typ der Entscheidung. Es sind die Werte »B« (Beschluss) und »U« (Urteil) vergeben. Wurde durch *regular expressions* aus der Variable »zitiervorschlag« berechnet.

Variable	Typ	Erläuterung
spruchkoerper_typ	String	Der Typ des Spruchkörpers. Es sind die Werte »K« (Kammer), »S« (Senat), »P« (Plenum) und »B« (Beschwerdekammer gem. § 97c BVerfGG) vergeben.
spruchkoerper_az	Natürliche Zahl	Der im Aktenzeichen angegebene Spruchkörper. Es sind nur die Werte »1« und »2« vergeben. Die Werte stehen für den 1. oder 2. Senat des Gerichts. Für Verzögerungsentscheidungen der Beschwerdekammer ist der Wert »NA«. **Achtung:** Um die Entscheidungen eines bestimmten Senats zu analysieren reicht es nicht, die Variable »spruchkoerper_az« zu nutzen, es muss zusätzlich noch die Variable »spruchkoerper_typ« auf »S« gesetzt werden, weil ansonsten noch mit dem Senat assoziierte Entscheidungen seiner Kammern und des Plenums mit ausgewählt werden.
registerzeichen	String	Das amtliche Registerzeichen. Es gibt die Verfahrensart an, in der die Entscheidung ergangen ist. Eine Erläuterung der Registerzeichen findet sich unter Punkt 7.
verfahrensart	String	Die ausführliche Beschreibung der Verfahrensart, die dem Registerzeichen zugeordnet ist. Eine Erläuterung der Registerzeichen und der zugehörigen Verfahrensarten findet sich unter Punkt 7.
eingangsnummer	Natürliche Zahl	Verfahren des gleichen Eingangsjahres erhalten vom Gericht eine Nummer in der Reihenfolge ihres Eingangs. Die Zahl ist in den Dateinamen mit führenden Nullen (falls <1000) codiert.
eingangsjahr_az	Natürliche Zahl	Das im Aktenzeichen angegebene Jahr in dem das Verfahren beim Gericht anhängig wurde. Das Format ist eine zweistellige Jahreszahl (YY).
eingangsjahr_iso	Natürliche Zahl	(Nur CSV-Datei) Das nach ISO-8601 codierte Jahr in dem das Verfahren beim Bundesverfassungsgericht anhängig wurde. Das Format ist eine vierstellige Jahreszahl (YYYY), um eine maschinenlesbare und eindeutige Jahreszahl für den Eingang zur Verfügung zu stellen. Wurde aus der Variable »eingangsjahr_az« durch den Autor des Datensatzes berechnet, unter der Annahme, dass Jahreszahlen über 50 dem 20. Jahrhundert zugeordnet sind und andere Jahreszahlen dem 21. Jahrhundert.

Variable	Typ	Erläuterung
entscheidungsjahr	Natürliche Zahl	(Nur CSV-Datei) Das Jahr in dem die Entscheidung ergangen ist. Das Format ist eine vierstellige Jahreszahl (YYYY). Wurde aus der Variable »datum« durch den Autor des Datensatzes berechnet.
kollision	String	In wenigen Fällen sind am gleichen Tag mehrere Entscheidungen zum gleichen Aktenzeichen ergangen. Diese werden ab der zweiten Entscheidung pro Tag durch eine Kollisions-ID mit einem Kleinbuchstaben ausdifferenziert. Für die erste Entscheidung ist der Wert der Variable »NA«, also nicht vorhanden. Die zweite Entscheidung ist mit »a« identifiziert, die dritte mit »b« und so fort. In der offiziellen Beschreibung der ECLI-Ordinalzahl wird diese Variable als »Kollisionsnummer« bezeichnet. Buchstaben sind allerdings keine Nummern, daher die abweichende Bezeichnung. In einem einzigen Fall ist tatsächlich eine »1« vergeben. Es handelt sich hier vermutlich um einen Fehler.
name	String	Der Name der Entscheidung. Für viele Entscheidungen aus der amtlichen Sammlung sind bekannte Namen vorhanden, diese wurden benutzt soweit möglich und auffindbar. Für weniger bekannte Entscheidungen wurde ein möglichst informativer Name vom Autor vergeben. Die konkrete Darstellung (ohne Leerzeichen, mit Bindestrichen usw.) ist Gründen der maschinellen Lesbarkeit geschuldet.
band	Natürliche Zahl	Der Band der amtlichen Sammlung in dem die Entscheidung veröffentlicht ist.
seite	Natürliche Zahl	Die genaue Fundstelle (Seitenzahl) der Entscheidung im jeweiligen Band der amtlichen Sammlung. Nur sinnvoll nutzbar im Zusammenspiel mit der Variable »band«.
aktenzeichen	String	(Nur CSV-Datei) Das amtliche Aktenzeichen. Die Variable wurde aus den Variablen »spruchkoerper_az«, »registerzeichen«, »eingangsnummer« und »eingangsjahr_az« durch den Autor des Datensatzes berechnet. Im Falle mehrere verbundener Verfahren mit einer einheitlichen Entscheidung ist dies das Aktenzeichen des Pilotverfahrens.

Variable	Typ	Erläuterung
aktenzeichen_alle	String	(Nur CSV-Datei) Alle Aktenzeichen der von der Entscheidung betroffenen Verfahren, falls es sich um verbunden Verfahren mit einheitlicher Entscheidung handelt. Ansonsten ist der Wert dieser Variable identisch mit der Variable »aktenzeichen«.
ecli	String	(Nur CSV-Datei) Der European Case Law Identifier (ECLI) der Entscheidung. Jeder Entscheidung ist eine einzigartige ECLI zugewiesen, ggf. mit Kollisisions-ID. Die ECLI ist vor allem dann hilfreich, wenn dieser Datensatz mit anderen Datensätzen zusammengeführt werden und Dopplungen vermieden werden sollen. Alle inhaltlichen Bestandteile der ECLI sind in diesem Datensatz zusätzlich auch anderen und besser verständlichen Variablen zugewiesen. Nutzen Sie bevorzugt diese anderen Variablen, statt Informationen aus der ECLI zu extrahieren. Die Variable wurde aus den Variablen »entscheidungsjahr«, »spruchkoerper_typ«, »datum«, »kollision«, »spruchkoerper_az«, »registerzeichen«, »eingangsnummer« und »eingangsjahr_az« durch den Autor des Datensatzes berechnet.
zitiervorschlag	String	Der vom BVerfG vorgegebene Zitiervorschlag.
kurzbeschreibung	String	Kurzbeschreibung des Inhalts des Verfahrens wie auf der Website des BVerfG angegeben.
pressemitteilung	String	Nummer und Datum der zugehörigen Pressemitteilung, falls vorhanden. Ansonsten »NA«.
praesi	String	(Nur CSV-Datei) Der Nachname des oder der Präsident:in in dessen/deren Amtszeit das Datum der Entscheidung fällt.
v_praesi	String	(Nur CSV-Datei) Der Nachname des oder der Vize-Präsident:in in dessen/deren Amtszeit das Datum der Entscheidung fällt.
richter	String	(Nur CSV-Datei) Die Nachnamen der Richter:innen, die die Entscheidung unterschrieben haben. Ggf. mit Angabe falls die Person verhindert war. Die einzelnen Namen sind jeweils durch vertikale Striche (»|«) voneinander getrennt.
zeichen	Natürliche Zahl	(Nur CSV-Datei) Die Anzahl Zeichen eines Dokumentes.

Variable	Typ	Erläuterung
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.
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.
saetze	Natürliche Zahl	(Nur CSV-Datei) Die Anzahl Sätze. Die Definition entspricht in etwa dem üblichen Verständnis eines Satzes. Die Regeln für die Bestimmung von Satzanfang und Satzende sind im Detail allerdings 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.
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 Gerichts abgerufen wurden.

Variable	Typ	Erläuterung
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.
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.
lizenz	String	Die Lizenz für den Gesamtdatensatz. In diesem Datensatz immer »Creative Commons Zero 1.0 Universal«.

6 Variablen (Linguistische Annotationen)

6.1 Datenstruktur

```
## Classes 'data.table' and 'data.frame':   25680501 obs. of  12 variables:
##  $ doc_id       : chr  "BVerfG_1951-09-09_S_2_BvQ_0001_51_NA_NA_NA_NA.txt" "
    BVerfG_1951-09-09_S_2_BvQ_0001_51_NA_NA_NA_NA.txt" "BVerfG_1951-09-09_S_2_BvQ
    _0001_51_NA_NA_NA_NA.txt" "BVerfG_1951-09-09_S_2_BvQ_0001_51_NA_NA_NA_NA.txt"
     ...
##  $ sentence_id  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ token_id     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ token        : chr  "BUNDESVERFASSUNGSGERICHT" "\n\n" "-" "2" ...
##  $ lemma        : chr  "BUNDESVERFASSUNGSGERICHT" "\n\n" "-" "2" ...
##  $ pos          : chr  "PROPN" "SPACE" "PUNCT" "NUM" ...
##  $ tag          : chr  "NE" "_SP" "$(" "CARD" ...
##  $ head_token_id: int  1 1 2 5 2 5 8 6 5 1 ...
##  $ dep_rel      : chr  "ROOT" "nk" "punct" "nk" ...
##  $ entity       : chr  "MISC_B" "MISC_I" "" "" ...
##  $ nounphrase   : chr  "beg_root" "" "" "" ...
##  $ whitespace   : logi  FALSE FALSE TRUE TRUE TRUE FALSE ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

6.2 Hinweise

Diese Variante des Datensatzes beruht nur auf den Variablen »doc_id« und »text« des regulären Datensatzes, die tokenisiert und mittels der Software »spacy«[7] mit linguistischen Annotationen versehen wurden.

