Corpus des Deutschen Bundesrechts

(C-DBR)

Codebook

Version 2022-08-05

DOI: 10.5281/zenodo.6959374

Titel	Corpus des Deutschen Bundesrechts
Abkürzung	C-DBR
Autor	Seán Fobbe
Version	2022-08-05
Download	https://doi.org/10.5281/zenodo.6959374
Lizenz	CC0 1.0 Universal

Zitiervorschlag

Seán Fobbe (2022). Corpus des Deutschen Bundesrechts (C-DBR). Version 2022-08-05. Zenodo. DOI: 10.5281/zenodo.6959374.

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 2022-08-05. 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. 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

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 vorgesehen 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.

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*. CSV[6] 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.

Hier 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.6959381

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 Addresse hinterlegt: https://doi.org/10.5281/zenodo.6959381

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

3.5 Grenzen des Datensatzes

Nutzer sollten folgende wichtige Grenzen beachten:

1. 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]
2. Es kann aufgrund technischer Grenzen bzw. Fehler sein, dass manche — im Grunde verfügbare — Rechtsakte nicht oder nicht korrekt abgerufen werden (*automation bias*).
3. 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 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 Webeite 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 versierte 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 hochexperimentell, 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!

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«.
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«.

Variable	Typ	Erläuterung
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 *regular expressions* 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 *regular expressions* 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 *regular expressions* extrahiert.
neufassung_datum	String	Das Datum an dem der Rechtsakt zuletzt neugefasst wurde. Das Format ist YYYY-MM-DD (Langform nach ISO-8601). Aus der Variable »neuf« mittels *regular expressions* extrahiert.
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 Jareszahl, zweistellige Monatszahl, zweistellige Tageszahl, zweistellige Stundenzahl, zweistellige Minutenzahl und eine zweistellige Sekundenzahl.

Variable	Typ	Erläuterung
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 Jareszahl, 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 »BGBl 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!
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!

Variable	Typ	Erläuterung
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	(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	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.

Variable	Typ	Erläuterung
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 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	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-Variable	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 drei verschiedene klassische linguistische Kennzahlen:

Variable	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 »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!

