

Corpus des Deutschen Bundesrechts (C-DBR)

CODEBOOK

Version 2023-01-05



DOI: [10.5281/zenodo.7494474](https://doi.org/10.5281/zenodo.7494474)

Titel	Corpus des Deutschen Bundesrechts
Abkürzung	C-DBR
Autor	Seán Fobbe
Version	2023-01-05
Download	https://doi.org/10.5281/zenodo.7494474
Lizenz	CC0 1.0 Universal

Zitiervorschlag

Seán Fobbe (2023). Corpus des Deutschen Bundesrechts (C-DBR). Version 2023-01-05. Zenodo. DOI: [10.5281/zenodo.7494474](https://doi.org/10.5281/zenodo.7494474).

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 2023-01-05. 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.3832111](https://doi.org/10.5281/zenodo.3832111). Die »Concept DOI« verlinkt immer die aktuellste Version.

Urheberrecht

Der Datensatz und dieses Dokument sind unter einer **Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication Lizenz** veröffentlicht. Ich stelle den Datensatz und das Codebook vollständig gemeinfrei und verzichte weltweit auf alle damit verbundenen Urheberrechte, einschließlich aller ähnlichen Rechte, soweit dies gesetzlich möglich ist.

Sie können die Werke kopieren, modifizieren, verteilen und aufführen ohne um Erlaubnis bitten zu müssen, selbst für kommerzielle Zwecke. Patente und Markenschutzrechte bleiben von CC0 unberührt. CC0 hat auch keine Auswirkungen auf etwaige Datenschutz- oder Persönlichkeitsrechte. Jegliche Haftung für die Benutzung dieses Werkes ist ausgeschlossen, bis zu dem maximalen Umfang in dem dies gesetzlich möglich ist.

Wenn Sie diese Werke nutzen oder zitieren sollten Sie nicht den Eindruck erwecken, der Autor unterstütze ihre Nutzung.

Dies ist nur eine unverbindliche deutsche Zusammenfassung der Lizenz, den vollständigen und rechtsverbindlichen Lizenztext finden Sie hier: <https://creativecommons.org/publicdomain/zero/1.0/legalcode>

Disclaimer

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

Inhaltsverzeichnis

1	Einführung	5
2	Nutzung	6
2.1	CSV-Dateien	6
2.2	XML-Dateien	6
2.3	TXT-Dateien	6
3	Konstruktion	7
3.1	Beschreibung des Datensatzes	7
3.2	Datenquellen	7
3.3	Sammlung der Daten	7
3.4	Source Code und Compilation Report	7
3.5	Grenzen des Datensatzes	9
3.6	Urheberrechtsfreiheit von Rohdaten und Datensatz	9
3.7	Metadaten	9
3.7.1	Schema für die Dateinamen (PDF, TXT, EPUB)	9
3.7.2	Beispiel eines Dateinamens	9
3.8	Qualitätsprüfung	10
4	Varianten und Zielgruppen	11
5	Variablen	14
5.1	Hinweise	14
5.2	Erläuterungen der einzelnen Variablen	14
5.3	Konkordanztabelle: XML-Struktur und CSV-Variablen	21
6	Linguistische Kennzahlen	22
6.1	Erläuterung der Kennzahlen	22
6.2	Kennzahlen: Einzelnormen	22
6.3	Kennzahlen: Rechtsakte	22
6.4	Verteilung Zeichen	23
6.5	Verteilung Tokens	24
6.6	Verteilung Typen	25
6.7	Verteilung Sätze	26
7	Inhalt	27
7.1	Nach Periodikum	27
7.1.1	Einzelnormen	27
7.1.2	Rechtsakte mit veröffentlichtem Normtext	29
7.1.3	Alle Rechtsakte (mit und ohne Normtext)	31
7.2	Nach Ausfertigungsjahr	34
7.2.1	Einzelnormen	34
7.2.2	Rechtsakte mit veröffentlichtem Normtext	39
7.2.3	Alle Rechtsakte (mit und ohne Normtext)	44
8	Dateigrößen	49
9	Kryptographische Signaturen	51
9.1	Zwei-Phasen-Signatur	51

9.2	Persönliche GPG-Signatur	51
10	Changelog	52
10.1	Version 2023-01-05	52
10.2	Version 2022-08-05	52
10.3	Version 2022-05-22	52
10.4	Version 2022-01-12	52
10.5	Version 2021-09-16	53
10.6	Version 2021-07-30	53
10.7	Version 2021-01-05	53
10.8	Version 2020-10-09	53
10.9	Version 2020-07-08	53
10.10	Version 2020-05-18	54
11	Parameter für strenge Replikationen	55
	Literaturverzeichnis	57

1 Einführung

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).

Bundesgesetze werden vom Bundestag im Zusammenwirken mit dem Bundesrat erlassen und vom Bundespräsidenten ausgefertigt (Art. 76 ff GG). Das Initiativrecht liegt bei Abgeordneten aus der Mitte des Bundestags, der Bundesregierung und dem Bundesrat (Art. 76 Abs. 1 GG). Der Bundesrat ist je nach Gesetzescharakter mit einem Zustimmungserfordernis oder einem Einspruchsrecht beteiligt (Art. 77, 78 GG).

Verordnungen werden in der Regel von der Exekutive erlassen, in seltenen Fällen vom Bundestag selbst. Durch Bundesgesetz können nur Bundesregierung, Bundesminister oder Landesregierungen hierzu ermächtigt werden (Art. 80 Abs. 1 S.1 GG), eine im Gesetz vorgesehene Sub-Delegation ist aber möglich (Art. 80 Abs. 1 S. 4 GG). Verordnungen müssen einem speziellen Bestimmtheitsgebot genügen und ihre Rechtsgrundlage in der Verordnung angeben (Art. 80 Abs. 1 S. 2 und 3). Der Erlass von Verordnungen erfordert zudem nicht selten die Zustimmung des Bundesrates, entweder aufgrund von Art. 80 Abs. 2 GG oder bedingt durch eine Regelung in einem einfachen Bundesgesetz.

Warum dieser Datensatz? Die quantitative Analyse von juristischen Texten, insbesondere von Gesetzen und Verordnungen, 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.⁴ Wissenschaftliche Initiativen wie der Juristische Referenzkorpus (JuReKo) sind nach jahrelanger Arbeit hinter verschlossenen Türen verschwunden.

In einem funktionierenden Rechtsstaat muss die Rechtsetzung 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.⁵

¹ Positive Ausnahmen finden sich vor allem unter: <https://www.quantitative-rechtswissenschaft.de/>

² <https://opinioiuris.de/>

³ <https://openjur.de/>

⁴ Openjur beabsichtigt eine API anzubieten, diese war aber im Januar 2021 immernoch nicht verfügbar. Openjur ist seit 2008 in Betrieb.

⁵ 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, XML, TXT, PDF, EPUB) 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. Die Nutzung der CSV-Varianten ist aus Qualitätsgründen und den umfangreicheren Metadaten *empfohlen*. CSV⁶ 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 Einzelnorm bzw. einem Rechtsakt (je nach Variante). Die Variablen sind unter Punkt 5 genauer erläutert.

Zum Einlesen empfehle ich für **R** 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 Beispiel:

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

2.2 XML-Dateien

Das Einlesen der **XML-Rohdaten** ist komplex und die Entscheidung welche XML-Nodes zu extrahieren sind wird ganz erheblich von der Forschungsfrage beeinflusst. Falls Sie über vertiefte XML-Kenntnisse verfügen, sollten Sie eine eigenständige Extraktion dennoch in Erwägung ziehen, weil sie so die Datenanalyse besser auf Ihre Bedürfnisse anpassen können. Lesen Sie hierfür bitte die Document Type Definition (DTD) genau und greifen Sie ggf. auf den im Source Code zur Verfügung gestellten XML Parser zurück.

2.3 TXT-Dateien

Die TXT-Dateien enthalten nur sehr rudimentäre Metadaten! Benutzen Sie daher für statistische Analysen vorzugsweise die CSV- oder XML-Dateien. Die **TXT-Dateien** inklusive Metadaten können zum Beispiel mit **R** und dem package **readtext** (via CRAN verfügbar) eingelesen werden. Ein Vorschlag:

```
library(readtext)
txt.dbr <- readtext("./*.txt",
                   docvarsfrom = "filenames",
                   docvarnames = c("kurztitel",
                                   "langtitel"),
                   dvsep = "_",
                   encoding = "UTF-8")
```

⁶ Das CSV-Format ist in RFC 4180 definiert, siehe <https://tools.ietf.org/html/rfc4180>

3 Konstruktion

3.1 Beschreibung des Datensatzes

Der Datensatz ist eine möglichst vollständige Sammlung der konsolidierten Fassungen aller Gesetze und Verordnungen auf Bundesebene. Änderungsgesetze und -verordnungen sind nicht enthalten. Er enthält alle Rechtsakte, die auf der amtlichen Webseite »Gesetze im Internet« des Bundesministerium des Justiz am jeweiligen Stichtag verfügbar waren. Die Stichtage für jede Version sind in der Versionsnummer festgehalten.

Zusätzlich zu den einfach maschinenlesbaren Formaten (CSV und TXT) sind die XML-, PDF- und EPUB-Rohdaten enthalten, damit Analysten gegebenenfalls ihre eigene Konvertierung vornehmen können. Die Rohdaten wurden inhaltlich nicht verändert. Die PDF- und EPUB-Varianten der Rechtsakte sollen primär traditionelle juristische Forschung und *mixed methods*-Ansätze unterstützen.

In diesem Datensatz sind nur Rechtsakte mit Außenwirkung (d.h. das Grundgesetz, Bundesgesetze und Bundesverordnungen) enthalten. Verwaltungsvorschriften sind nicht Teil des Datensatzes.

3.2 Datenquellen

Datenquelle	Fundstelle
Primäre Datenquelle	https://www.gesetze-im-internet.de/
Source Code	https://doi.org/10.5281/zenodo.7494477

3.3 Sammlung der Daten

Die Daten wurden vollautomatisiert gesammelt und mit Abschluss der Verarbeitung kryptographisch signiert. Die Webseite des Justizministeriums ist laut dem Reiter »Hinweise«⁷ ausdrücklich für die vollautomatisierte Datensammlung freigegeben. Der Abruf geschieht ausschließlich über TLS-verschlüsselte Verbindungen.

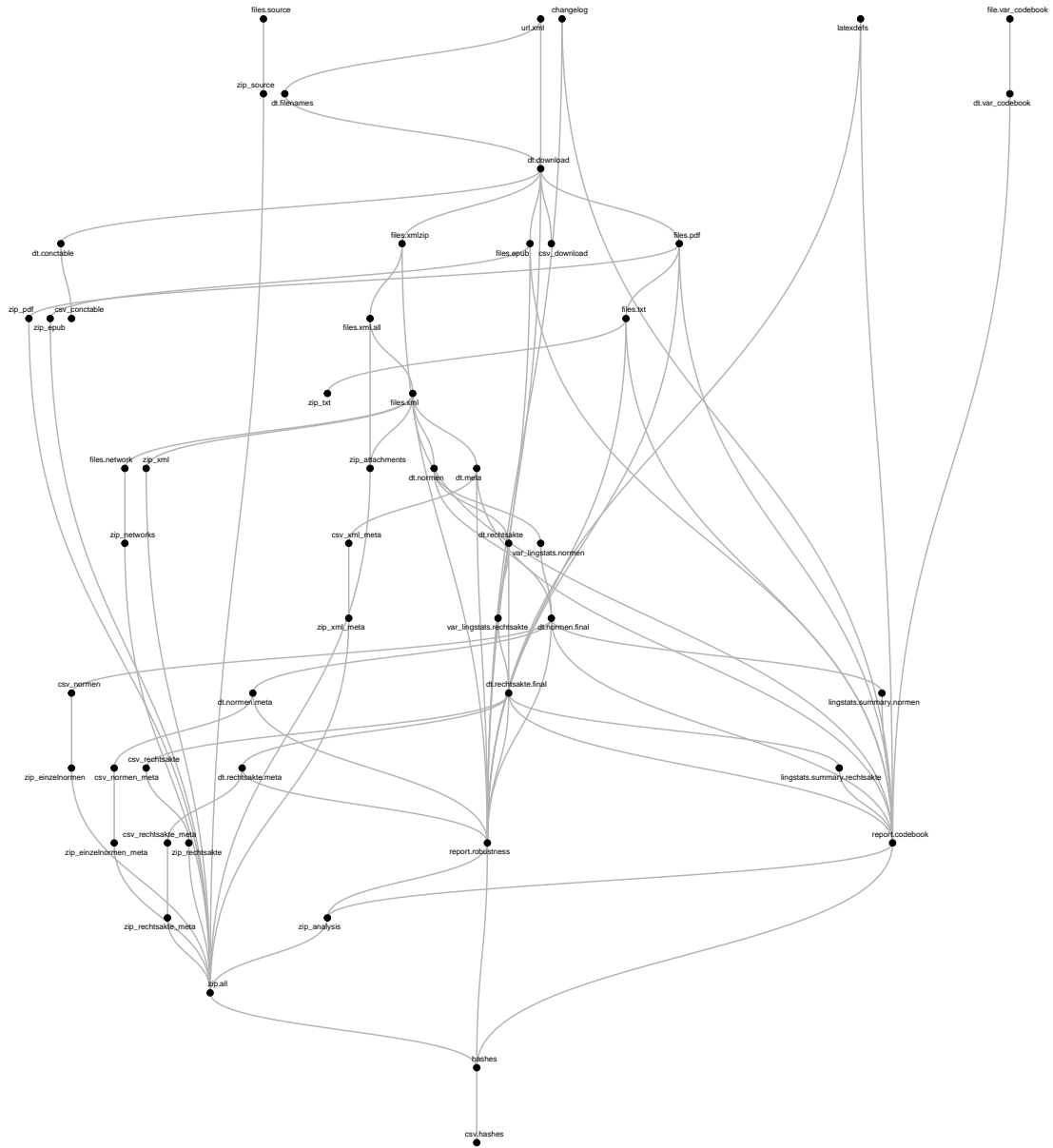
3.4 Source Code und Compilation Report

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 unter dieser Adresse hinterlegt: <https://doi.org/10.5281/zenodo.7494477>

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.

⁷ <https://www.gesetze-im-internet.de/hinweise.html>

C-DBR | Version 2023-01-05 | Vollständiger Prozess der Datensatz-Kompilierung



Föbbe | DOI: 10.5281/zenodo.7494474

Abbildung 1: Der vollständige Prozess der Datensatz-Kompilierung.