Die Metadaten des Gesamtdatensatzes sind nicht in der linguistische annotierten Fassung enthalten, weil die Bereitstellung von Metadaten für jedes Token die Dateigröße und damit auch den RAM-Bedarf für Analysen gewaltig steigern würde. Um anhand der Metadaten Teilmengen der linguistischen Annotationen zu bilden, gehen sie bitte wie folgt vor:

1. CSV-Datei mit Metadaten einlesen
2. Anhand der Metadaten die gewünschte Teilmenge der Dokumente bilden
3. CSV-Datei mit Linguistischen Annotationen einlesen
4. Die Werte der Variable »doc_id« der Teilmenge nutzen um aus den annotierten Daten nur diejenigen herauszufiltern, deren Variable »doc_id« mit der Teilmenge übereinstimmt

[7] Die den Annotationen zugrundeliegende Software ist *spacy*, die hier verfügbar ist https://spacy.io/. Diese wird in R mittels des *spacyr* packages integriert: https://spacyr.quanteda.io/.

6.3 Erläuterung der Variablen

Variable	Typ	Erläuterung
doc_id	String	Der Dateiname des Dokumentes, aus dem die Tokens stammen. Identische Werte wie im Hauptdatensatz. Geeignet um Metadaten mit den linguistischen Annotationen zu verbinden.
sentence_id	Natürliche Zahl	Die Ordinalzahl des Satzes in dem Dokument, dem das Token zugeordnet ist.
token_id	Natürliche Zahl	Die Nummer des Tokens in einem Dokument.
token	String	Einzelne Tokens in der Reihenfolge ihres Vorkommens im jeweiligen Dokument.
lemma	String	Die lemmatisierte Form des Tokens.
pos	String	Grobe Annotation des einzelnen Tokens nach dem universal dependency POS tagset, siehe auch http://universaldependencies.org/u/pos/.
tag	String	Feine Annotation des einzelnen Tokens mit dem »de_core_news_sm«-Modell von spacy. Für eine detaillierte Erläuterung der Annotationen siehe: https://spacy.io/models/de
head_token_id	Natürliche Zahl	Das führende Token.
dep_rel	String	Die *dependency relation* zum head_token.
entity	String	Erkennung von benannten Entitäten (Named Entity Recoginition).
nounphrase	String	Erkennung von Nominalphrasen.
whitespace	Logisch	Ob es sich bei dem Token um Whitespace handelt oder nicht.

7 Registerzeichen

Die Tabelle der Registerzeichen und der ihnen zugeordneten Verfahrensarten stammt aus dem folgenden Datensatz: “Seán Fobbe (2021). Aktenzeichen der Bundesrepublik Deutschland (AZ-BRD). Version 1.0.1. Zenodo. DOI: 10.5281/zenodo.4569564.”

Registerzeichen	Verfahrensart
AR	Allgemeines Register: Vorverfahren oder sonstige Verfahrensarten
BvA	Verwirkung von Grundrechten
BvB	Verfassungswidrigkeit von Parteien
BvC	Wahlprüfungsverfahren
BvD	Anklage des Bundespräsidenten
BvE	Organstreitverfahren
BvF	Abstrakte Normenkontrolle
BvG	Bund-Länder-Streitigkeiten
BvH	Andere Streitigkeiten zwischen Bund und Ländern
BvJ	Anklage von Richtern des Bundesverfassungsgerichts
BvK	Landesverfassungsstreitigkeiten
BvL	Konkrete Normenkontrolle
BvM	Feststellung der Anwendbarkeit einer Regel des Völkergewohnheitsrechts
BvN	Divergenzvorlagen eines Landesverfassungsgerichts zur Auslegung des Grundgesetzes
BvO	Fortgeltung vorkonstitutionellen Rechts als Bundesrecht
BvP	Sonstige durch Bundesrecht zugewiesene Verfahren
BvQ	Einstweilige Anordnungen
BvR	Verfassungsbeschwerden; Kommunalverfassungsbeschwerden
BvT	Sonstige Verfahren
PBvS	Beendigung des Richteramtes am Bundesverfassungsgericht
PBvU	Plenarentscheidungen
PBvV	Rechtsgutachten
PKH	Prozesskostenhilfe
Vz	Verzögerungsbeschwerde

8 Präsident:innen

8.1 Hinweise

- Die Personaldaten stammen aus folgendem Datensatz: »Seán Fobbe and Tilko Swalve (2021). Presidents and Vice-Presidents of the Federal Courts of Germany (PVP-FCG). Version 2021-04-08. Zenodo. DOI: 10.5281/zenodo.4568682«.
- Das Datum bezieht sich jeweils auf das Amt als Präsident:in, nicht auf die Amtszeit als Richter:in.

8.2 Lebensdaten

Nachname	Vorname	Amtsantritt	Amtsende	Geboren	Gestorben
Höpker-Aschoff	Hermann	1951-09-07	1954-01-15	1883-01-31	1954-01-15
VACANCY-1	VACANCY-1	1954-01-16	1954-03-22	NA	NA
Wintrich	Josef	1954-03-23	1958-10-19	1891-02-15	1958-10-19
VACANCY-2	VACANCY-2	1958-10-20	1959-01-07	NA	NA
Müller	Gebhard	1959-01-08	1971-12-07	1900-04-17	1990-08-07
Benda	Ernst	1971-12-08	1983-12-19	1925-01-15	2009-03-02
Zeidler	Wolfgang	1983-12-20	1987-11-15	1924-09-02	1987-12-31
Herzog	Roman	1987-11-16	1994-06-30	1934-04-05	2017-01-10
VACANCY-3	VACANCY-3	1994-07-01	1994-09-13	NA	NA
Limbach	Jutta	1994-09-14	2002-04-09	1934-03-27	2016-09-10
Papier	Hans-Jürgen	2002-04-10	2010-03-15	1943-07-06	NA
Voßkuhle	Andreas	2010-03-16	2020-06-21	1963-12-21	NA
Harbarth	Stephan	2020-06-22	NA	1971-12-19	NA

8.3 Dienstalter und Lebensalter

Nachname	Vorname	Alter (Amtsantritt)	Alter (Amtsende)	Alter (Tod)
Höpker-Aschoff	Hermann	68	70	70
Wintrich	Josef	63	67	67
Müller	Gebhard	58	71	90
Benda	Ernst	46	58	84
Zeidler	Wolfgang	59	63	63
Herzog	Roman	53	60	82
Limbach	Jutta	60	68	82
Papier	Hans-Jürgen	58	66	NA
Voßkuhle	Andreas	46	56	NA
Harbarth	Stephan	48	NA	NA

9 Vize-Präsident:innen

9.1 Hinweise

- Die Personaldaten stammen aus folgendem Datensatz: »Seán Fobbe and Tilko Swalve (2021). Presidents and Vice-Presidents of the Federal Courts of Germany (PVP-FCG). Version 2021-04-08. Zenodo. DOI: 10.5281/zenodo.4568682«.
- Das Datum bezieht sich jeweils auf das Amt als Vize-Präsident:in, nicht auf die Amtszeit als Richter:in.

9.2 Lebensdaten

Nachname	Vorname	Amtsantritt	Amtsende	Geboren	Gestorben
Katz	Rudolf	1951-09-07	1961-07-23	1895-11-23	1961-07-23
Wagner	Friedrich Wilhelm	1961-12-19	1967-10-18	1894-02-28	1971-03-27
Seuffert	Walter	1967-10-18	1975-11-07	1907-02-04	1989-12-28
Zeidler	Wolfgang	1975-11-07	1983-12-20	1924-09-02	1987-12-31
Herzog	Roman	1983-12-20	1987-11-16	1934-04-05	2017-01-10
Mahrenholz	Ernst Gottfried	1987-11-16	1994-03-24	1929-06-18	2021-01-28
Limbach	Jutta	1994-03-24	1994-09-14	1934-03-27	2016-09-10
Henschel	Johann Friedrich	1994-09-29	1995-10-13	1931-06-10	2007-03-19
Seidl	Otto	1995-10-13	1998-02-27	1931-12-11	NA
Papier	Hans-Jürgen	1998-02-27	2002-04-10	1943-07-06	NA
Hassemer	Winfried	2002-04-10	2008-05-07	1940-02-17	2014-01-09
Voßkuhle	Andreas	2008-05-07	2010-03-16	1963-12-21	NA
Kirchhof	Ferdinand	2010-03-16	2018-11-30	1950-06-21	NA
Harbarth	Stephan	2018-11-30	2020-06-22	1971-12-19	NA
König	Doris	2020-06-22	NA	1957-06-25	NA

9.3 Dienstalter und Lebensalter

Nachname	Vorname	Alter (Amtsantritt)	Alter (Amtsende)	Alter (Tod)
Katz	Rudolf	55	65	65
Wagner	Friedrich Wilhelm	67	73	77
Seuffert	Walter	60	68	82
Zeidler	Wolfgang	51	59	63
Herzog	Roman	49	53	82
Mahrenholz	Ernst Gottfried	58	64	91
Limbach	Jutta	59	60	82
Henschel	Johann Friedrich	63	64	75
Seidl	Otto	63	66	NA
Papier	Hans-Jürgen	54	58	NA
Hassemer	Winfried	62	68	73
Voßkuhle	Andreas	44	46	NA
Kirchhof	Ferdinand	59	68	NA
Harbarth	Stephan	46	48	NA
König	Doris	62	NA	NA

10 Linguistische Kennzahlen

10.1 Erläuterung der Kennzahlen und Diagramme

Zur besseren Einschätzung des inhaltlichen Umfangs des Korpus dokumentiere ich an dieser Stelle die Verteilung der Werte für einige klassische linguistische 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 gelegentlich auch sinnlose Zeichenfolgen enthalten, weil es rein syntaktisch berechnet wird.