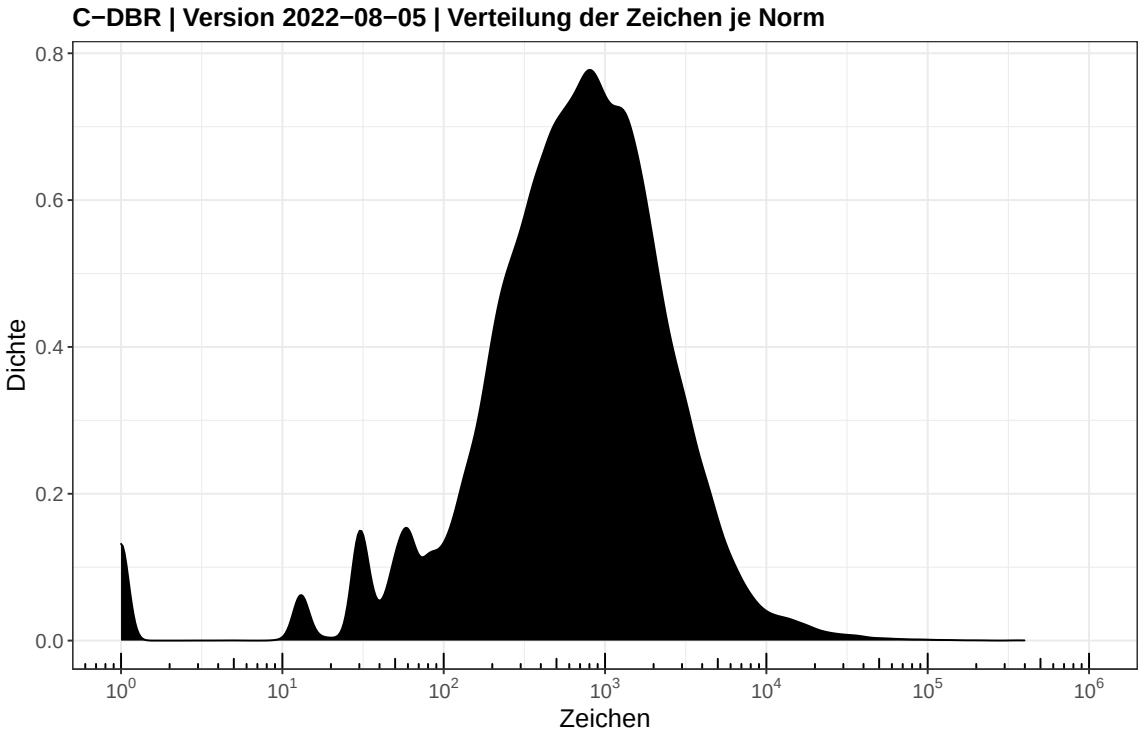
6.2 Kennzahlen: Einzelnormen

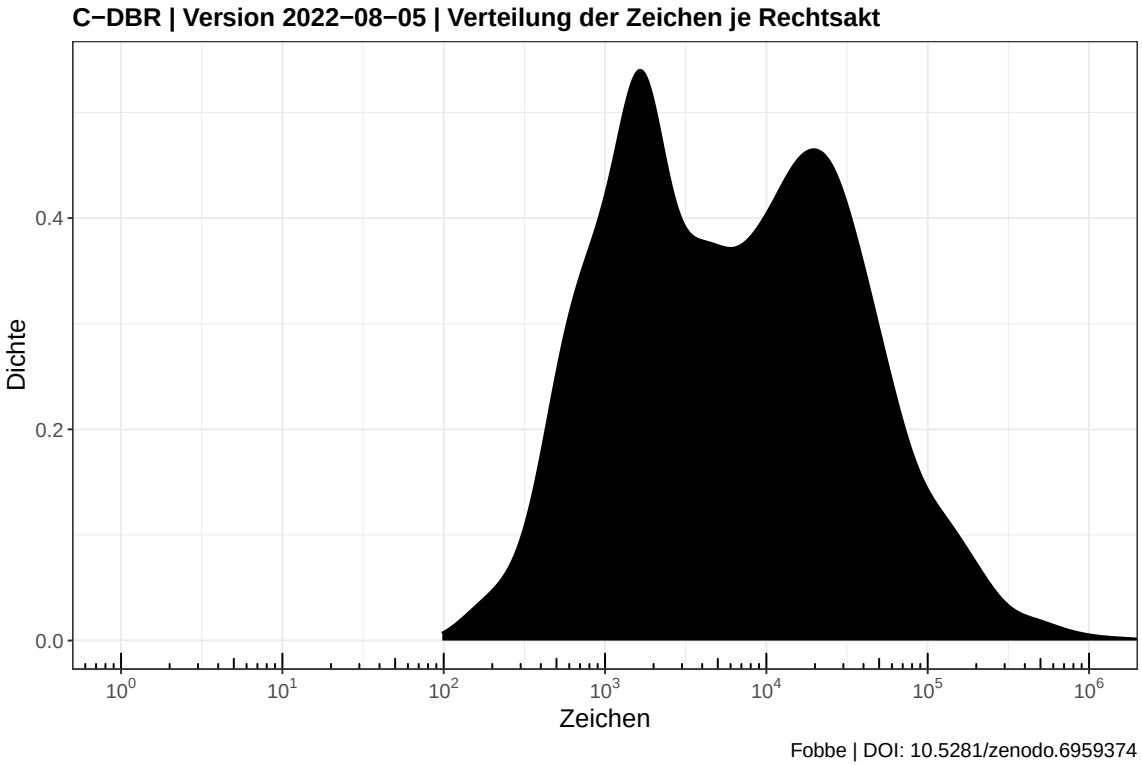
Kennzahl	Gesamt	Min	1. Quartil	Median	Mittel	3. Quartil	Max
Zeichen	144,030,059	1	294	690	1,421.19	1,493	398,079
Tokens	22,529,761	1	47	109	222.31	234	56,584
Typen	421,806	1	36	66	93.36	115	23,559
Sätze	752,160	1	2	4	7.42	8	2,666