3.5 Grenzen des Datensatzes

Nutzer sollten folgende wichtige Grenzen beachten:

- Der Datensatz enthält nur das, was das Bundesjustizministerium auch tatsächlich veröffentlicht (*publication bias*). Es fehlen insbesondere Änderungsgesetze und -verordnungen. Manche Rechtsakte sind zudem nur mit den Metadaten nachgewiesen — ihr Inhalt fehlt aus technischen Gründen oder weil in der Bereinigten Sammlung Bundesgesetzblatt Teil III nur bibliographische Angaben enthalten sind (§ 3 Abs. 2 BRSG).⁸
- Es kann aufgrund technischer Grenzen bzw. Fehler sein, dass manche — im Grunde verfügbare — Rechtsakte nicht oder nicht korrekt abgerufen werden (*automation bias*).
- Es sind nur am Tag des Abrufs veröffentlichte konsolidierte Rechtsakte enthalten, eine diachronische Untersuchung muss somit mehrere verfügbare Versionen auswerten (*temporal bias*).

3.6 Urheberrechtsfreiheit von Rohdaten und Datensatz

An den Rechtsakten und Rechtsnormen 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«).

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

Alle Metadaten wurden aus den XML-Rohdaten zeitgleich mit dem Text der Normen extrahiert. Der volle Satz an Metadaten ist nur in den CSV-Dateien enthalten. Bitte beachten Sie, dass bei weitem nicht alle XML-Nodes ausgewertet wurden. Viele Nodes enthalten nur optische Informationen und wurden deshalb ignoriert. Manche Nodes (z.B. einzelne Absätze, Listen) wurden nicht extrahiert, weil nicht alle Normen in Absätze und Listen unterteilt sind und die Bereitstellung in einem nicht-hierarchischen Format wie CSV keine Vorteile gegenüber dem XML-Format bringen würde.

Die Dateinamen der PDF-, TXT und EPUB-Dateien enthalten nur eine Abkürzung und einen modifizierten Langtitel (auf 200 Zeichen gekürzt und um Sonderzeichen bereinigt). Diese wurden aus den jeweiligen Header-Markierungen der HTML-Seiten extrahiert.

3.7.1 Schema für die Dateinamen (PDF, TXT, EPUB)

[Abkürzung]_[modifizierter_Langtitel]

3.7.2 Beispiel eines Dateinamens

2.WasSV_ZweiteWassersicherstellungsverordnung.pdf

⁸ <https://www.gesetze-im-internet.de/hinweise.html>

3.8 Qualitätsprüfung

Insgesamt werden zusammen mit jeder Kompilierung Dutzende Tests zur Qualitätsprüfung durchgeführt. Alle Ergebnisse der Qualitätsprüfungen sind aggregiert im Compilation Report und im Robustess Checks-Bericht, sowie einzeln im Archiv »Analyse« zusammen mit dem Datensatz veröffentlicht.

4 Varianten und Zielgruppen

Dieser Datensatz ist in verschiedenen Varianten verfügbar, die sich an unterschiedliche Zielgruppen richten. Zielgruppe sind nicht nur quantitativ forschende RechtswissenschaftlerInnen, sondern auch traditionell arbeitende JuristInnen. Idealerweise müssen quantitative Methoden ohnehin immer durch qualitative Interpretation, Bildung von Theorien 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
CSV_Einzelnormen_Datensatz	Legal Tech/Quantitative Forschung. Diese CSV-Datei ist eine der für statistische Analysen empfohlenen Varianten des Datensatzes. Sie enthält den Volltext aller Rechtsakte, disaggregiert nach Einzelnormen, sowie alle in diesem Codebook beschriebenen Metadaten. Enthält nur Rechtsakte, für die mindestens eine Einzelnorm mit Normtext veröffentlicht wurde!
CSV_Einzelnormen_Metadaten	Legal Tech/Quantitative Forschung. Wie die andere CSV-Datei mit Einzelnormen, nur ohne die Normtexte. Sinnvoll für Analyst:innen, die sich nur für die Metadaten interessieren und Speicherplatz sparen wollen. Enthält nur Rechtsakte, für die mindestens eine Einzelnorm mit Normtext veröffentlicht wurde!
CSV_Rechtsakte_Datensatz	Legal Tech/Quantitative Forschung. Diese CSV-Datei ist eine der für statistische Analysen empfohlenen Varianten des Datensatzes. Sie enthält den Volltext aller Rechtsakte, sowie fast alle in diesem Codebook beschriebenen Metadaten. Die gegenüber den Einzelnormen fehlenden Metadaten betreffen vor allem Gliederungsdaten (z.B. Gliederungsüberschrift), die auf Rechtsakts-Ebene keinen Sinn ergeben. Wurde durch ein Zusammenfügen der Einzelnorm-Variante erstellt. Enthält nur Rechtsakte, für die mindestens eine Einzelnorm mit Normtext veröffentlicht wurde!
CSV_Rechtsakte_Metadaten	Legal Tech/Quantitative Forschung. Wie die andere CSV-Datei mit Rechtsakten, nur ohne die Normtexte. Sinnvoll für Analyst:innen, die sich nur für die Metadaten interessieren und Speicherplatz sparen wollen. Enthält nur Rechtsakte, für die mindestens eine Einzelnorm mit Normtext veröffentlicht wurde!
CSV_MetadatenXML	Legal Tech/Quantitative Forschung. Diese CSV-Datei enthält Metadaten für jeden auf der amtlichen Webseite nachgewiesenen Rechtsakt, unabhängig davon, ob mit oder ohne Normtext veröffentlicht. Die Zahl der Rechtsakte ist daher um etwa 1000 höher als bei den anderen CSV-Dateien, es sind aber keine Normtexte enthalten.

Variante	Zielgruppe und Beschreibung
XML_Datensatz	Legal Tech/Quantitative Forschung. Die XML-Rohdaten. Alle CSV-Dateien wurden aus diesen Rohdaten extrahiert. XML ist ein komplexes Format und daher nur für entsprechend verteilte Forscher:innen geeignet.
XML_Anlagen	Legal Tech/Quantitative Forschung. Manche XML-Dateien verweisen auf Anlagen, vorwiegend Bild-Dateien. Diese sind hier zusammengefasst.
PDF_Datensatz	Traditionelle juristische Forschung. Die PDF-Dokumente wie sie vom Bundesjustizministerium auf der amtlichen Webseite bereitgestellt werden, jedoch verbessert durch semantisch hochwertige Dateinamen, die sowohl die Abkürzung, als auch einen modifizierten Langtitel enthalten. Die Dateinamen sind so konzipiert, dass sie auch für traditionelle qualitative juristische Arbeit einen erheblichen Mehrwert bieten. Im Vergleich zu den CSV-Dateien enthalten die Dateinamen nur einen drastisch reduzierten Umfang an Metadaten, um Kompatibilitätsprobleme unter Windows zu vermeiden und die Lesbarkeit zu verbessern. Besonders geeignet für die Arbeit an Desktop PCs.
EPUB_Datensatz	Traditionelle juristische Forschung. Die PDF-Dokumente wie sie vom Bundesjustizministerium auf der amtlichen Webseite bereitgestellt werden, jedoch verbessert durch semantisch hochwertige Dateinamen, die sowohl die Abkürzung, als auch einen modifizierten Langtitel enthalten. Die Dateinamen sind so konzipiert, dass sie auch für traditionelle qualitative juristische Arbeit einen erheblichen Mehrwert bieten. Im Vergleich zu den CSV-Dateien enthalten die Dateinamen nur einen drastisch reduzierten Umfang an Metadaten, um Kompatibilitätsprobleme unter Windows zu vermeiden und die Lesbarkeit zu verbessern. Besonders geeignet für die Arbeit an mobilen Endgeräten, weil sich das Format der Bildschirmgröße anpassen kann.
TXT_Datensatz	Subsidiär. Diese Variante enthält die vollständigen aus den PDF-Dateien extrahierten Normtexte der Rechtsakte, aber nur einen drastisch 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, die nur wenig Speicherplatz oder eine langsame Internetverbindung zur Verfügung haben und für quantitative Forscher, die beim Einlesen der CSV-Dateien Probleme haben.

Variante	Zielgruppe und Beschreibung
Netzwerke	Experimentell. Die Gliederungshierarchie aller Rechtsakte wurde in eine Netzwerkstruktur übersetzt und ist in verschiedenen Formaten bereitgestellt (GraphML, Adjazenzmatrizen und Edge Lists). Aus dieser Netzwerkstruktur wurden zudem hierarchische Dendrogramme erstellt um einen visuellen Überblick zu bieten. Einzelnormen sind in den Netzwerkstrukturen aktuell noch nicht berücksichtigt. Diese Variante ist noch hoch-experimentell, sollte also nicht ohne genaue Prüfung für die eigene Forschung verwendet werden.
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 die einzelne Inhalte aus dem Codebook für andere Zwecke (z.B. eigene Publikationen) weiterverwenden möchten.

5 Variablen

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.
- Alle Variablen sind in der hier beschriebenen Form nur in der CSV-Datei enthalten. Die meisten davon sind jedoch aus gleichlautenden oder ähnlich lautenden Nodes in den XML-Daten vorhanden.

5.2 Erläuterungen der einzelnen Variablen

Variable	Typ	Erläuterung
doc_id	String	Ein einzigartiger Identifikator für jede Einzelnorm bzw. jeden Rechtsakt. Für Rechtsakte entspricht diese Variable dem Namen der extrahierten XML-Datei. Einzelnormen enthalten den Namen der XML-Datei und jeweils eine fortlaufende Zahl. Bei Einzelnormen nicht notwendigerweise stabil zwischen den Versionen des C-DBR. Bei Rechtsakten vermutlich schon.
dateiname	String	(Nur Einzelnormen-Variante). Der Dateiname der XML-Datei aus dem die Einzelnormen extrahiert wurden.
text	String	Der vollständige Normtext der Einzelnorm oder des Rechtsaktes, so wie er in den XML-Dateien dokumentiert ist. Nur die Varianten »Einzelnormen« und »Rechtsakte« enthalten Textdaten. Hierzu schreibt das Ministerium in den Hinweisen: »Einzelne Vorschriften sind nur mit der Überschrift aufgenommen. In einigen Fällen hat dies technische Gründe, in anderen Fällen ist dies dadurch bedingt, dass die Vorschrift nur mit ihren bibliographischen Angaben in der Bereinigten Sammlung Bundesgesetzblatt Teil III enthalten ist (§ 3 Abs. 2 BRSG).«
amtabk	String	Die amtliche Abkürzung des Rechtsaktes. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!

(continued)

Variable	Typ	Erläuterung
jurabk	String	Unter »Hinweise« schreibt das Ministerium: »Zu einigen Gesetzen und Verordnungen existieren keine amtlichen Abkürzungen. In diesen Fällen sind die Vorschriften in den alphabetischen Listen von ›Gesetze im Internet‹ anhand der von der Dokumentationsstelle im BfJ gebildeten und in der Bundesrechtsdatenbank verwendeten Abkürzungen eingeordnet. Diese nichtamtlichen Abkürzungen können von Abkürzungen, die andere Anbieter verwenden, abweichen.« Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
langue	String	Die Langform des Namens (Langüberschrift) eines Rechtsaktes. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
kurzue	String	Die Kurzform des Namens (Kurzüberschrift) eines Rechtsaktes. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
gliederungskennzahl	String	(Nur Einzelnormen-Variante). Die Kennzahl der jeweiligen Gliederungsebene. Das zugrundeliegende System ist vom Ministerium nicht dokumentiert. Beispielsweise »010050030«. Vermutlich ist jede Gliederungsebene mit drei Zahlen definiert: die ersten beiden Zahlen bilden die Ordinalzahl innerhalb der Ebene (ggf. mit vorangestellter Null falls kleiner 10), die dritte Zahl ist meistens eine Null und wird ggf. erhöht falls die Gliederungsbezeichnung mit Buchstaben ausdifferenziert wurde (nachträgliche Einfügung). Das Beispiel enthält also drei Ebenen: erste Überschrift der 1. Ebene, fünfte Überschrift der 2. Ebene, dritte Überschrift der 3. Ebene.
gliederungsbez	String	(Nur Einzelnormen-Variante). Die Bezeichnung der Gliederungsebene. Beispielsweise »Titel 3«.
gliederungstitel	String	(Nur Einzelnormen-Variante). Der Titel der Gliederungsebene. Beispielsweise »Rechtsfolgen der Verjährung«.
enbez	String	(Nur Einzelnormen-Variante). Die Bezeichnung der Einzelnorm. Beispielsweise »§ 214«.

(continued)

Variable	Typ	Erläuterung
bezkette	String	Die volle Hierarchie der Bezeichnungen der Gliederungsebenen in Form einer Kette. Beispiel: »Buch 2 Abschnitt 8 Titel 5 Untertitel 2 Kapitel 5 Unterkapitel 4«.
titelkette	String	Die volle Hierarchie der Titel der Gliederungsebenen in Form einer Kette. Beispiel: »Recht der Schuldverhältnisse Einzelne Schuldverhältnisse Mietvertrag, Pachtvertrag Mietverhältnisse über Wohnraum Beendigung des Mietverhältnisses Werkwohnungen«.
ausfertigung_datum	Datum (ISO)	Das Datum an dem der Rechtsakt ausgefertigt wurde. Das Format ist YYYY-MM-DD (Langform nach ISO-8601). Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt! Daher sind beispielsweise für das BGB alle Normen mit dem Ausfertigungsjahr 1896 versehen, auch wenn die Einzelnorm später erlassen wurde.
ausfertigung_jahr	Natürliche Zahl	Das Jahr in dem der Rechtsakt ausgefertigt wurde. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt. Das Format ist eine vierstellige Jahreszahl (YYYY). Wurde durch den Autor aus dem Ausfertigungsdatum berechnet.
aenderung_datum	Datum (ISO)	Das Datum der letzten Änderung des Rechtsakts. Das Format ist YYYY-MM-DD (Langform nach ISO-8601). Aus der Variable »stand« mittels <i>regular expressions</i> extrahiert.
Aufhebung_ verkuendung_datum	Datum (ISO)	Das Datum an dem ein etwaiger aufhebender Rechtsakt verkündet wurde. Das Format ist YYYY-MM-DD (Langform nach ISO-8601). Aus der Variable »aufh« mittels <i>regular expressions</i> extrahiert.
aufhebung_wirkung_ datum	Datum (ISO)	Das Datum an dem ein etwaiger aufhebender Rechtsakt wirksam wird. Das Format ist YYYY-MM-DD (Langform nach ISO-8601). Aus der Variable »aufh« mittels <i>regular expressions</i> extrahiert.
neufassung_datum	String	Das Datum an dem der Rechtsakt zuletzt neu gefasst wurde. Das Format ist YYYY-MM-DD (Langform nach ISO-8601). Aus der Variable »neuf« mittels <i>regular expressions</i> extrahiert.