Typen	Einzigartige Tokens. Beispiel: wenn das Token »Verfassungsrecht« zehnmal in einer Entscheidung 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.

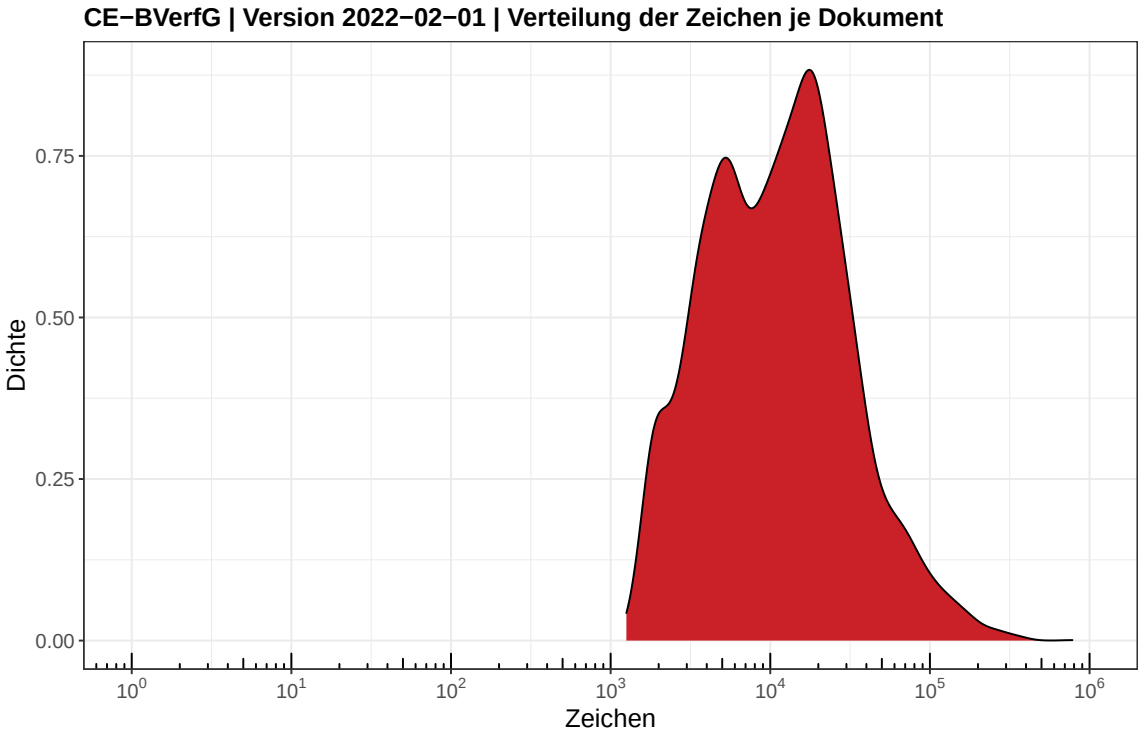
Es handelt sich bei den Diagrammen jeweils um »Density Charts«, die sich besonders dafür eignen die Schwerpunkte von Variablen mit stark schwankenden numerischen Werten zu visualisieren. Die Interpretation ist denkbar einfach: je höher die Kurve, desto dichter sind in diesem Bereich die Werte der Variable. 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.

10.2 Werte der Kennzahlen

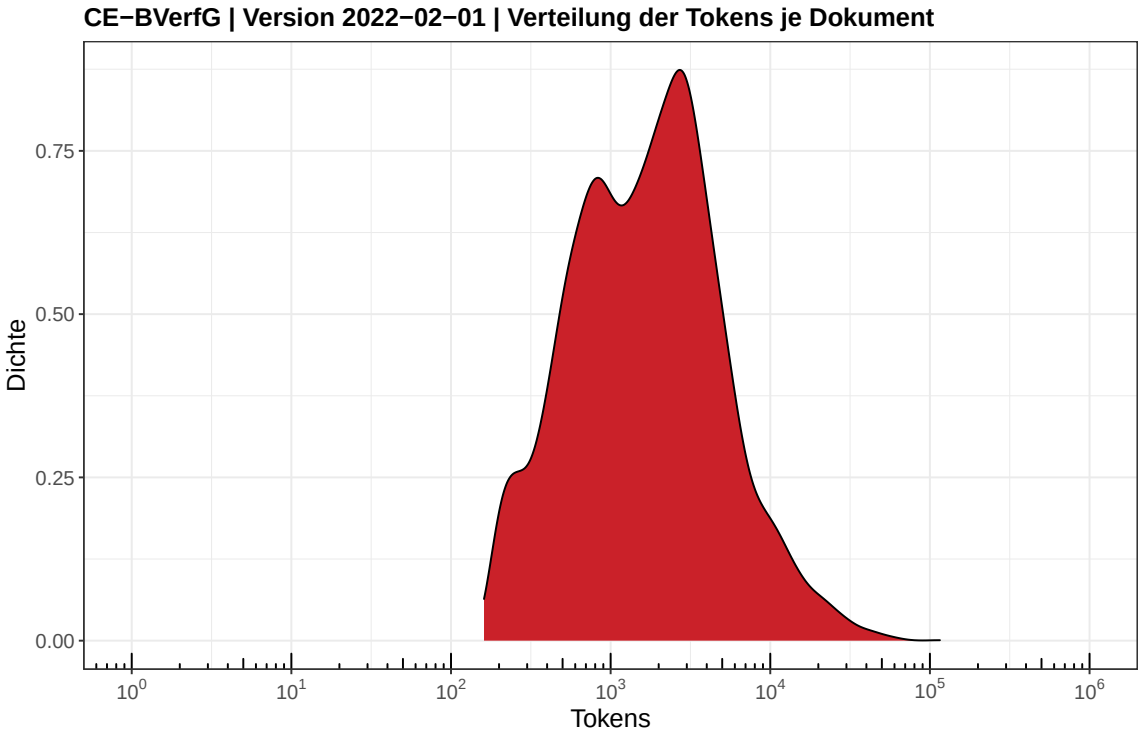
Kennzahl	Summe	Min	Quart1	Median	Mittel	Quart3	Max
zeichen	160,301,985	1,255	4,951	11,198	19,518.08	21,961	781,226
tokens	24,755,119	161	751	1,720	3,014.14	3,395	115,540
typen	281,164	91	307	591	777.20	983	13,491
saetze	1,411,651	6	45	105	171.88	199	5,282

10.3 Verteilung Zeichen



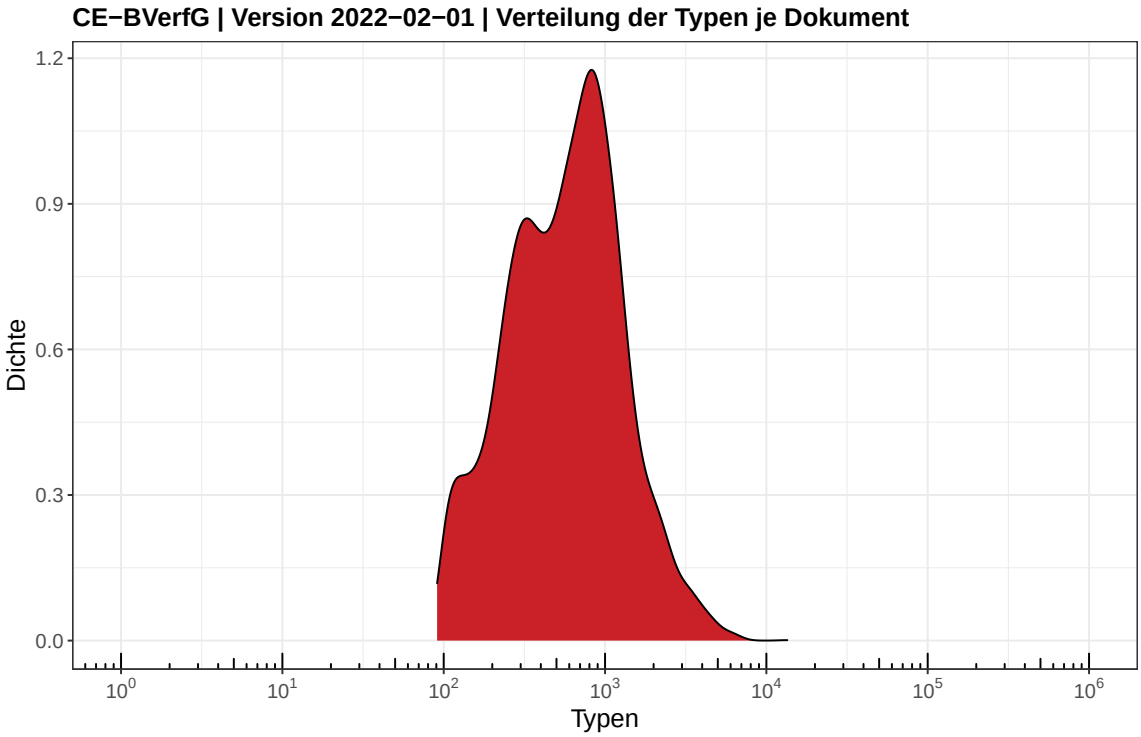
Fobbe | DOI: 10.5281/zenodo.5910152

10.4 Verteilung Tokens



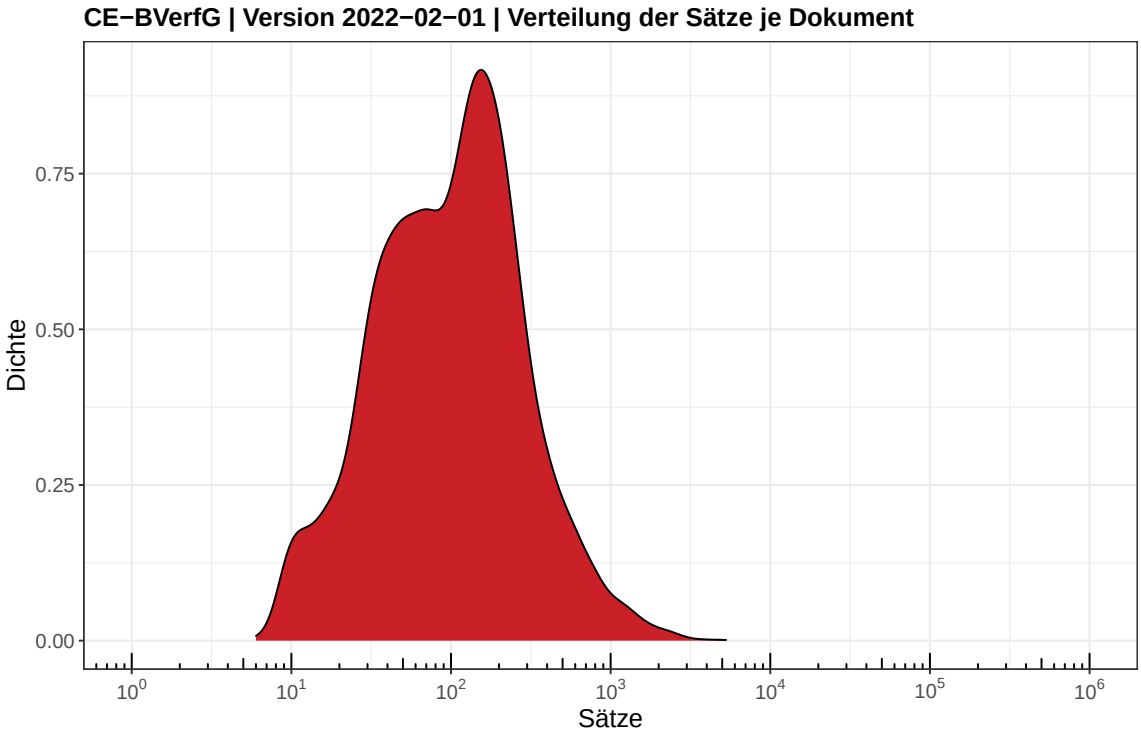
Fobbe | DOI: 10.5281/zenodo.5910152

10.5 Verteilung Typen



Fobbe | DOI: 10.5281/zenodo.5910152

10.6 Verteilung Sätze



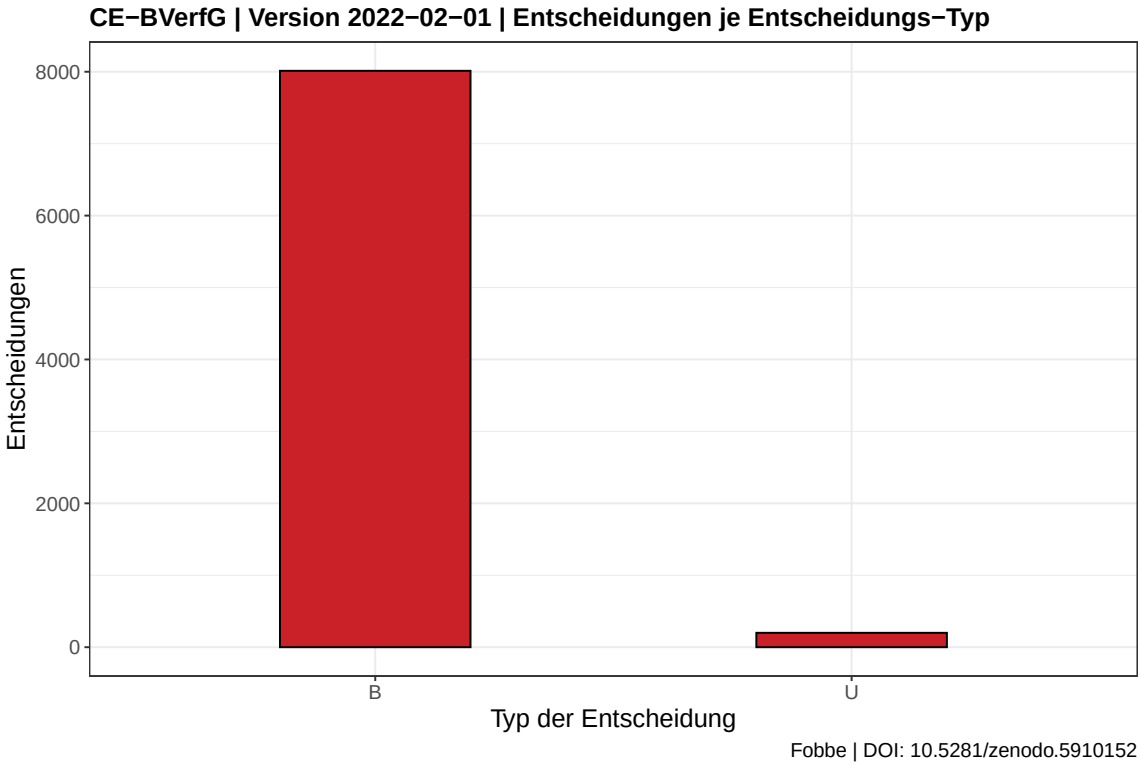
Fobbe | DOI: 10.5281/zenodo.5910152