6.3 Kennzahlen: Rechtsakte

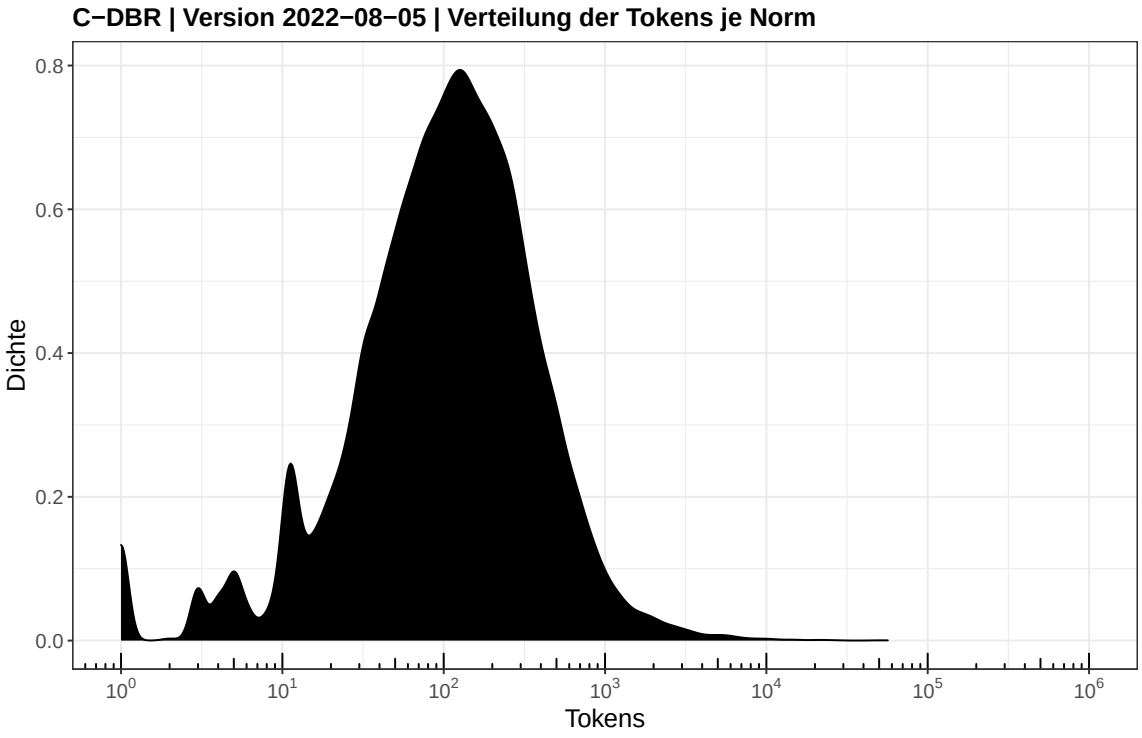
Kennzahl	Gesamt	Min	1. Quartil	Median	Mittel	3. Quartil	Max
Zeichen	144,125,735	98	1,569	5,717	25,423.48	22,801	2,087,193
Tokens	22,529,761	17	265	925	3,974.20	3,457	306,803
Typen	421,806	15	130	305	646.17	813	23,606
Sätze	740,726	1	13	36	130.66	115	14,306