(continued)

Variable	Typ	Erläuterung
builddate_original	String	Datum und Uhrzeit an dem die XML-Repräsentation der Norm konstruiert wurde, eine Serie von Zahlen ohne Interpunktion. Das genaue Format ist nicht dokumentiert, es ist aber sehr wahrscheinlich so aufgebaut: vierstellige Jahreszahl, zweistellige Monatszahl, zweistellige Tageszahl, zweistellige Stundenzahl, zweistellige Minutenzahl und eine zweistellige Sekundenzahl.
builddate_iso	Zeitstempel (ISO)	Eine Interpretation der builddate-Variable im ISO 8601-Format (z.B. 2016-09-12T18:12:16Z). Das genaue Original-Format ist nicht dokumentiert, die Variable wurde aber unter folgenden Annahmen extrahiert: vierstellige Jahreszahl, zweistellige Monatszahl, zweistellige Tageszahl, zweistellige Stundenzahl, zweistellige Minutenzahl und eine zweistellige Sekundenzahl.
fundstellentyp	String	Ob es sich um eine amtliche Fundstelle handelt oder nicht. Mögliche Werte sind »amtlich« oder »nichtamtlich«. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
periodikum	String	Die Abkürzung des Periodikums in dem die amtliche Fassung des Rechtsaktes erschienen ist, beispielsweise »BGBI I« (Bundesgesetzblatt I). Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
zitstelle	String	Die genaue Fundstelle im jeweiligen Periodikum. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
stand	String	Informationen zum aktuellen Stand des Rechtsaktes, als Fließtext. Enthält insbesondere Informationen zur letzten Änderung und dem letzten Änderungsrechtsakt. Jeweils durch einen vertikalen Strich » « getrennt, falls mehr als eine Bemerkung vorhanden ist. Falls nicht vorhanden ist der Wert »NA«. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!

(continued)

Variable	Typ	Erläuterung
aufh	String	Informationen zur etwaigen Aufhebung des Rechtsaktes, als Fließtext. Jeweils durch einen vertikalen Strich » « getrennt, falls mehr als eine Bemerkung vorhanden ist. Falls nicht vorhanden ist der Wert »NA«. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
neuf	String	Informationen zur letzten Neufassung des Rechtsaktes, als Fließtext. Jeweils durch einen vertikalen Strich » « getrennt, falls mehr als eine Bemerkung vorhanden ist. Falls nicht vorhanden ist der Wert »NA«. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
hinweis	String	Hinweise zur dokumentarischen Bearbeitung des Rechtsaktes, als Fließtext. Jeweils durch einen vertikalen Strich » « getrennt, falls mehr als eine Bemerkung vorhanden ist. Falls nicht vorhanden ist der Wert »NA«. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
sonst	String	Sonstige Informationen zum Stand des Rechtsaktes, als Fließtext. Jeweils durch einen vertikalen Strich » « getrennt, falls mehr als eine Bemerkung vorhanden ist. Falls nicht vorhanden ist der Wert »NA«. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
check_*	String	Ob die Angabe der jeweiligen Stand-Variable geprüft wurde. Es ist unklar, welche Prüfung hier vom Ministerium vorgenommen wurde. Mögliche Werte sind »ja« oder »NA«. Für Einzelnormen bezieht sich die Angabe auf den gesamten Rechtsakt!
tokens	Natürliche Zahl	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.

(continued)

Variable	Typ	Erläuterung
typen	Natürliche Zahl	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	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.
version	Datum	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.
doi_concept	String	Der Digital Object Identifier (DOI) des Gesamtkonzeptes des Datensatzes. Dieser ist langzeitstabil (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.

(continued)

Variable	Typ	Erläuterung
doi_version	String	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 des Datensatzes. In diesem Fall immer »Creative Commons Zero 1.0 Universal«.

5.3 Konkordanztabelle: XML-Struktur und CSV-Variablen

CSV-Variablen	XPath	Attribut
text	/norm/textdaten/text/Content	-
builddate_original	/norm	builddate
fundstellentyp	/norm/metadaten/fundstelle	typ
periodikum	/norm/metadaten/fundstelle/periodikum	-
zitstelle	/norm/metadaten/fundstelle/zitstelle	-
stand	/norm/metadaten/standangabe/standtyp	-
	/norm/metadaten/standangabe/standkommentar	-
aufh	/norm/metadaten/standangabe/standtyp	-
	/norm/metadaten/standangabe/standkommentar	-
neuf	/norm/metadaten/standangabe/standtyp	-
	/norm/metadaten/standangabe/standkommentar	-
hinweis	/norm/metadaten/standangabe/standtyp	-
	/norm/metadaten/standangabe/standkommentar	-
sonst	/norm/metadaten/standangabe/standtyp	-
	/norm/metadaten/standangabe/standkommentar	-
check_*	/norm/metadaten/standangabe	checked
amtabk	/norm/metadaten/amtabk	-
jurabk	/norm/metadaten/jurabk	-
langue	/norm/metadaten/langue	-
kurzue	/norm/metadaten/kurzue	-
gliederungskennzahl	/norm/metadaten/gliederungseinheit/gliederungskennzahl	-
gliederungsbez	/norm/metadaten/gliederungseinheit/gliederungsbez	-
gliederungstitel	/norm/metadaten/gliederungseinheit/gliederungstitel	-
enbez	/norm/metadaten/enbez	-
ausfertigung_datum	/norm/metadaten/ausfertigung-datum	-

6 Linguistische Kennzahlen

6.1 Erläuterung der Kennzahlen

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

Variable	Definition
Zeichen	Zeichen entsprechen grob den <i>Graphemen</i> , 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 »gewerblich« mehrmals in einer Norm 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. Für Rechtsnormen ist diese Zählweise vermutlich nicht robust genug, interpretieren Sie die Ergebnisse mit großer Vorsicht!

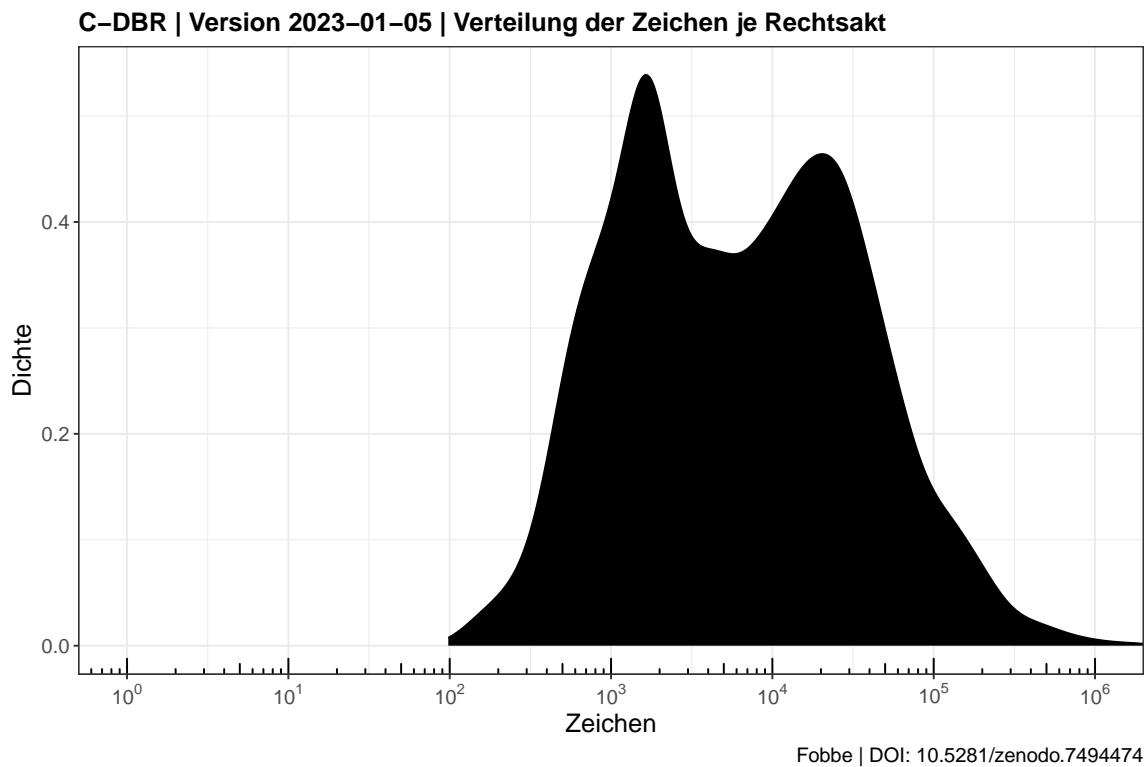
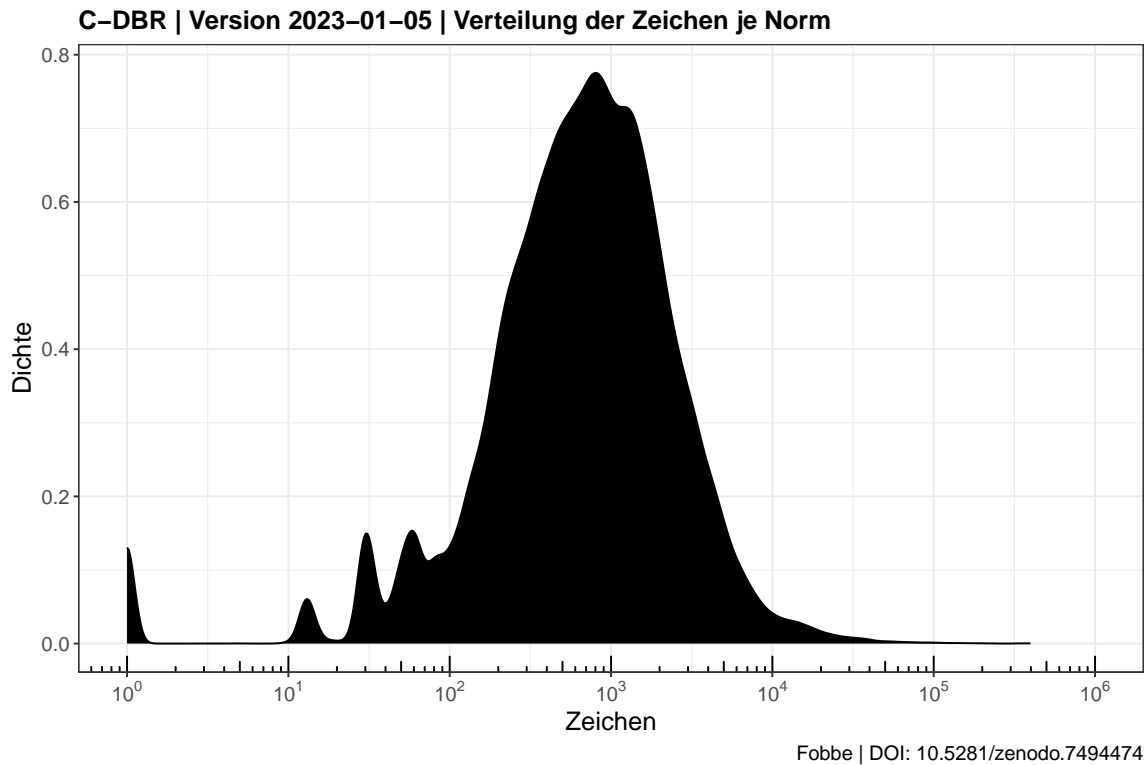
6.2 Kennzahlen: Einzelnormen

Variable	Gesamt	Min	1. Quartil	Median	Mittel	3. Quartil	Max
Zeichen	146,474,799	1	296	694	1,431.17	1,499	398,079
Tokens	22,911,664	1	47	110	223.86	235	56,584
Typen	424,769	1	36	67	93.67	115	23,559
Sätze	763,273	1	2	4	7.46	8	2,666

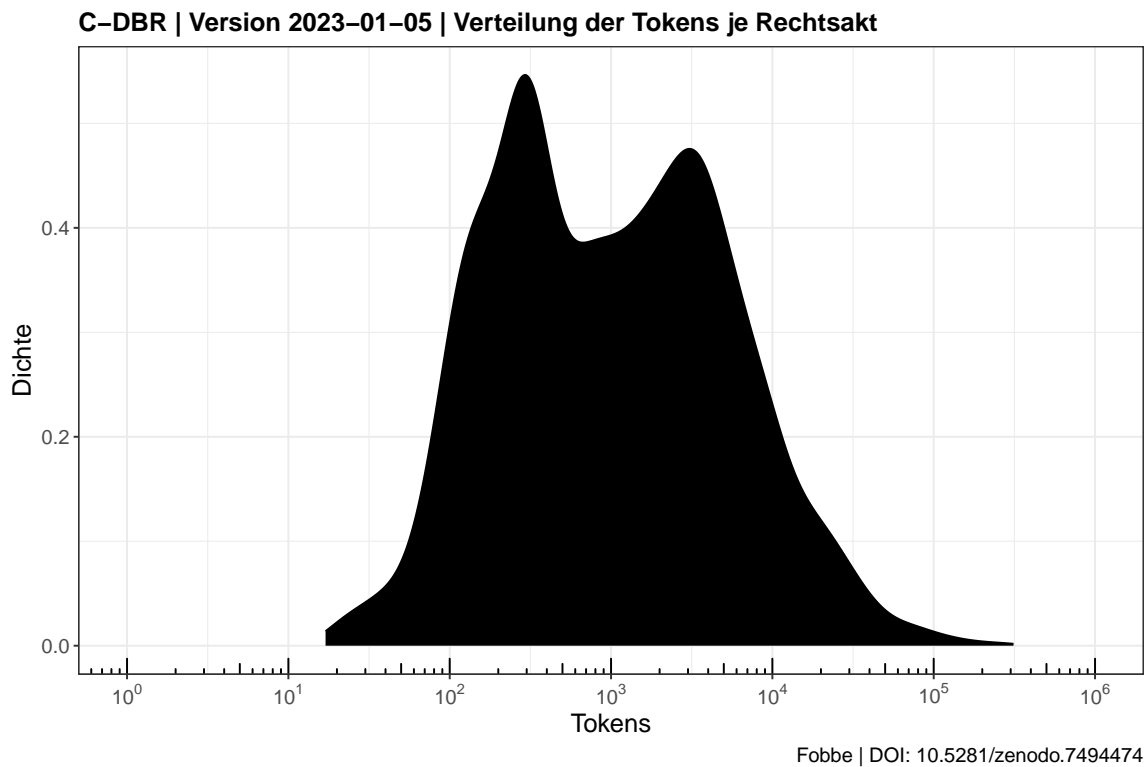
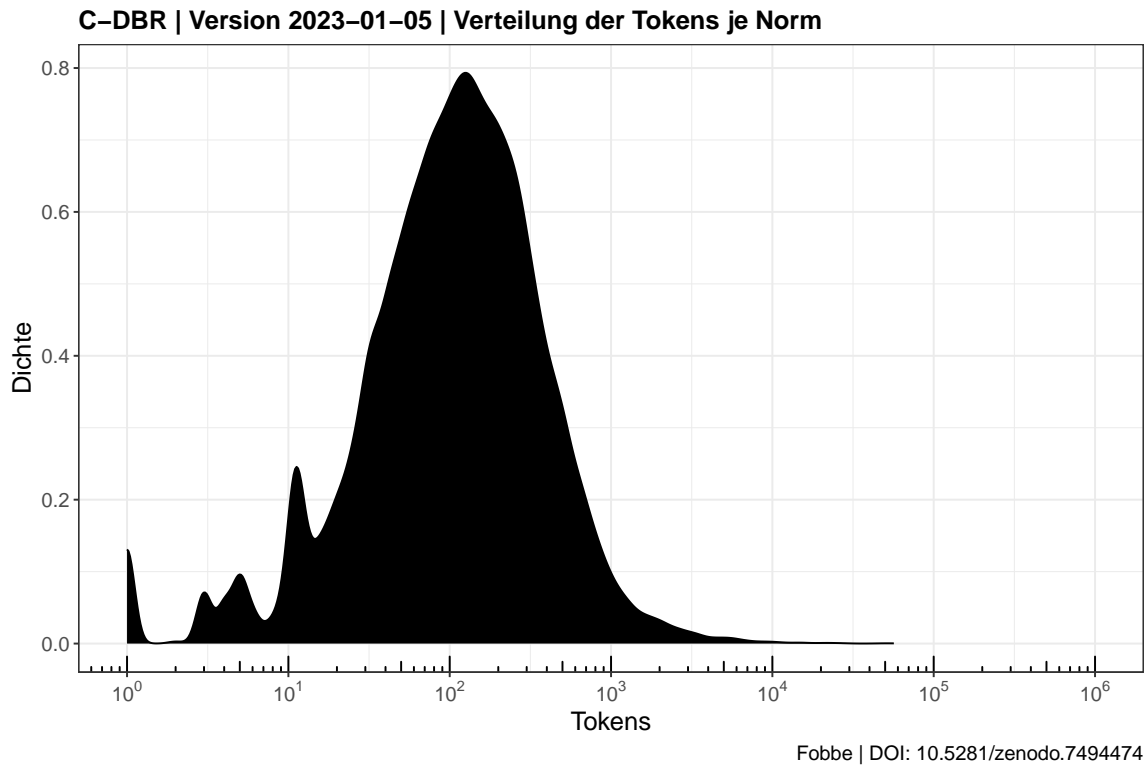
6.3 Kennzahlen: Rechtsakte