11 Inhalt des Korpus

11.1 Zusammenfassung

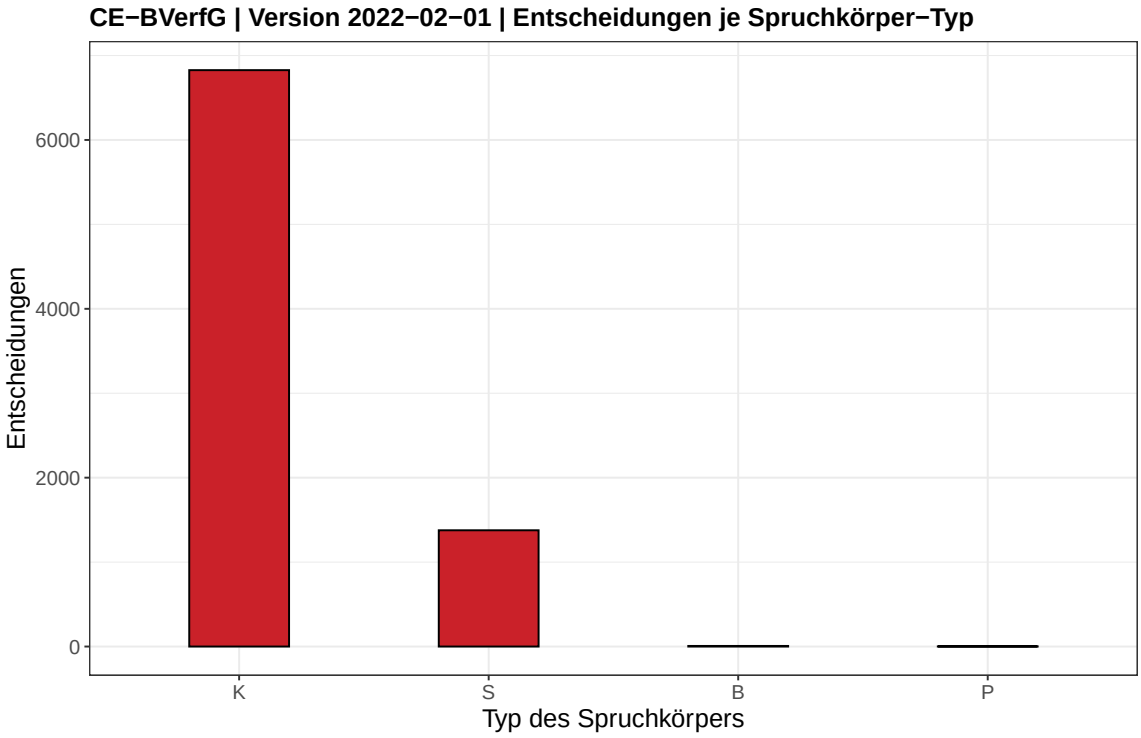
Variable	Anzahl	Min	Quart1	Median	Mittel	Quart3	Max
entscheidungsjahr	39	1951	2002	2009	2008.86	2016	2022
eingangsjahr_iso	46	1951	2001	2008	2007.30	2014	2022
band	67	1	106	120	121.10	136	155
eingangsnummer	2716	1	289	1054	1146.09	1866	3634

11.2 Nach Typ der Entscheidung



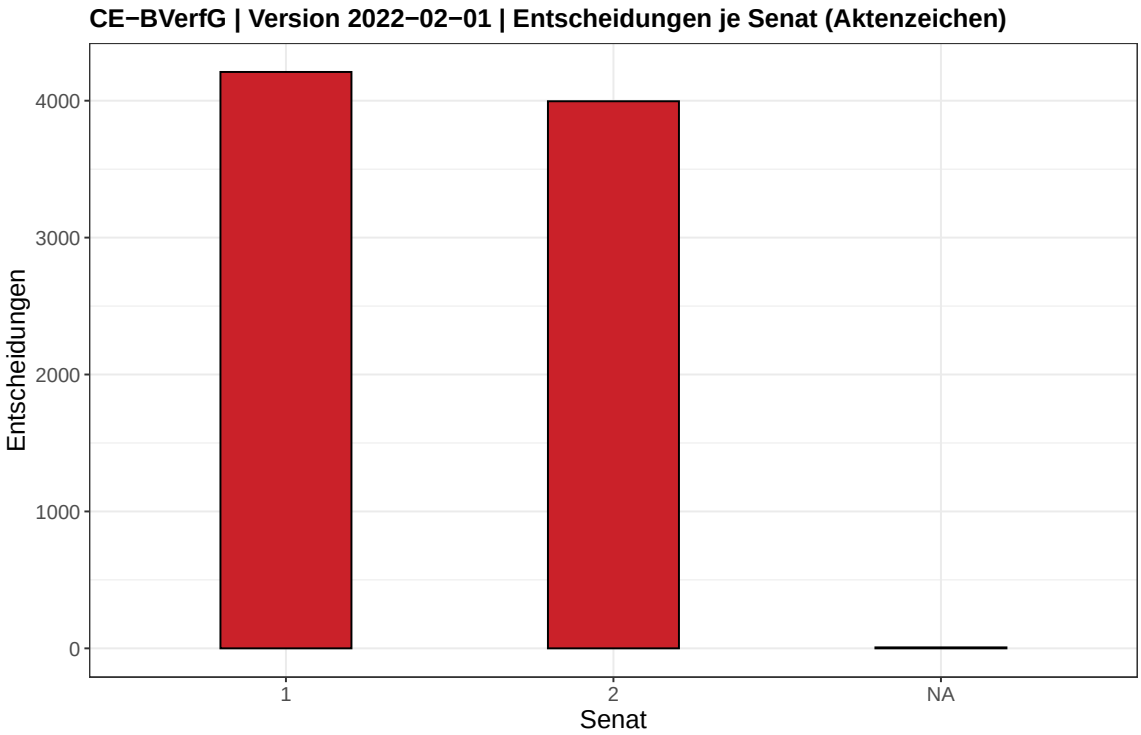
Typ	Entscheidungen	% Gesamt	% Kumulativ
B	8013	97.56	97.56
U	200	2.44	100.00
Total	8213	100.00	100.00

11.3 Nach Typ des Spruchkörpers



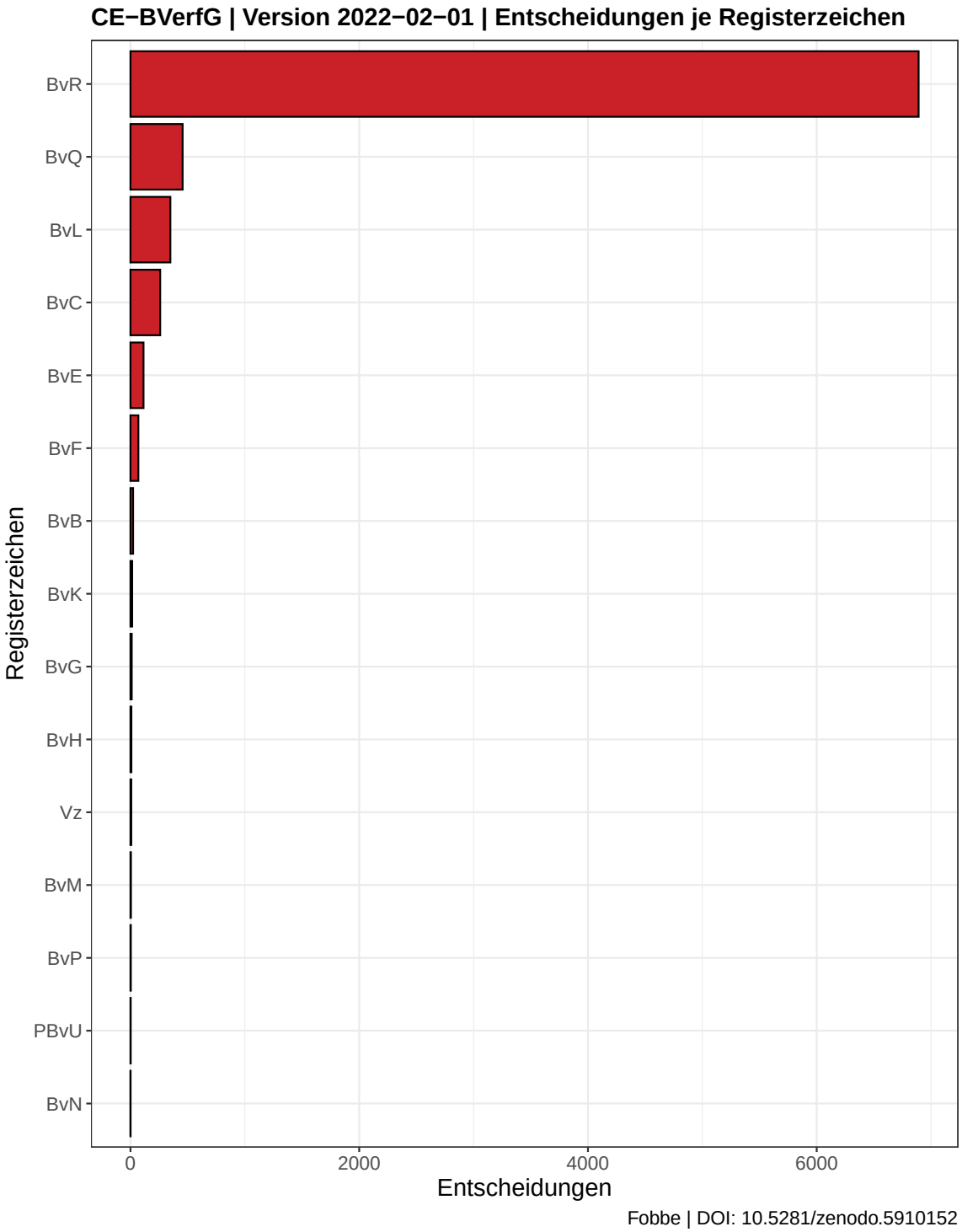
Typ	Entscheidungen	% Gesamt	% Kumulativ
B	7	0.09	0.09
K	6827	83.12	83.21
P	2	0.02	83.23
S	1377	16.77	100.00
Total	8213	100.00	100.00

11.4 Nach Spruchkörper (Aktenzeichen)



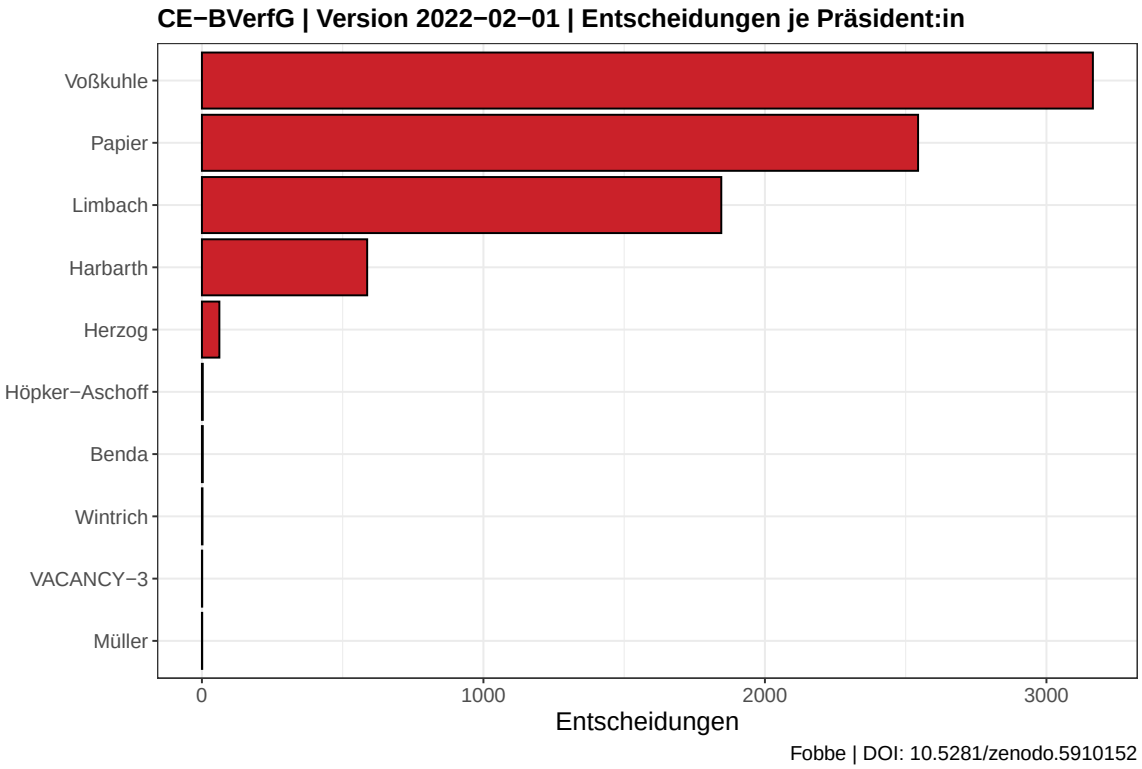
Senat	Entscheidungen	% Gesamt	% Kumulativ
NA	7	0.09	0.09
1	4210	51.26	51.35
2	3996	48.65	100.00
Total	8213	100.00	100.00

11.5 Nach Registerzeichen



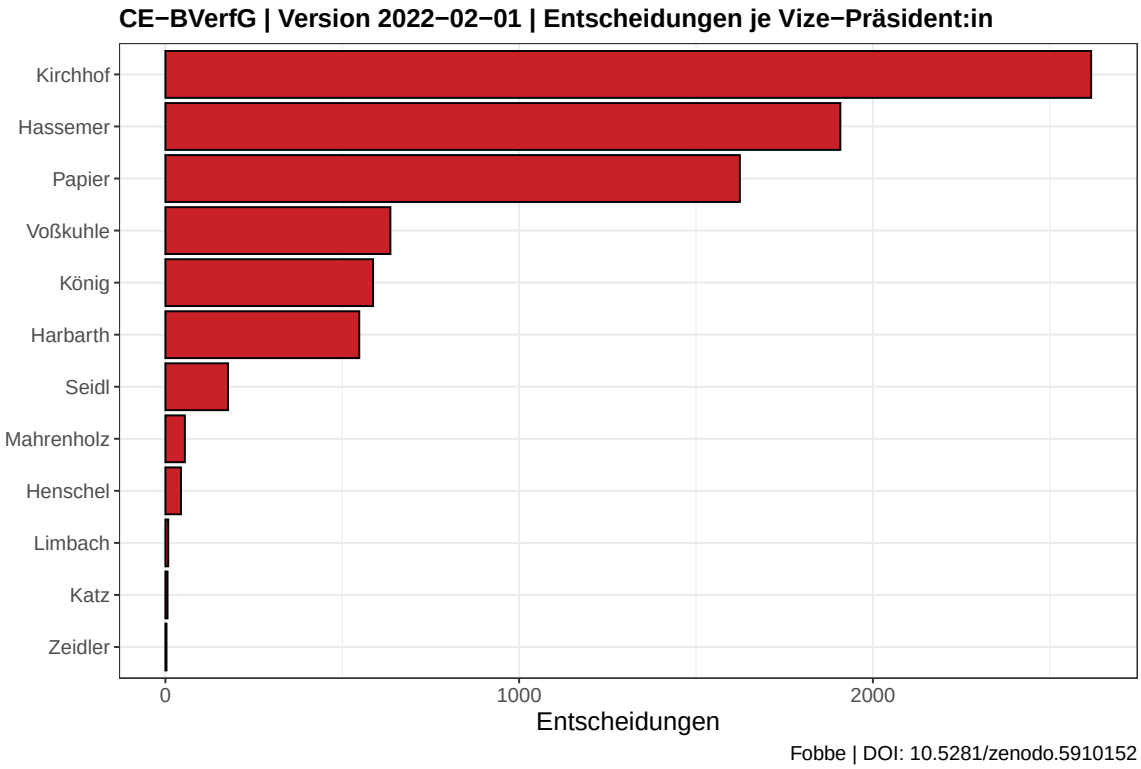
Registerzeichen	Entscheidungen	% Gesamt	% Kumulativ
BvB	23	0.28	0.28
BvC	260	3.17	3.45
BvE	114	1.39	4.83
BvF	69	0.84	5.67
BvG	11	0.13	5.81
BvH	8	0.10	5.91
BvK	15	0.18	6.09
BvL	349	4.25	10.34
BvM	4	0.05	10.39
BvN	1	0.01	10.40
BvP	3	0.04	10.43