Variable	Gesamt	Min	1. Quartil	Median	Mittel	3. Quartil	Max
Zeichen	146,571,447	98	1,570	5,847	25,723.31	23,117.00	2,121,463
Tokens	22,911,664	17	266	934	4,021.00	3,500.25	311,979
Typen	424,769	15	130	307	650.64	820.00	23,606
Sätze	751,831	1	13	36	131.95	117.00	14,306

6.4 Verteilung Zeichen

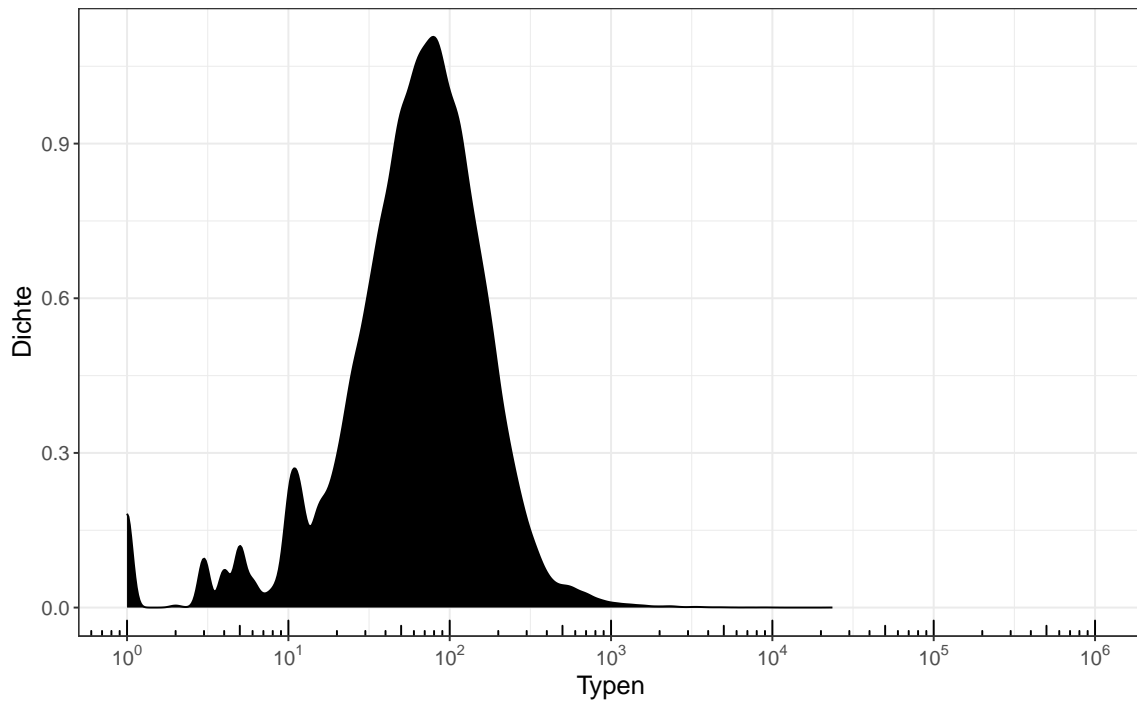


6.5 Verteilung Tokens



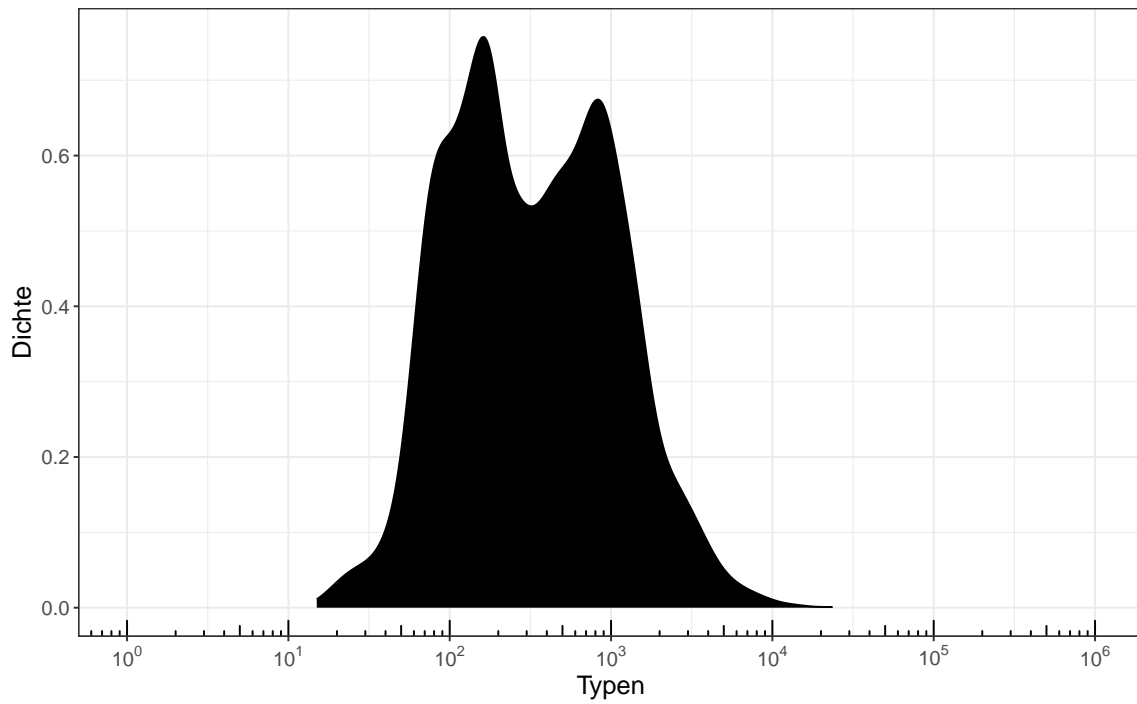
6.6 Verteilung Typen

C-DBR | Version 2023-01-05 | Verteilung der Typen je Norm



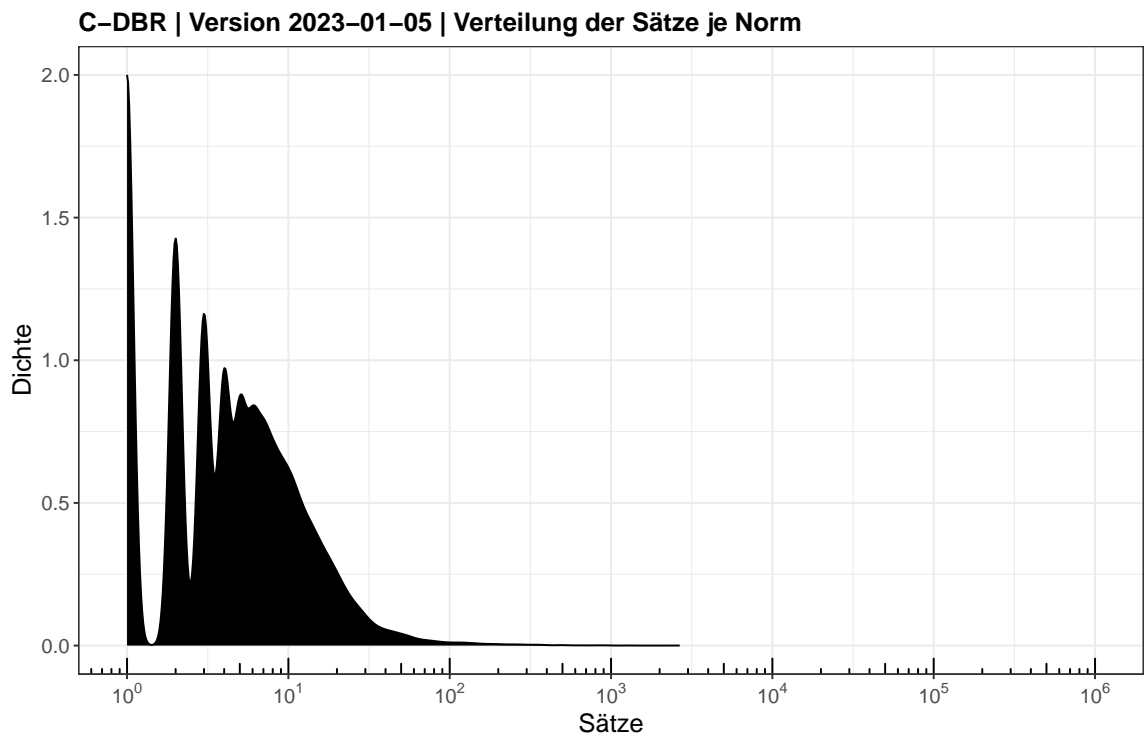
Fobbe | DOI: 10.5281/zenodo.7494474

C-DBR | Version 2023-01-05 | Verteilung der Typen je Rechtsakt

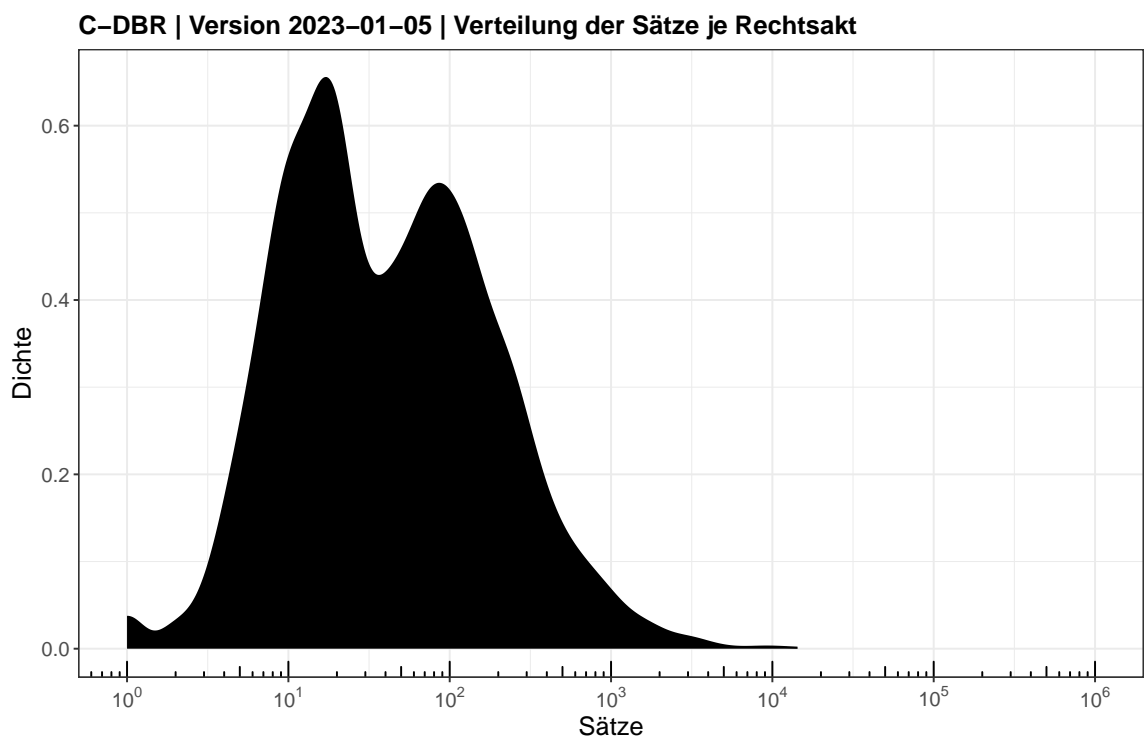


Fobbe | DOI: 10.5281/zenodo.7494474

6.7 Verteilung Sätze



Fobbe | DOI: 10.5281/zenodo.7494474

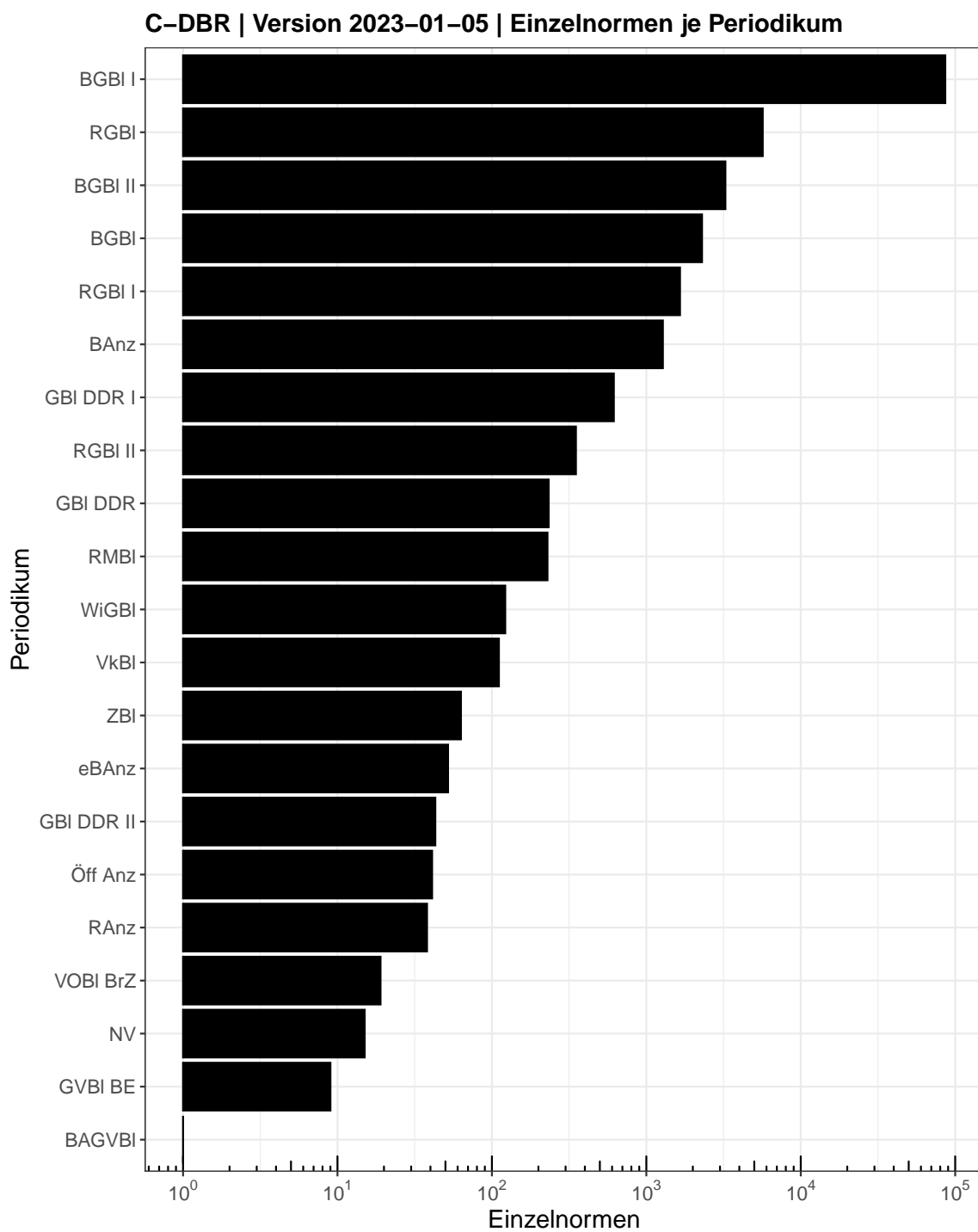


Fobbe | DOI: 10.5281/zenodo.7494474

7 Inhalt

7.1 Nach Periodikum

7.1.1 Einzelnormen

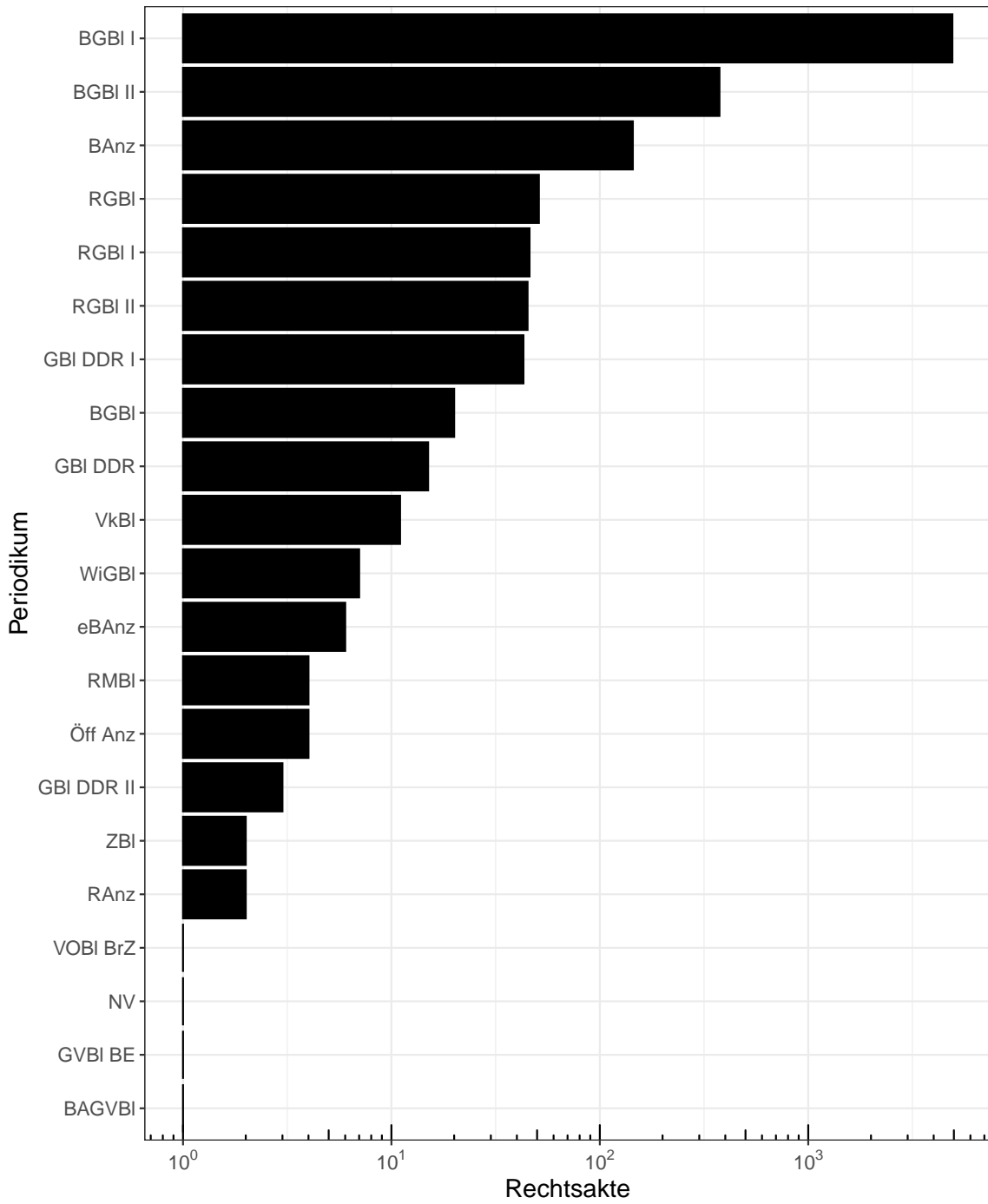


Fobbe | DOI: 10.5281/zenodo.7494474

Periodikum	Einzelnormen	% Gesamt	% Kumulativ
BAGVBl	1	0.00	0.00
BAnz	1280	1.25	1.25
BGBI	2294	2.24	3.49
BGBI I	86255	84.28	87.77
BGBI II	3251	3.18	90.95
GBl DDR	233	0.23	91.18
GBl DDR I	616	0.60	91.78
GBl DDR II	43	0.04	91.82
GVBl BE	9	0.01	91.83
NV	15	0.01	91.84
RAnz	38	0.04	91.88
RGBI	5672	5.54	97.42
RGBI I	1652	1.61	99.04
RGBI II	350	0.34	99.38
RMBI	229	0.22	99.60
VOBl BrZ	19	0.02	99.62
VkBl	111	0.11	99.73
WiGBI	122	0.12	99.85
ZBl	63	0.06	99.91
eBAnz	52	0.05	99.96
Öff Anz	41	0.04	100.00
Total	102346	100.00	100.00

7.1.2 Rechtsakte mit veröffentlichtem Normtext

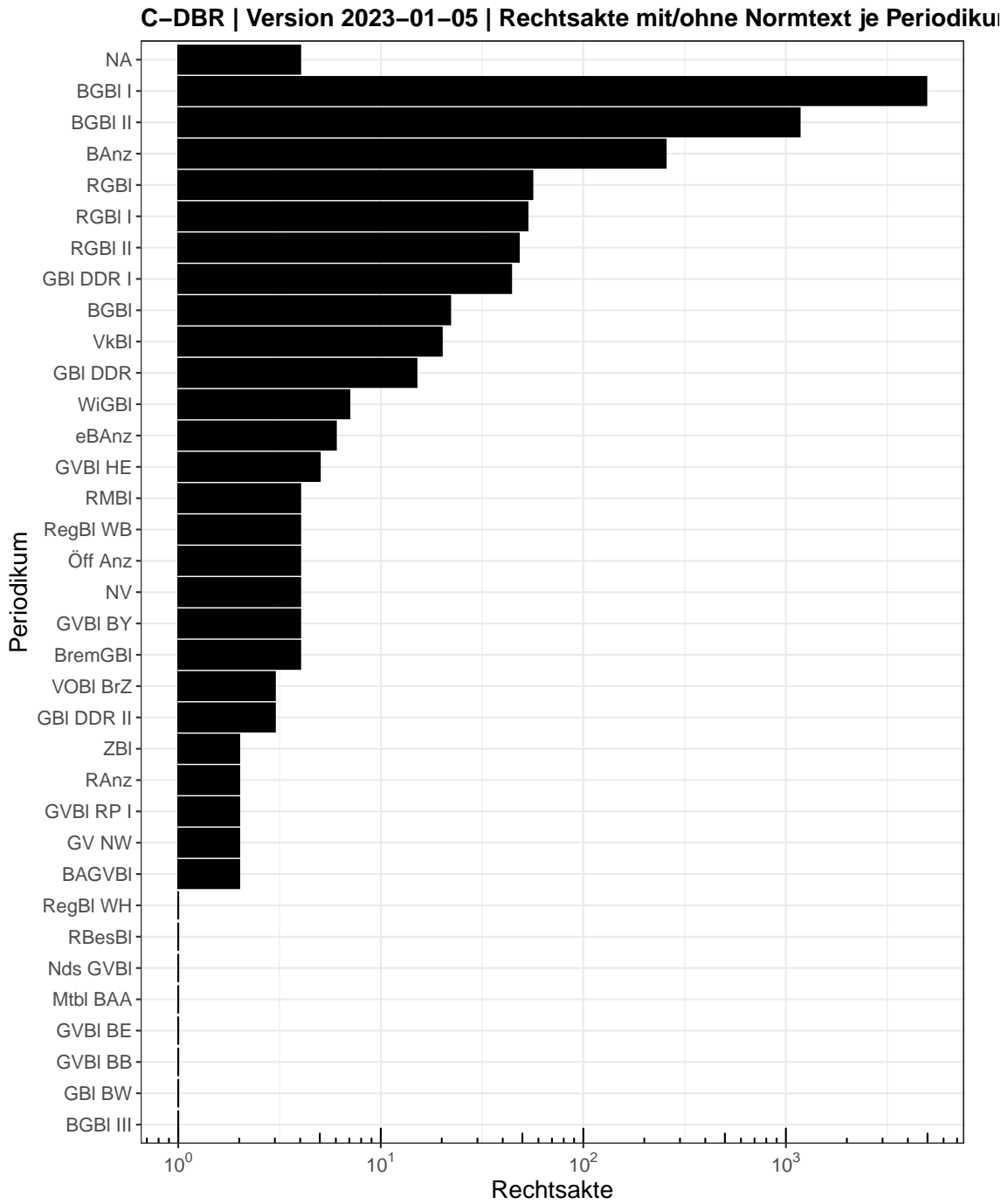
C-DBR | Version 2023-01-05 | Rechtsakte mit Normtext je Periodikum



Fobbe | DOI: 10.5281/zenodo.7494474

Periodikum	Rechtsakte	% Gesamt	% Kumulativ
BAGVBl	1	0.02	0.02
BAnz	144	2.53	2.54
BGBI	20	0.35	2.90
BGBI I	4916	86.28	89.17
BGBI II	375	6.58	95.75
GBl DDR	15	0.26	96.02
GBl DDR I	43	0.75	96.77
GBl DDR II	3	0.05	96.82
GVBl BE	1	0.02	96.84
NV	1	0.02	96.86
RAnz	2	0.04	96.89
RGBI	51	0.90	97.79
RGBI I	46	0.81	98.60
RGBI II	45	0.79	99.39
RMBI	4	0.07	99.46
VOBl BrZ	1	0.02	99.47
VkBl	11	0.19	99.67
WiGBI	7	0.12	99.79
ZBl	2	0.04	99.82
eBAnz	6	0.11	99.93
Öff Anz	4	0.07	100.00
Total	5698	100.00	100.00

7.1.3 Alle Rechtsakte (mit und ohne Normtext)



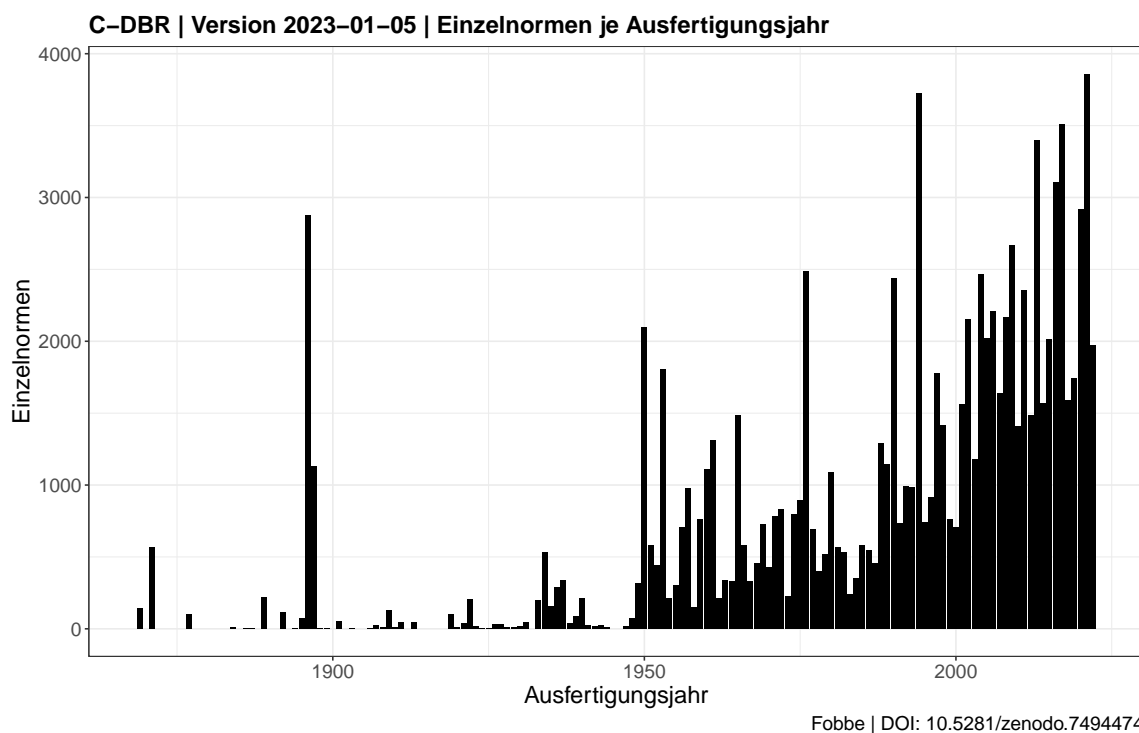
Periodikum	Rechtsakte	% Gesamt	% Kumulativ
NA	4	0.06	0.06
BAGVBl	2	0.03	0.09
BAnz	255	3.81	3.90
BGBI	22	0.33	4.23
BGBI I	4935	73.80	78.03
BGBI II	1169	17.48	95.51
BGBI III	1	0.01	95.53
BremGBI	4	0.06	95.59
GBI BW	1	0.01	95.60
GBI DDR	15	0.22	95.83
GBI DDR I	44	0.66	96.49
GBI DDR II	3	0.04	96.53
GV NW	2	0.03	96.56
GVB I BB	1	0.01	96.58
GVB I BE	1	0.01	96.59
GVB I BY	4	0.06	96.65
GVB I HE	5	0.07	96.72
GVB I RP I	2	0.03	96.75
Mtbl BAA	1	0.01	96.77
NV	4	0.06	96.83
Nds GVB I	1	0.01	96.84
RAnz	2	0.03	96.87
RBesBl	1	0.01	96.89
RGBI	56	0.84	97.73
RGBI I	53	0.79	98.52
RGBI II	48	0.72	99.24
RMBI	4	0.06	99.30
RegBl WB	4	0.06	99.36
RegBl WH	1	0.01	99.37
VOBl BrZ	3	0.04	99.42
VkBl	20	0.30	99.72
WiGBI	7	0.10	99.82