BvQ	456	5.55	15.99
BvR	6891	83.90	99.89
PBvU	2	0.02	99.91
Vz	7	0.09	100.00
Total	8213	100.00	100.00

11.6 Nach Präsident:in



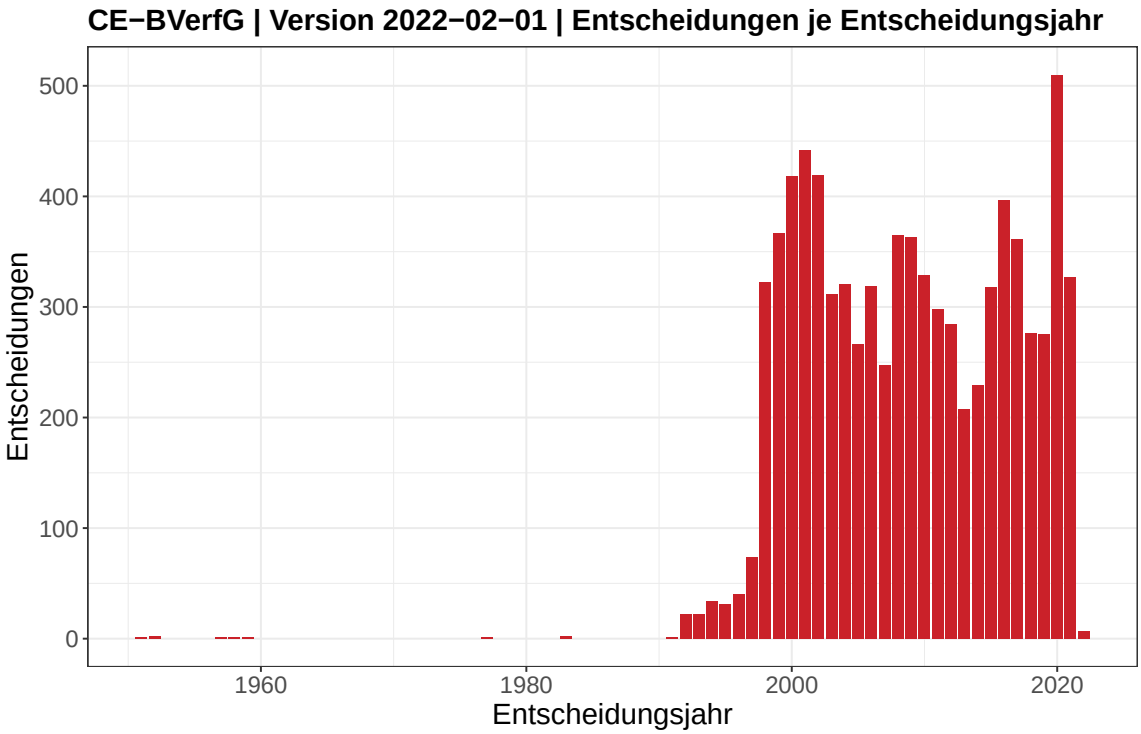
Präsident:in	Entscheidungen	% Gesamt	% Kumulativ
Benda	3	0.04	0.04
Harbarth	587	7.15	7.18
Herzog	62	0.75	7.94
Höpker-Aschoff	3	0.04	7.98
Limbach	1845	22.46	30.44
Müller	1	0.01	30.45
Papier	2544	30.98	61.43
VACANCY-3	1	0.01	61.44
Voßkuhle	3165	38.54	99.98
Wintrich	2	0.02	100.00
Total	8213	100.00	100.00

11.7 Nach Vize-Präsident:in



Vize-Präsident:in	Entscheidungen	% Gesamt	% Kumulativ
Harbarth	548	6.67	6.67
Hassemer	1908	23.23	29.90
Henschel	44	0.54	30.44
Katz	6	0.07	30.51
Kirchhof	2617	31.86	62.38
König	587	7.15	69.52
Limbach	8	0.10	69.62
Mahrenholz	55	0.67	70.29
Papier	1624	19.77	90.06
Seidl	177	2.16	92.22
Voßkuhle	636	7.74	99.96
Zeidler	3	0.04	100.00
Total	8213	100.00	100.00

11.8 Nach Entscheidungsjahr

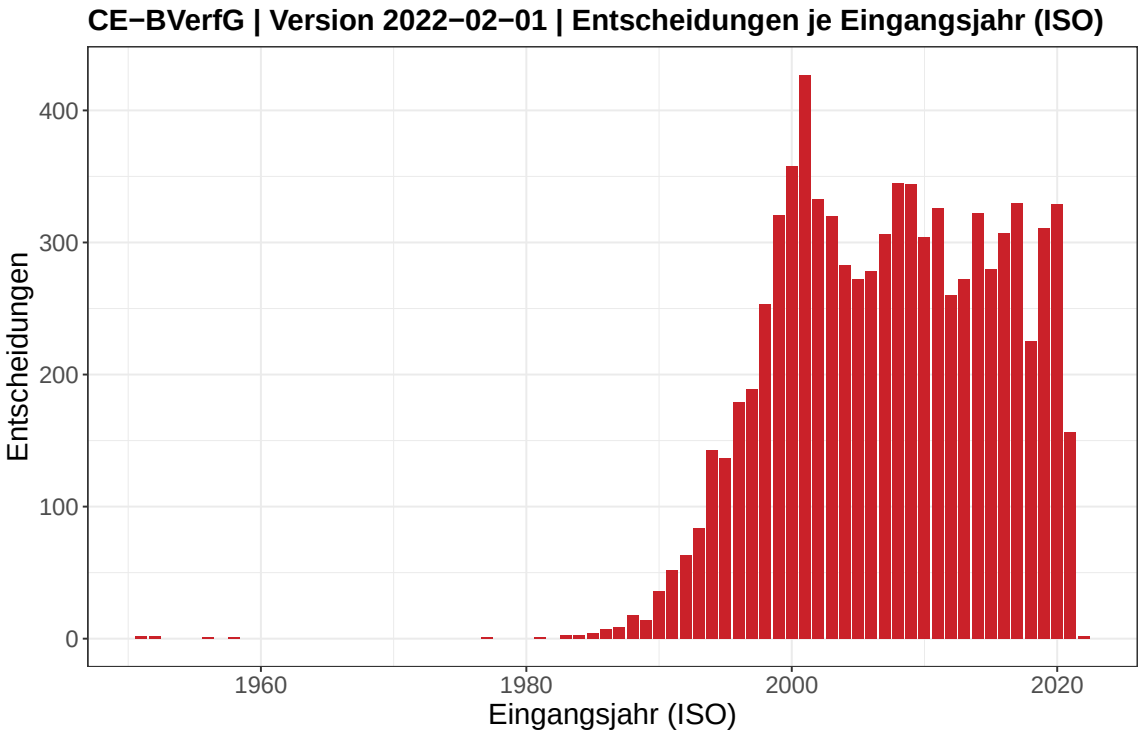


Jahr	Entscheidungen	% Gesamt	% Kumulativ
1951	1	0.01	0.01
1952	2	0.02	0.04
1957	1	0.01	0.05
1958	1	0.01	0.06
1959	1	0.01	0.07
1977	1	0.01	0.09
1983	2	0.02	0.11
1991	1	0.01	0.12
1992	22	0.27	0.39
1993	22	0.27	0.66
1994	34	0.41	1.07
1995	31	0.38	1.45
1996	40	0.49	1.94
1997	74	0.90	2.84

(continued)

Jahr	Entscheidungen	% Gesamt	% Kumulativ
1998	322	3.92	6.76
1999	367	4.47	11.23
2000	418	5.09	16.32
2001	442	5.38	21.70
2002	419	5.10	26.80
2003	312	3.80	30.60
2004	321	3.91	34.51
2005	266	3.24	37.75
2006	319	3.88	41.63
2007	247	3.01	44.64
2008	365	4.44	49.08
2009	363	4.42	53.50
2010	329	4.01	57.51
2011	298	3.63	61.13
2012	284	3.46	64.59
2013	208	2.53	67.13
2014	229	2.79	69.91
2015	318	3.87	73.79
2016	397	4.83	78.62
2017	361	4.40	83.01
2018	276	3.36	86.38
2019	275	3.35	89.72
2020	510	6.21	95.93
2021	327	3.98	99.91
2022	7	0.09	100.00
Total	8213	100.00	100.00

11.9 Nach Eingangsjahr (ISO)



Jahr	Entscheidungen	% Gesamt	% Kumulativ
1951	2	0.02	0.02
1952	2	0.02	0.05
1956	1	0.01	0.06
1958	1	0.01	0.07
1977	1	0.01	0.09
1981	1	0.01	0.10
1983	3	0.04	0.13
1984	3	0.04	0.17
1985	4	0.05	0.22
1986	7	0.09	0.30
1987	9	0.11	0.41
1988	18	0.22	0.63
1989	14	0.17	0.80
1990	36	0.44	1.24

(continued)

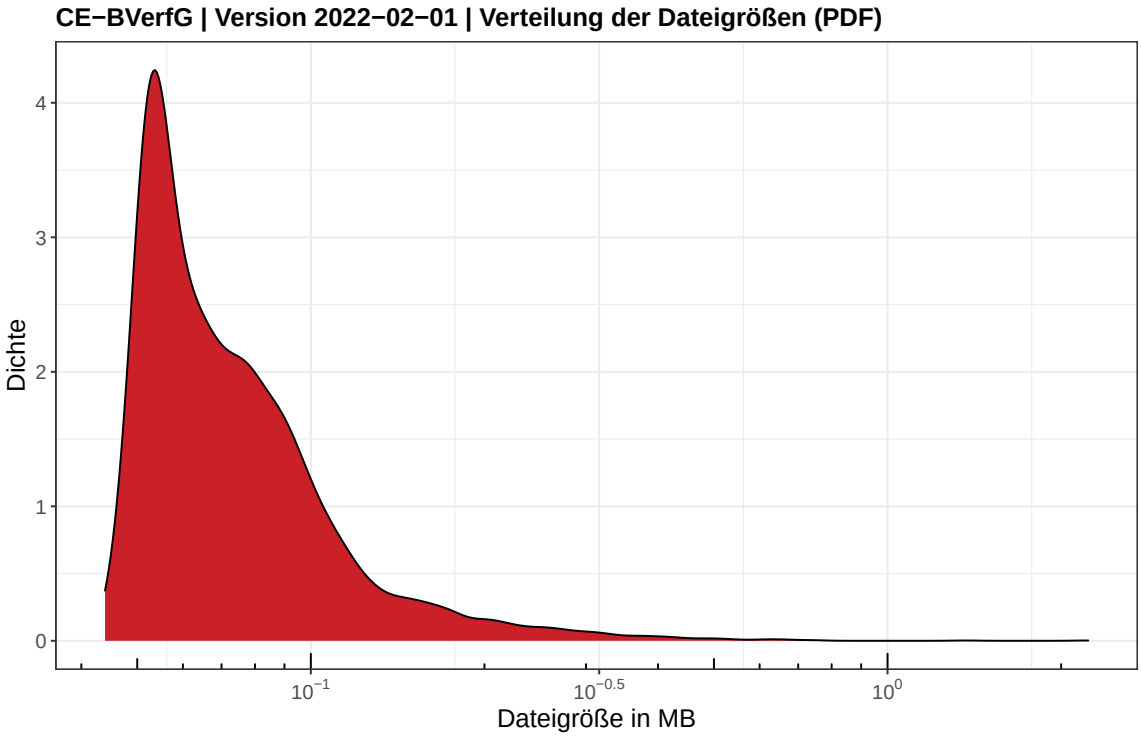
Jahr	Entscheidungen	% Gesamt	% Kumulativ
1991	52	0.63	1.88
1992	63	0.77	2.64
1993	84	1.02	3.66
1994	143	1.74	5.41
1995	137	1.67	7.07
1996	179	2.18	9.25
1997	189	2.30	11.55
1998	253	3.08	14.64
1999	321	3.91	18.54
2000	358	4.36	22.90
2001	427	5.20	28.10
2002	333	4.05	32.16
2003	320	3.90	36.05
2004	283	3.45	39.50
2005	272	3.31	42.81