(continued)

Periodikum	Rechtsakte	% Gesamt	% Kumulativ
ZBl	2	0.03	99.85
eBAnz	6	0.09	99.94
Öff Anz	4	0.06	100.00
Total	6687	100.00	100.00

7.2 Nach Ausfertigungsjahr

7.2.1 Einzelnormen



Jahr	Einzelnormen	% Gesamt	% Kumulativ
1869	144	0.14	0.14
1871	564	0.55	0.69
1877	99	0.10	0.79
1884	10	0.01	0.80
1886	3	0.00	0.80
1887	4	0.00	0.81
1889	218	0.21	1.02
1892	117	0.11	1.13
1894	4	0.00	1.14
1895	74	0.07	1.21
1896	2877	2.81	4.02
1897	1131	1.11	5.12
1898	3	0.00	5.13
1899	2	0.00	5.13
1901	51	0.05	5.18

(continued)

Jahr	Einzelnormen	% Gesamt	% Kumulativ
1903	1	0.00	5.18
1906	1	0.00	5.18
1907	26	0.03	5.21
1908	8	0.01	5.21
1909	128	0.13	5.34
1910	11	0.01	5.35
1911	47	0.05	5.40
1913	46	0.04	5.44
1919	99	0.10	5.54
1920	7	0.01	5.54
1921	40	0.04	5.58
1922	206	0.20	5.79
1923	15	0.01	5.80
1924	2	0.00	5.80
1925	2	0.00	5.80
1926	31	0.03	5.83
1927	32	0.03	5.87
1928	8	0.01	5.87
1929	8	0.01	5.88
1930	14	0.01	5.89
1931	46	0.04	5.94
1933	196	0.19	6.13
1934	530	0.52	6.65
1935	156	0.15	6.80
1936	287	0.28	7.08
1937	334	0.33	7.41
1938	41	0.04	7.45
1939	90	0.09	7.54
1940	210	0.21	7.74
1941	25	0.02	7.77
1942	20	0.02	7.79

(continued)

Jahr	Einzelnormen	% Gesamt	% Kumulativ
1943	24	0.02	7.81
1944	12	0.01	7.82
1947	19	0.02	7.84
1948	73	0.07	7.91
1949	314	0.31	8.22
1950	2095	2.05	10.26
1951	583	0.57	10.83
1952	445	0.43	11.27
1953	1806	1.76	13.03
1954	209	0.20	13.24
1955	303	0.30	13.53
1956	705	0.69	14.22
1957	975	0.95	15.17
1958	150	0.15	15.32
1959	765	0.75	16.07
1960	1107	1.08	17.15
1961	1312	1.28	18.43
1962	210	0.21	18.64
1963	336	0.33	18.97
1964	332	0.32	19.29
1965	1486	1.45	20.74
1966	580	0.57	21.31
1967	330	0.32	21.63
1968	456	0.45	22.08
1969	725	0.71	22.79
1970	425	0.42	23.20
1971	780	0.76	23.96
1972	831	0.81	24.77
1973	228	0.22	25.00
1974	796	0.78	25.78
1975	897	0.88	26.65

(continued)

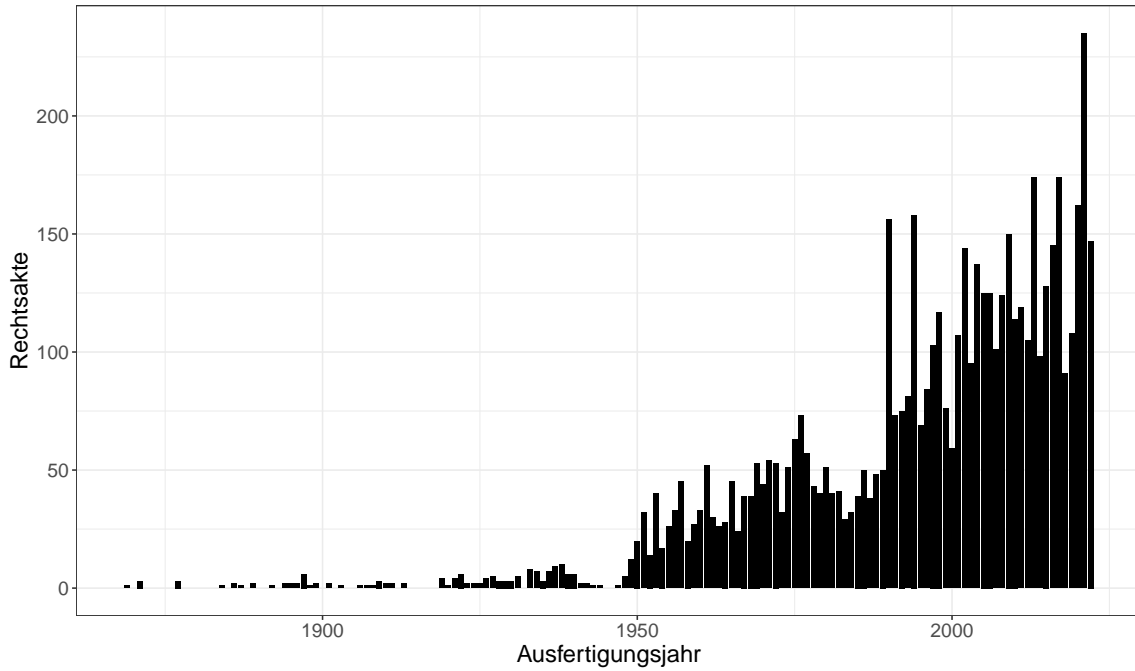
Jahr	Einzelnormen	% Gesamt	% Kumulativ
1976	2486	2.43	29.08
1977	690	0.67	29.75
1978	399	0.39	30.14
1979	521	0.51	30.65
1980	1088	1.06	31.72
1981	569	0.56	32.27
1982	535	0.52	32.80
1983	243	0.24	33.03
1984	351	0.34	33.38
1985	580	0.57	33.94
1986	548	0.54	34.48
1987	455	0.44	34.92
1988	1288	1.26	36.18
1989	1143	1.12	37.30
1990	2436	2.38	39.68
1991	734	0.72	40.40
1992	993	0.97	41.37
1993	981	0.96	42.32
1994	3726	3.64	45.96
1995	738	0.72	46.69
1996	917	0.90	47.58
1997	1780	1.74	49.32
1998	1416	1.38	50.70
1999	765	0.75	51.45
2000	706	0.69	52.14
2001	1559	1.52	53.67
2002	2150	2.10	55.77
2003	1177	1.15	56.92
2004	2467	2.41	59.33
2005	2021	1.97	61.30
2006	2210	2.16	63.46

(continued)

Jahr	Einzelnormen	% Gesamt	% Kumulativ
2007	1636	1.60	65.06
2008	2169	2.12	67.18
2009	2666	2.60	69.78
2010	1407	1.37	71.16
2011	2354	2.30	73.46
2012	1482	1.45	74.91
2013	3396	3.32	78.22
2014	1567	1.53	79.75
2015	2017	1.97	81.73
2016	3107	3.04	84.76
2017	3512	3.43	88.19
2018	1590	1.55	89.75
2019	1744	1.70	91.45
2020	2919	2.85	94.30
2021	3859	3.77	98.07
2022	1972	1.93	100.00
Total	102346	100.00	100.00

7.2.2 Rechtsakte mit veröffentlichtem Normtext

C-DBR | Version 2023-01-05 | Rechtsakte mit Normtext je Ausfertigungsjahr



Fobbe | DOI: 10.5281/zenodo.7494474

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1869	1	0.02	0.02
1871	3	0.05	0.07
1877	3	0.05	0.12
1884	1	0.02	0.14
1886	2	0.04	0.18
1887	1	0.02	0.19
1889	2	0.04	0.23
1892	1	0.02	0.25
1894	2	0.04	0.28
1895	2	0.04	0.32
1896	2	0.04	0.35
1897	6	0.11	0.46
1898	1	0.02	0.47
1899	2	0.04	0.51
1901	2	0.04	0.54
1903	1	0.02	0.56

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1906	1	0.02	0.58
1907	1	0.02	0.60
1908	1	0.02	0.61
1909	3	0.05	0.67
1910	2	0.04	0.70
1911	2	0.04	0.74
1913	2	0.04	0.77
1919	4	0.07	0.84
1920	1	0.02	0.86
1921	4	0.07	0.93
1922	6	0.11	1.04
1923	2	0.04	1.07
1924	2	0.04	1.11
1925	2	0.04	1.14
1926	4	0.07	1.21
1927	5	0.09	1.30
1928	3	0.05	1.35
1929	3	0.05	1.40
1930	3	0.05	1.46
1931	5	0.09	1.54
1933	8	0.14	1.68
1934	7	0.12	1.81
1935	3	0.05	1.86
1936	7	0.12	1.98
1937	9	0.16	2.14
1938	10	0.18	2.32
1939	6	0.11	2.42
1940	6	0.11	2.53
1941	2	0.04	2.56
1942	2	0.04	2.60
1943	1	0.02	2.61

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1944	1	0.02	2.63
1947	1	0.02	2.65
1948	5	0.09	2.74
1949	12	0.21	2.95
1950	20	0.35	3.30
1951	32	0.56	3.86
1952	14	0.25	4.11
1953	40	0.70	4.81
1954	17	0.30	5.11
1955	26	0.46	5.56
1956	33	0.58	6.14
1957	45	0.79	6.93
1958	20	0.35	7.28
1959	27	0.47	7.76
1960	33	0.58	8.34
1961	52	0.91	9.25
1962	30	0.53	9.78
1963	26	0.46	10.23
1964	28	0.49	10.72
1965	45	0.79	11.51
1966	24	0.42	11.93
1967	39	0.68	12.62
1968	39	0.68	13.30
1969	53	0.93	14.23
1970	44	0.77	15.01
1971	54	0.95	15.95
1972	53	0.93	16.88
1973	32	0.56	17.44
1974	51	0.90	18.34
1975	63	1.11	19.45
1976	73	1.28	20.73

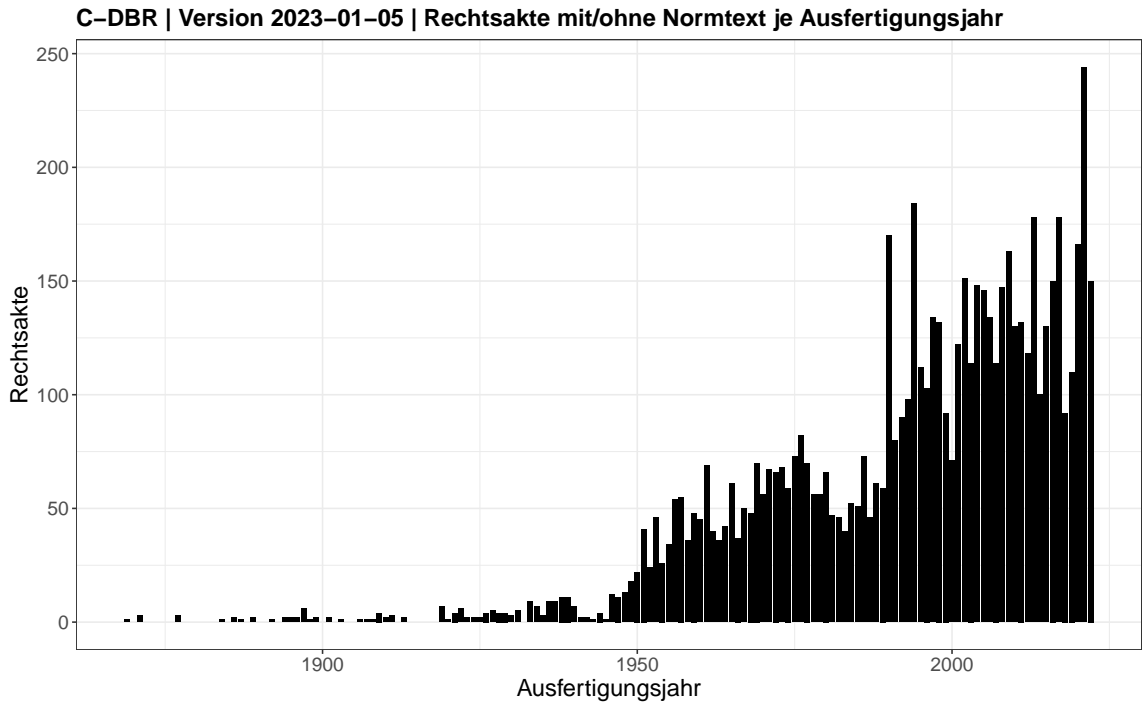
(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1977	57	1.00	21.73
1978	43	0.75	22.48
1979	40	0.70	23.18
1980	51	0.90	24.08
1981	40	0.70	24.78
1982	41	0.72	25.50
1983	29	0.51	26.01
1984	32	0.56	26.57
1985	39	0.68	27.26
1986	50	0.88	28.13
1987	38	0.67	28.80
1988	48	0.84	29.64
1989	50	0.88	30.52
1990	156	2.74	33.26
1991	73	1.28	34.54
1992	75	1.32	35.85
1993	81	1.42	37.28
1994	158	2.77	40.05
1995	69	1.21	41.26
1996	84	1.47	42.73
1997	103	1.81	44.54
1998	117	2.05	46.60
1999	76	1.33	47.93
2000	59	1.04	48.96
2001	107	1.88	50.84
2002	144	2.53	53.37
2003	95	1.67	55.04
2004	137	2.40	57.44
2005	125	2.19	59.63
2006	125	2.19	61.83
2007	101	1.77	63.60