2006	278	3.38	46.20
2007	306	3.73	49.92
2008	345	4.20	54.12
2009	344	4.19	58.31
2010	304	3.70	62.01
2011	326	3.97	65.98
2012	260	3.17	69.15
2013	272	3.31	72.46
2014	322	3.92	76.38
2015	280	3.41	79.79
2016	307	3.74	83.53
2017	330	4.02	87.54
2018	225	2.74	90.28
2019	311	3.79	94.07
2020	329	4.01	98.08
2021	156	1.90	99.98

(continued)

Jahr	Entscheidungen	% Gesamt	% Kumulativ
2022	2	0.02	100.00
Total	8213	100.00	100.00

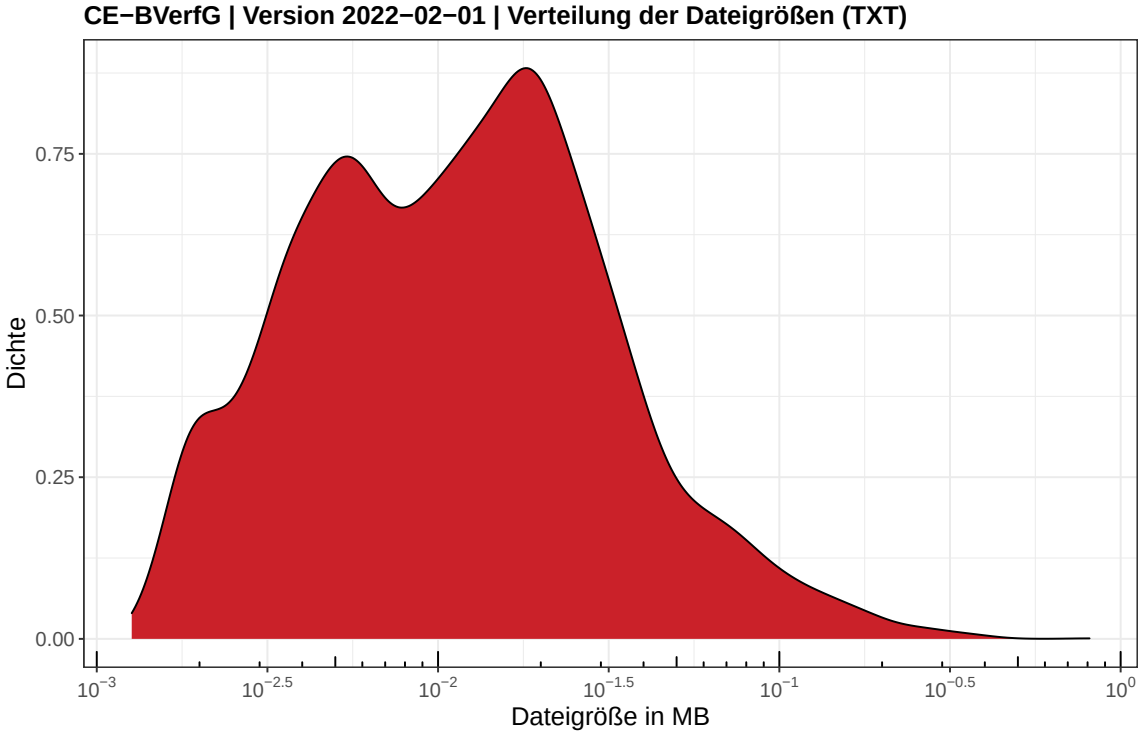
12 Dateigrößen

12.1 Verteilung PDF-Dateigrößen



Fobbe | DOI: 10.5281/zenodo.5910152

12.2 Verteilung TXT-Dateigrößen



Fobbe | DOI: 10.5281/zenodo.5910152

12.3 Gesamtgröße je ZIP-Archiv

Datei	Größe in MB
CE-BVerfG_2022-02-01_DE_ANALYSE.zip	1.41
CE-BVerfG_2022-02-01_DE_CSV_Annotiert.zip	299.01
CE-BVerfG_2022-02-01_DE_CSV_Datensatz.zip	46.38
CE-BVerfG_2022-02-01_DE_CSV_Metadaten.zip	0.90
CE-BVerfG_2022-02-01_DE_CSV_Segmentiert.zip	50.21
CE-BVerfG_2022-02-01_DE_HTML_Datensatz.zip	103.03
CE-BVerfG_2022-02-01_DE_PDF_Datensatz.zip	631.76
CE-BVerfG_2022-02-01_DE_TXT_Datensatz.zip	56.97
CE-BVerfG_2022-02-01_Source_Files.zip	0.12

13 Signaturprüfung

13.1 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 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 ist.