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
2008	124	2.18	65.78
2009	150	2.63	68.41
2010	114	2.00	70.41
2011	119	2.09	72.50
2012	105	1.84	74.34
2013	174	3.05	77.40
2014	98	1.72	79.12
2015	128	2.25	81.36
2016	145	2.54	83.91
2017	174	3.05	86.96
2018	91	1.60	88.56
2019	108	1.90	90.45
2020	162	2.84	93.30
2021	235	4.12	97.42
2022	147	2.58	100.00
Total	5698	100.00	100.00

7.2.3 Alle Rechtsakte (mit und ohne Normtext)



Jahr	Rechtsakte	% Gesamt	% Kumulativ
NA	1	0.01	0.01
1869	1	0.01	0.03
1871	3	0.04	0.07
1877	3	0.04	0.12
1884	1	0.01	0.13
1886	2	0.03	0.16
1887	1	0.01	0.18
1889	2	0.03	0.21
1892	1	0.01	0.22
1894	2	0.03	0.25
1895	2	0.03	0.28
1896	2	0.03	0.31
1897	6	0.09	0.40
1898	1	0.01	0.42
1899	2	0.03	0.45
1901	2	0.03	0.48

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1903	1	0.01	0.49
1906	1	0.01	0.51
1907	1	0.01	0.52
1908	1	0.01	0.54
1909	4	0.06	0.60
1910	2	0.03	0.63
1911	3	0.04	0.67
1913	2	0.03	0.70
1919	7	0.10	0.81
1920	1	0.01	0.82
1921	4	0.06	0.88
1922	6	0.09	0.97
1923	2	0.03	1.00
1924	2	0.03	1.03
1925	2	0.03	1.06
1926	4	0.06	1.12
1927	5	0.07	1.20
1928	4	0.06	1.26
1929	4	0.06	1.32
1930	3	0.04	1.36
1931	5	0.07	1.44
1933	9	0.13	1.57
1934	7	0.10	1.67
1935	3	0.04	1.72
1936	9	0.13	1.85
1937	9	0.13	1.99
1938	11	0.16	2.15
1939	11	0.16	2.32
1940	7	0.10	2.42
1941	2	0.03	2.45
1942	2	0.03	2.48

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1943	1	0.01	2.50
1944	4	0.06	2.56
1945	1	0.01	2.57
1946	12	0.18	2.75
1947	11	0.16	2.92
1948	13	0.19	3.11
1949	18	0.27	3.38
1950	22	0.33	3.71
1951	41	0.61	4.32
1952	24	0.36	4.68
1953	46	0.69	5.37
1954	26	0.39	5.76
1955	34	0.51	6.27
1956	54	0.81	7.07
1957	55	0.82	7.90
1958	36	0.54	8.43
1959	48	0.72	9.15
1960	45	0.67	9.83
1961	69	1.03	10.86
1962	40	0.60	11.46
1963	36	0.54	11.99
1964	42	0.63	12.62
1965	61	0.91	13.53
1966	37	0.55	14.09
1967	50	0.75	14.83
1968	48	0.72	15.55
1969	70	1.05	16.60
1970	56	0.84	17.44
1971	67	1.00	18.44
1972	66	0.99	19.43
1973	68	1.02	20.44

(continued)

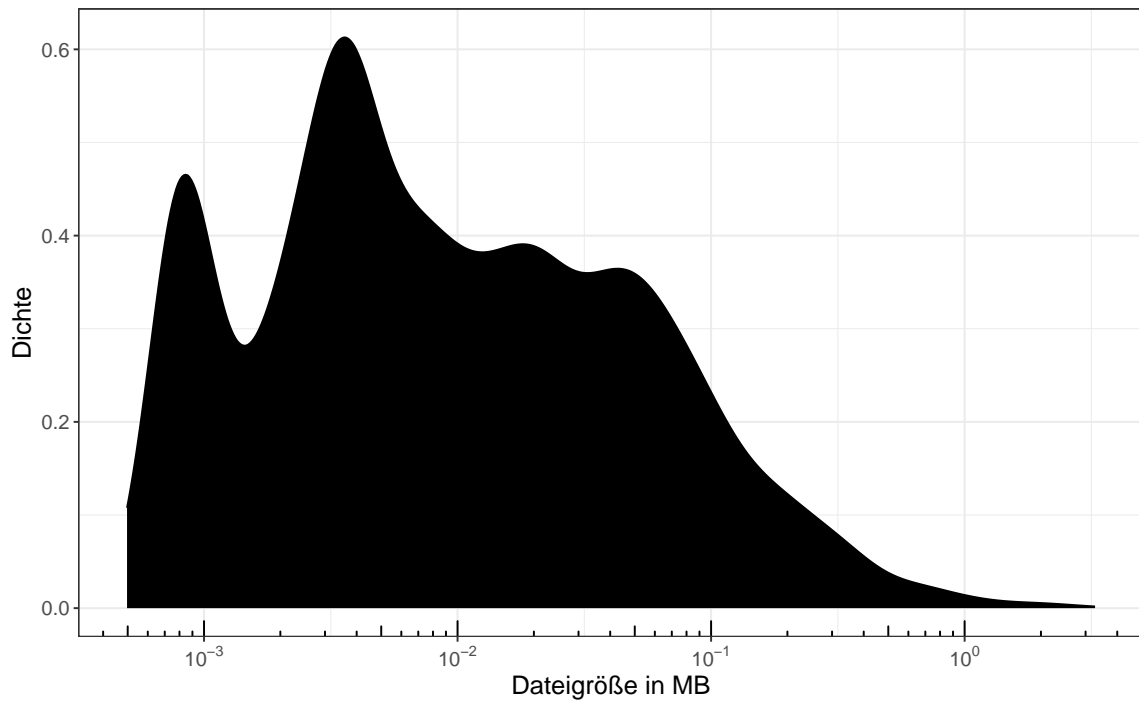
Jahr	Rechtsakte	% Gesamt	% Kumulativ
1974	59	0.88	21.32
1975	73	1.09	22.42
1976	82	1.23	23.64
1977	70	1.05	24.69
1978	56	0.84	25.53
1979	56	0.84	26.36
1980	66	0.99	27.35
1981	47	0.70	28.05
1982	46	0.69	28.74
1983	40	0.60	29.34
1984	52	0.78	30.12
1985	51	0.76	30.88
1986	73	1.09	31.97
1987	46	0.69	32.66
1988	61	0.91	33.57
1989	59	0.88	34.45
1990	170	2.54	37.00
1991	80	1.20	38.19
1992	90	1.35	39.54
1993	98	1.47	41.00
1994	184	2.75	43.76
1995	112	1.67	45.43
1996	103	1.54	46.97
1997	134	2.00	48.98
1998	132	1.97	50.95
1999	92	1.38	52.33
2000	71	1.06	53.39
2001	122	1.82	55.21
2002	151	2.26	57.47
2003	114	1.70	59.17
2004	148	2.21	61.39

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
2005	146	2.18	63.57
2006	134	2.00	65.57
2007	114	1.70	67.28
2008	147	2.20	69.48
2009	163	2.44	71.92
2010	130	1.94	73.86
2011	132	1.97	75.83
2012	118	1.76	77.60
2013	178	2.66	80.26
2014	100	1.50	81.76
2015	130	1.94	83.70
2016	150	2.24	85.94
2017	178	2.66	88.60
2018	92	1.38	89.98
2019	110	1.64	91.63
2020	166	2.48	94.11
2021	244	3.65	97.76
2022	150	2.24	100.00
Total	6687	100.00	100.00

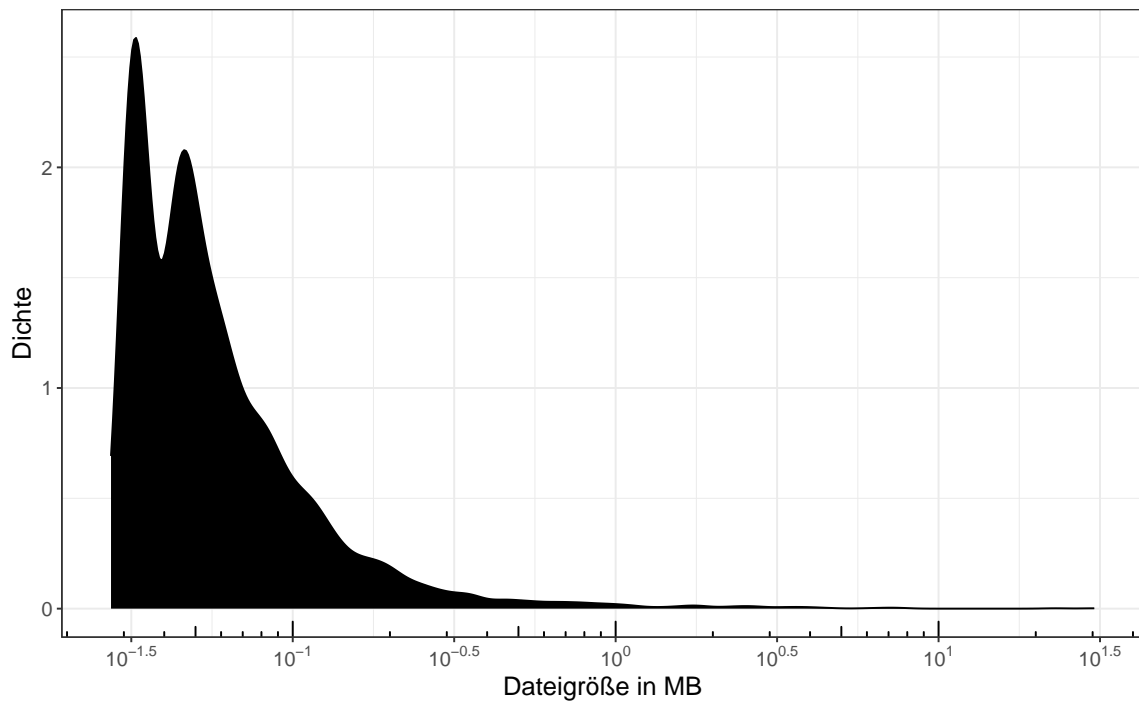
8 Dateigrößen

C-DBR | Version 2023-01-05 | Verteilung der Dateigrößen (XML)



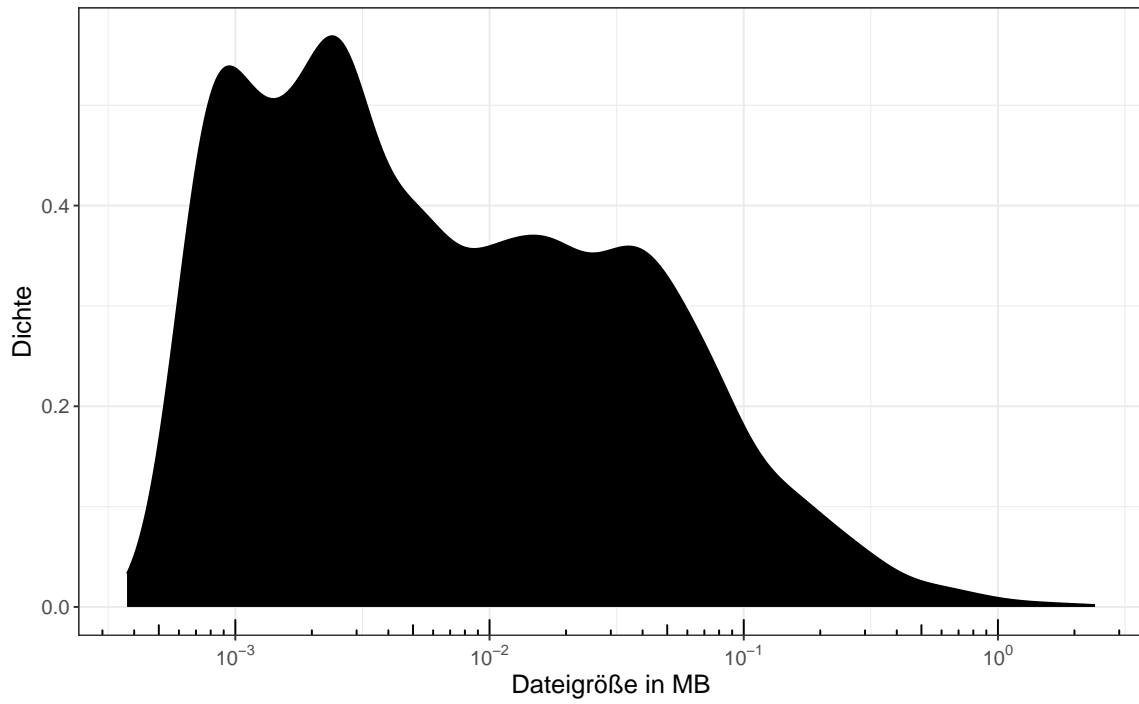
Fobbe | DOI: 10.5281/zenodo.7494474

C-DBR | Version 2023-01-05 | Verteilung der Dateigrößen (PDF)



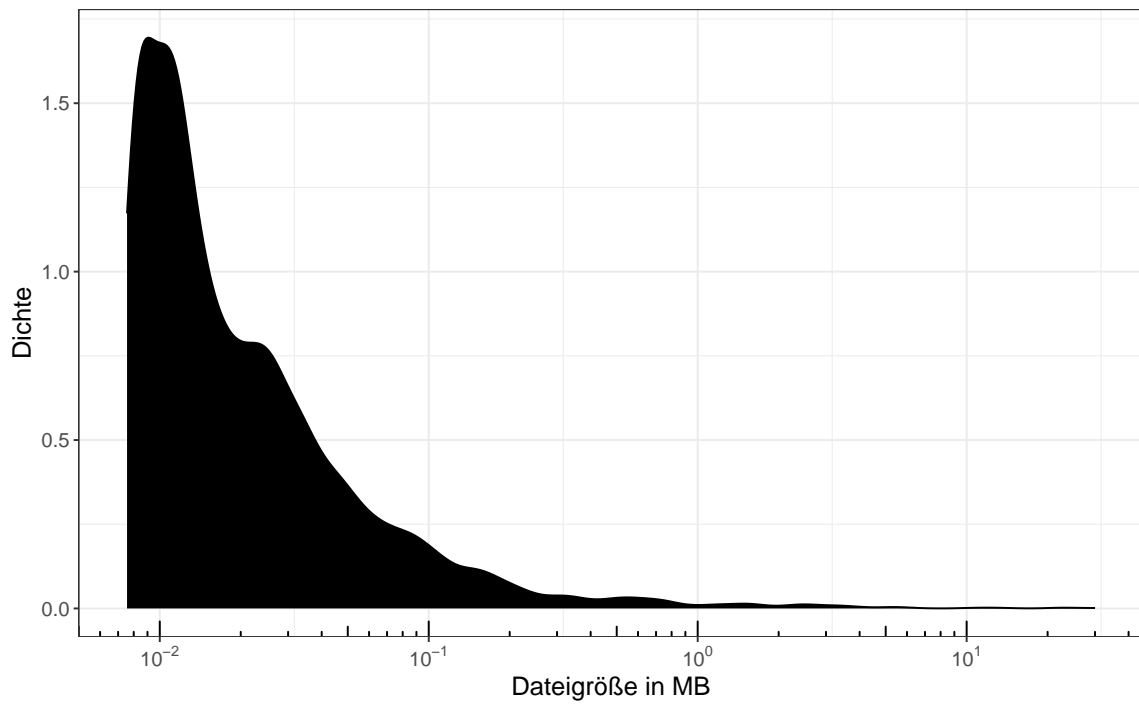
Fobbe | DOI: 10.5281/zenodo.7494474

C-DBR | Version 2023-01-05 | Verteilung der Dateigrößen (TXT)



Fobbe | DOI: 10.5281/zenodo.7494474

C-DBR | Version 2023-01-05 | Verteilung der Dateigrößen (EPUB)



Fobbe | DOI: 10.5281/zenodo.7494474

9 Kryptographische Signaturen

9.1 Zwei-Phasen-Signatur

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, das Codebook und die Robustness Checks Hash-Werte in zwei verschiedenen Verfahren (SHA2-256 und SHA3-512) berechnet und in einer CSV-Datei dokumentiert.

In **Phase II** werden diese CSV-Datei und der Compilation Report 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.

9.2 Persönliche GPG-Signatur

Die während der Kompilierung des Datensatzes erstellte CSV-Datei mit den Hash-Prüfsummen und der Compilation Report sind 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

10 Changelog

10.1 Version 2023-01-05

- Vollständige Aktualisierung der Daten
- Neuer Entwurf des gesamten Source Codes im {targets} Framework
- Zusätzliche Netzwerk-Diagramme für alle Rechtsakte: Sunburst und Circlepacking
- Reguläre Netzwerk-Diagramme nun in blau auf schwarzem Hintergrund
- Manche finale Dateinamen nun mit Trennstrichen statt Pascal Case
- TXT-Konvertierung bricht bei Fehler nicht ab, dokumentiert aber fehlende TXT-Dateien
- Einführung eines separaten Berichts für Robustness Checks

10.2 Version 2022-08-05

- Vollständige Aktualisierung der Daten
- Wenn der Download einer Datei scheitert wird der Kompilierungs-Prozess nicht mehr abgebrochen; Kontrolle über Datenabgleich im Compilation Report
- Diagramme für Norm/Rechtsakt/Metadaten je Periodikum sind nun logarithmisch skaliert
- Technischer Bugfix bei der Berechnung von Netzwerkdiagrammen
- Neuer Unit Test um identische Länge von HTML-Links und extrahierten PDF- und EPUB-Dateinamen
- Fehlende PDF- oder EPUB-Dateien führen nun nicht mehr zu Fehlern in der Pipeline
- Unterscheidung zwischen VBVG 2005 und VBVG 2023

10.3 Version 2022-05-22

- Vollständige Aktualisierung der Daten
- README und CHANGELOG sind nun externe Dateien die bei der Kompilierung automatisch eingebunden werden
- Das für *renv* notwendige Skript *activate.R* ist im ZIP-Archiv in den Ordner “*renv*” sortiert

10.4 Version 2022-01-12

- Vollständige Aktualisierung der Daten
- Strenge Versionskontrolle aller R packages
- Der Prozess der Kompilierung ist jetzt detailliert konfigurierbar, insbesondere die Parallelisierung
- Parallelisierung der XML-Parser deaktiviert, weil instabil
- Parallelisierung nun vollständig mit *future* statt mit *foreach* und *doParallel*
- Fehlerhafte Kompilierungen werden beim vor der nächsten Kompilierung vollautomatisch aufgeräumt
- Alle Ergebnisse werden automatisch fertig verpackt in den Ordner »output« sortiert
- Source Code des Changelogs zu Markdown konvertiert
- Einführung eines Debugging-Modus um die Entwicklung zu beschleunigen

10.5 Version 2021-09-16

- Vollständige Aktualisierung der Daten
- Einfügung von Kurzbezeichnungen der Rechtsakte in die Dateinamen der Netzwerkanalysen
- Einfügung der ID der Rechtsakte in die CSV-Tabelle aller Kurz- und Langtitel

10.6 Version 2021-07-30

- Vollständige Aktualisierung der Daten
- Einführung von neuen Variablen für letzte Änderung (Datum), Neufassung (Datum), Aufhebung (Datum jeweils für Verkündung und Wirkung), Lizenz und hierarchische Ketten von Gliederungsbezeichnungen und -titeln
- Parallelisierung der Downloads um Kompilierung des Korpus zu beschleunigen
- Korrektur bei den Dateinamen der Allgemeinen Eisenbahngesetze: GII weist zwei gleichnamige Rechtsakte (»Allgemeines Eisenbahngesetz«) nach. Beide werden nun mit dem Jahr ihrer Ausfertigung 1951 und 1993 im Langtitel differenziert. In der Vorversion wurde das neuere AEG noch mit dem Jahr 1994 (Inkrafttreten) beschriftet und das andere AEG ohne Jahreszahl.
- Einführung von Netzwerkanalysen (experimentell!)
- Variablen in CSV-Dateien sind nun semantisch sortiert
- Neues Diagramm für Verteilung von Zeichen
- Falls die XML-Datei mehrere Bemerkungen für Hinweise, Änderung, Neufassung, den Stand oder sonstige Angaben aufweist werden diese nun durch einen vertikalen Strich getrennt (vorher nur mehrere Leerzeichen).
- Kleinere Korrekturen und Ergänzungen im Codebook

10.7 Version 2021-01-05

- Vollständige Aktualisierung der Daten
- Komplette Überarbeitung des Source Codes
- Erstveröffentlichung eines Codebooks
- Einführung der vollautomatischen Erstellung von Datensatz und Codebook
- Einführung von Compilation Reports um den Erstellungsprozess exakt zu dokumentieren
- CSV-Dateien werden durch Parsing der XML-Dateien erstellt
- Automatisierung und deutliche Erweiterung der Qualitätskontrolle
- Einführung von Diagrammen zur Visualisierung von Prüfergebnissen
- Einführung kryptographischer Signaturen

10.8 Version 2020-10-09

- Vollständige Aktualisierung der Daten
- Erstveröffentlichung des Source Codes
- XML-Daten nun fehlerfrei. In Version 2020-07-08 waren XML-Dateien mit Anhängen fehlerhaft.

10.9 Version 2020-07-08

- Vollständige Aktualisierung der Daten

10.10 Version 2020-05-18

- Erstveröffentlichung

11 Parameter für strenge Replikationen

```
## [1] "OpenSSL 3.0.5 5 Jul 2022 (Library: OpenSSL 3.0.5 5 Jul 2022)"
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora Linux 36 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libflexiblas.so.3.2
##
## 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] stats      graphics  grDevices datasets  utils      methods    base
##
## other attached packages:
## [1] future.apply_1.10.0 future_1.29.0      quanteda_3.2.3
## [4] readtext_0.81      data.table_1.14.6 scales_1.2.1
## [7] igraph_1.3.5       ggraph_2.1.0      ggplot2_3.4.0
## [10] pdftools_3.3.2     kableExtra_1.3.4  knitr_1.41
## [13] rvest_1.0.3        xml2_1.3.3         httr_1.4.4
## [16] mgsub_1.7.3        zip_2.2.2          fs_1.5.2
## [19] testthat_3.1.5     RcppTOML_0.1.7     tarchetypes_0.7.2
## [22] targets_0.14.0
##
## loaded via a namespace (and not attached):
## [1] webshot_0.5.4      rprojroot_2.0.3    tools_4.1.3
## [4] backports_1.4.1    utf8_1.2.2         R6_2.5.1
## [7] colorspace_2.0-3  withr_2.5.0        tidyselect_1.2.0
## [10] gridExtra_2.3      processx_3.8.0     compiler_4.1.3
## [13] cli_3.4.1          desc_1.4.2         labeling_0.4.2
## [16] stringfish_0.15.7 callr_3.7.3         askpass_1.1
## [19] systemfonts_1.0.4 stringr_1.4.1       digest_0.6.30
## [22] rmarkdown_2.18     svglite_2.1.0      pkgconfig_2.0.3
## [25] htmltools_0.5.4    parallely_1.32.1   fastmap_1.1.0
## [28] rlang_1.0.6        rstudioapi_0.14    RApiSerialize_0.1.2
## [31] farver_2.1.1       generics_0.1.3     dplyr_1.0.10
## [34] magrittr_2.0.3     Matrix_1.4-0       Rcpp_1.0.9
## [37] munsell_0.5.0      fansi_1.0.3        viridis_0.6.2
## [40] lifecycle_1.0.3   stringi_1.6.2      yaml_2.2.1
## [43] MASS_7.3-55        brio_1.1.3         grid_4.1.3
## [46] parallel_4.1.3     listenv_0.8.0      ggrepel_0.9.2
## [49] lattice_0.20-45    graphlayouts_0.8.3 magick_2.7.3
## [52] ps_1.7.2           pillar_1.8.1       base64url_1.4
## [55] pkgload_1.3.2      codetools_0.2-18   stopwords_2.3
## [58] fastmatch_1.1-3    glue_1.6.2         evaluate_0.18
```

```
## [61] qpdf_1.3.0          renv_0.16.0          RcppParallel_5.1.5
## [64] vctrs_0.5.1         tweenr_2.0.2         selectr_0.4-2
## [67] gtable_0.3.1        purrr_0.3.5          polyclip_1.10-4
## [70] tidyr_1.2.1         qs_0.25.4            xfun_0.35
## [73] ggforce_0.4.1       tidygraph_1.2.2     viridisLite_0.4.1
## [76] tibble_3.1.8        globals_0.16.1
```


Literaturverzeichnis

- Allaire, JJ, Jeffrey Horner, Yihui Xie, Vicent Marti, and Natacha Porte. 2019. *Markdown: Render Markdown with the c Library Sundown*. <https://github.com/rstudio/markdown>.
- Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2022. *Rmarkdown: Dynamic Documents for r*. <https://CRAN.R-project.org/package=rmarkdown>.
- Almende B.V. and Contributors, and Benoit Thieurmél. 2022. *visNetwork: Network Visualization Using Vis.js Library*. <https://datastorm-open.github.io/visNetwork/>.
- Bengtsson, Henrik. 2021. “A Unifying Framework for Parallel and Distributed Processing in r Using Futures.” *The R Journal* 13 (2): 208–27. <https://doi.org/10.32614/RJ-2021-048>.
- . 2022a. *Future: Unified Parallel and Distributed Processing in r for Everyone*. <https://CRAN.R-project.org/package=future>.
- . 2022b. *Future.apply: Apply Function to Elements in Parallel Using Futures*. <https://CRAN.R-project.org/package=future.apply>.
- 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. 2022. *Quanteda: Quantitative Analysis of Textual Data*. <https://quanteda.io>.
- Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2022. *Shiny: Web Application Framework for r*. <https://shiny.rstudio.com/>.
- Ching, Travers. 2022. *Qs: Quick Serialization of r Objects*. <https://github.com/traversc/qs>.
- Csardi, Gabor, and Tamas Nepusz. 2006. “The Igraph Software Package for Complex Network Research.” *InterJournal Complex Systems*: 1695. <https://igraph.org>.
- Csárdi, Gábor. 2022. *Pingr: Check If a Remote Computer Is up*. <https://github.com/r-lib/pingr#readme>.
- Csárdi, Gábor, Kuba Podgórski, and Rich Geldreich. 2022. *Zip: Cross-Platform Zip Compression*. <https://github.com/r-lib/zip#readme>.
- Dowle, Matt, and Arun Srinivasan. 2022. *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>.
- file., See AUTHORS. 2022. *Igraph: Network Analysis and Visualization*. <https://CRAN.R-project.org/package=igraph>.
- Granjon, David. 2022. *bs4Dash: A Bootstrap 4 Version of Shinydashboard*. <https://CRAN.R-project.org/package=bs4Dash>.
- Hester, Jim, Hadley Wickham, and Gábor Csárdi. 2021. *Fs: Cross-Platform File System Operations Based on Libuv*. <https://CRAN.R-project.org/package=fs>.
- Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, and JooYoung Seo. 2022. *Gt: Easily Create Presentation-Ready Display Tables*. <https://CRAN.R-project.org/package=gt>.
- Landau, William Michael. 2021a. *Tarchetypes: Archetypes for Targets*.

- . 2021b. “The Targets r Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing.” *Journal of Open Source Software* 6 (57): 2959. <https://doi.org/10.21105/joss.02959>.
- . 2022a. *Tarchetypes: Archetypes for Targets*. <https://CRAN.R-project.org/package=tarchetypes>.
- . 2022b. *Targets: Dynamic Function-Oriented Make-Like Declarative Workflows*. <https://CRAN.R-project.org/package=targets>.
- Meyer, Fanny, and Victor Perrier. 2022. *Shinybusy: Busy Indicators and Notifications for Shiny Applications*. <https://github.com/dreamRs/shinybusy>.
- Ooms, Jeroen. 2021. *Magick: Advanced Graphics and Image-Processing in r*. <https://CRAN.R-project.org/package=magick>.
- . 2022a. *Openssl: Toolkit for Encryption, Signatures and Certificates Based on OpenSSL*. <https://github.com/jeroen/openssl>.
- . 2022b. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. <https://CRAN.R-project.org/package=pdfutils>.
- Pedersen, Thomas Lin. 2022. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. <https://CRAN.R-project.org/package=ggraph>.
- Perrier, Victor, Fanny Meyer, and David Granjon. 2022. *shinyWidgets: Custom Inputs Widgets for Shiny*. <https://CRAN.R-project.org/package=shinyWidgets>.
- Schubert, Michael. 2022. *Clustermq: Evaluate Function Calls on HPC Schedulers (LSF, SGE, SLURM, PBS/Torque)*. <https://mschubert.github.io/clustermq/>.
- Ushey, Kevin. 2022. *Renv: Project Environments*. <https://rstudio.github.io/renv/>.
- Ushey, Kevin, JJ Allaire, Hadley Wickham, and Gary Ritchie. 2022. *Rstudioapi: Safely Access the RStudio API*. <https://CRAN.R-project.org/package=rstudioapi>.
- Wickham, Hadley. 2011. “Testthat: Get Started with Testing.” *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.
- . 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- . 2022a. *Httr: Tools for Working with URLs and HTTP*. <https://CRAN.R-project.org/package=httr>.
- . 2022b. *Rvest: Easily Harvest (Scrape) Web Pages*. <https://CRAN.R-project.org/package=rvest>.
- . 2022c. *Testthat: Unit Testing for r*. <https://CRAN.R-project.org/package=testthat>.
- Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2022. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. <https://CRAN.R-project.org/package=ggplot2>.
- Wickham, Hadley, Jim Hester, and Jeroen Ooms. 2021. *Xml2: Parse XML*. <https://CRAN.R-project.org/package=xml2>.
- Wickham, Hadley, and Dana Seidel. 2022. *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/>.
- . 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://yihui.org/knitr/>.

- Xie, Yihui, J. J. Allaire, and Garrett Golemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>.
- Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown-cookbook>.
- Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.