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 »Signatur verifiziert!«. Eine gescheiterte Prüfung meldet »FEHLER!«

13.2 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

13.3 Import: Public Key

```
system2("gpg2", "--import gpg/PublicKey_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"
```

13.4 Prüfung: GPG-Signatur der Hash-Datei

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

```
## [1] "output/CE-BVerfG_2022-02-01_KryptographischeHashes.csv"
```

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

```
## [1] "output/CE-BVerfG_2022-02-01_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 Tue 01 Feb 2022 09:05:33 PM CET
gpg: using RSA key FE6FB888F0E5656C1D253B9A50C41384F44A4E42
gpg: Good signature from Sean Fobbe <fobbe-data@posteo.de> [full]

13.5 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 <- file.path("output", 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(basename(filename), sha3.512.result)
```

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

Datei	Ergebnis
CE-BVerfG_2022-02-01_DE_ANALYSE.zip	Signatur verifiziert!
CE-BVerfG_2022-02-01_DE_CSV_Annotiert.zip	Signatur verifiziert!
CE-BVerfG_2022-02-01_DE_CSV_Datensatz.zip	Signatur verifiziert!
CE-BVerfG_2022-02-01_DE_CSV_Metadaten.zip	Signatur verifiziert!
CE-BVerfG_2022-02-01_DE_CSV_Segmentiert.zip	Signatur verifiziert!
CE-BVerfG_2022-02-01_DE_HTML_Datensatz.zip	Signatur verifiziert!
CE-BVerfG_2022-02-01_DE_PDF_Datensatz.zip	Signatur verifiziert!
CE-BVerfG_2022-02-01_DE_TXT_Datensatz.zip	Signatur verifiziert!
CE-BVerfG_2022-02-01_Source_Files.zip	Signatur verifiziert!

14 Changelog

14.1 Version 2022-02-01

- Vollständige Aktualisierung der Daten
- Strenge Versionskontrolle von R packages mit **renv**
- Kompilierung jetzt detailliert konfigurierbar, insbesondere die Parallelisierung
- Parallelisierung nun vollständig mit *future* statt mit *foreach* und *doParallel*
- Codebook-Erstellung stark beschleunigt durch Verwendung vorberechneter Diagramme
- Fehlerhafte Kompilierungen werden vor der nächsten Kompilierung vollautomatisch aufgeräumt
- Alle Ergebnisse werden automatisch fertig verpackt in den Ordner 'output' sortiert
- README und CHANGELOG sind jetzt externe Markdown-Dateien, die bei der Kompilierung automatisiert eingebunden werden
- Source Code des Changelogs zu Markdown konvertiert
- REGEX-Tests im Detail kommentiert

14.2 Version 2021-09-19

- Vollständige Aktualisierung der Daten
- Neue Variablen: Pressemitteilung, Zitiervorschlag, Aktenzeichen (alle), Kurzbeschreibung und Richter
- Neue Variante: Segmentiert
- Neue Variante: HTML
- Erweiterung der Codebook-Dokumentation
- Strenge Kontrolle und semantische Sortierung der Variablen-Namen
- Abgleich der selbst berechneten ECLI mit der in der HTML-Fassung dokumentierten ECLI
- Variable für Entscheidungstyp wird nun aus dem Zitiervorschlag berechnet um eine höhere Genaugikeit zu gewährleisten

14.3 Version 2021-05-20

- Vollständige Aktualisierung der Daten
- Einführung eines Debugging-Modus
- Einführung von Variablen für Verfahrensart, Lizenz, Typ der Entscheidung und Zeichenzahl
- Zusätzliche Diagramme für Typ der Entscheidung, Verteilung der Zeichen und Verteilung der Dateigrößen (TXT)
- Neue Datenquellen für Präsident:in, Vize-Präsident:in und für Registerzeichen/Verfahrensarten
- Zusammenfügen von über Zeilengrenzen getrennten Wörtern in der Variable »text« (nur CSV-Formate)
- Einige Verbesserungen im Codebook

14.4 Version 2021-01-08

- Vollständige Aktualisierung der Daten

- Veröffentlichung des vollständigen Source Codes
- Deutliche Erweiterung des inhaltlichen Umfangs des Codebooks
- Einführung der vollautomatischen Erstellung von Datensatz und Codebook
- Einführung von Compilation Reports um den Erstellungsprozess exakt zu dokumentieren
- Einführung von Variablen für Versionsnummer, Concept DOI, Version DOI, ECLI, Entscheidungsnamen, BVerfGE-Band, BVerfGE-Seite, Typ des Spruchkörpers, Präsident:in, Vize-Präsident:in und linguistische Kennzahlen (Tokens, Typen, Sätze)
- Automatisierung und Erweiterung der Qualitätskontrolle
- Einführung von Diagrammen zur Visualisierung von Prüfergebnissen
- Einführung kryptographischer Signaturen
- Alle Variablen sind nun in Kleinschreibung und Snake Case gehalten
- Variable »Suffix« in »kollision« umbenannt.
- Variable »Ordinalzahl« in »eingangsnummer« umbenannt.

14.5 Version 2020-08-03

- Vollständige Aktualisierung der Daten
- Angleichung der Variablen-Namen an andere Datensätze der CE-Serie[8]
- Einführung der Variable »Suffix« um weitere Entscheidungen korrekt erfassen zu können; aufgrund der fehlenden Berücksichtigung des Suffix sind die Metadaten von 36 Entscheidungen der Version 2020-06-20 fehlerhaft. Bitte verwenden Sie daher nur die neue Version. Alternativ können Sie die fehlerhaften Dateien (erkennbar an einem dreistelligen Eingangsjahr) aus der Analyse ausschließen oder per Hand korrigieren.

14.6 Version 2020-06-20

- Erstveröffentlichung

[8] Siehe: https://zenodo.org/communities/sean-fobbe-data/

15 Parameter für strenge Replikationen

```
## [1] "OpenSSL 1.1.1l  24 Aug 2021"
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora 34 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libflexiblas.so.3.0
##
## 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_2.7.3       future.apply_1.8.1 future_1.23.0      spacyr_1.2.1
##  [5] quanteda_3.2.0     readtext_0.81      data.table_1.14.2  scales_1.1.1
##  [9] ggplot2_3.3.5      pdftools_3.0.1     kableExtra_1.3.4   knitr_1.33
## [13] rvest_1.0.2        httr_1.4.2         mgsub_1.7.3        RcppTOML_0.1.7
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.7          here_1.0.1          svglite_2.0.0      lattice_0.20-45
##  [5] listenv_0.8.0       png_0.1-7           rprojroot_2.0.2    digest_0.6.29
##  [9] utf8_1.2.2          parallelly_1.30.0   R6_2.5.1           evaluate_0.14
## [13] highr_0.9           pillar_1.6.4        rlang_0.4.12       curl_4.3.2
## [17] rstudioapi_0.13     Matrix_1.4-0        reticulate_1.22    rmarkdown_2.11
## [21] qpdf_1.1            labeling_0.4.2      webshot_0.5.2      stringr_1.4.0
## [25] selectr_0.4-2       tinytex_0.36        munsell_0.5.0      compiler_4.0.5
## [29] xfun_0.29           pkgconfig_2.0.3     askpass_1.1        systemfonts
    _1.0.3
## [33] globals_0.14.0      htmltools_0.5.2     tibble_3.1.6       codetools_0.2-18
## [37] fansi_1.0.0         viridisLite_0.4.0   crayon_1.4.2       withr_2.4.3
## [41] rappdirs_0.3.3      grid_4.0.5          jsonlite_1.7.2     gtable_0.3.0
## [45] lifecycle_1.0.1     magrittr_2.0.1      RcppParallel_5.1.4 stringi_1.7.6
## [49] farver_2.1.0        renv_0.15.0         xml2_1.3.3         ellipsis_0.3.2
## [53] stopwords_2.3       vctrs_0.3.8         fastmatch_1.1-3    tools_4.0.5
## [57] glue_1.6.0          fastmap_1.1.0       yaml_2.2.1         colorspace_2.0-2
```

Literaturverzeichnis

Bengtsson, Henrik. 2021a. “A Unifying Framework for Parallel and Distributed Processing in R Using Futures.” https://journal.r-project.org/archive/2021/RJ-2021-048/index.html.

———. 2021b. “A Unifying Framework for Parallel and Distributed Processing in R Using Futures.” https://journal.r-project.org/archive/2021/RJ-2021-048/index.html.

———. 2021c. *Future.apply: Apply Function to Elements in Parallel Using Futures*. https://CRAN.R-project.org/package=future.apply.

———. 2021d. *Future: Unified Parallel and Distributed Processing in R for Everyone*. https://CRAN.R-project.org/package=future.

Benoit, Kenneth, and Akitaka Matsuo. 2020. *Spacyr: Wrapper to the spaCy 'Nlp' Library*. https://spacyr.quanteda.io.

Benoit, Kenneth, and Adam Obeng. 2021. *Readtext: Import and Handling for Plain and Formatted Text Files*. https://github.com/quanteda/readtext.

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, and William Lowe. 2021. *Quanteda: Quantitative Analysis of Textual Data*. https://quanteda.io.

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

Eddelbuettel, Dirk. 2020. *RcppTOML: Rcpp Bindings to Parser for Tom's Obvious Markup Language*. http://dirk.eddelbuettel.com/code/rcpp.toml.html.

Ewing, Mark. 2021. *Mgsub: Safe, Multiple, Simultaneous String Substitution*. https://CRAN.R-project.org/package=mgsub.

Ooms, Jeroen. 2021. *Pdftools: Text Extraction, Rendering and Converting of Pdf Documents*. https://CRAN.R-project.org/package=pdftools.

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

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

———. 2020. *Httr: Tools for Working with Urls and Http*. https://CRAN.R-project.org/package=httr.

———. 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. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

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.