6.4 Verteilung Zeichen

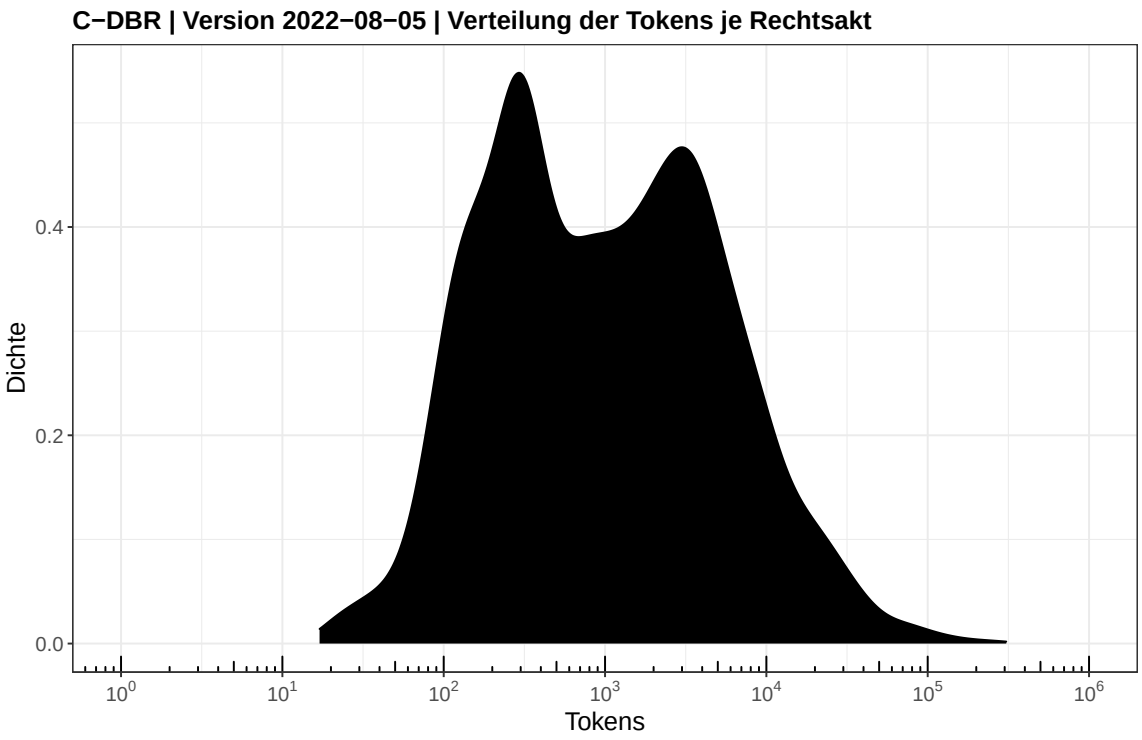




6.5 Verteilung Tokens

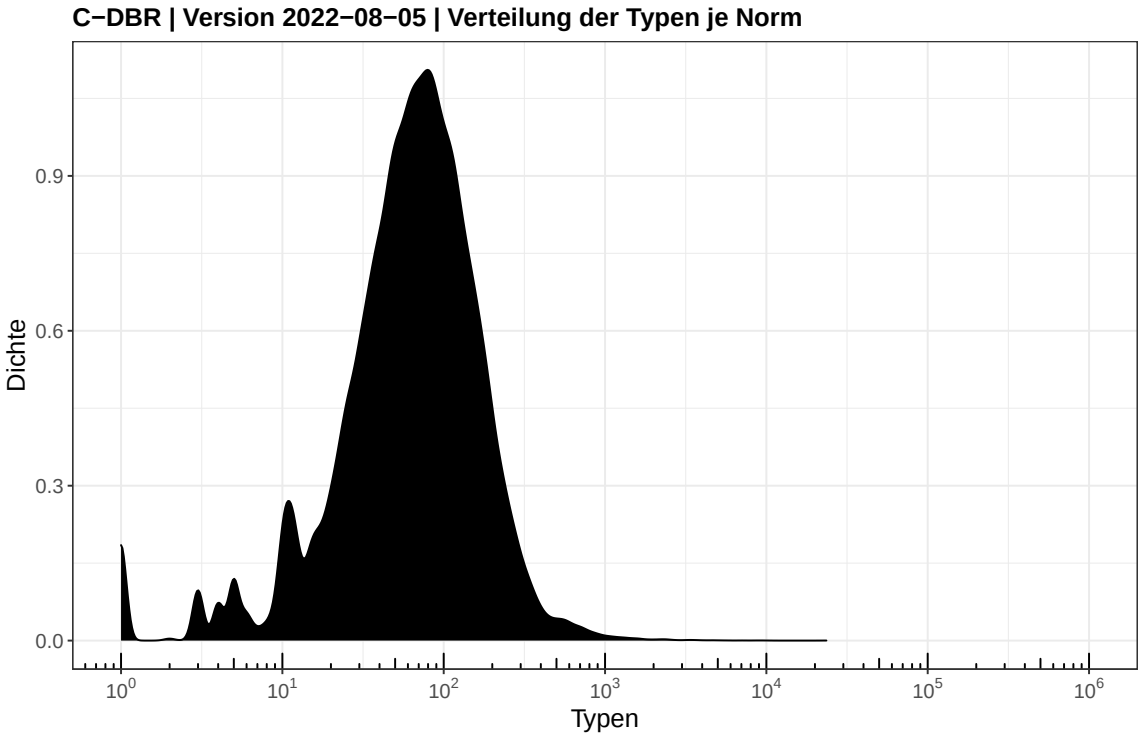


Fobbe | DOI: 10.5281/zenodo.6959374

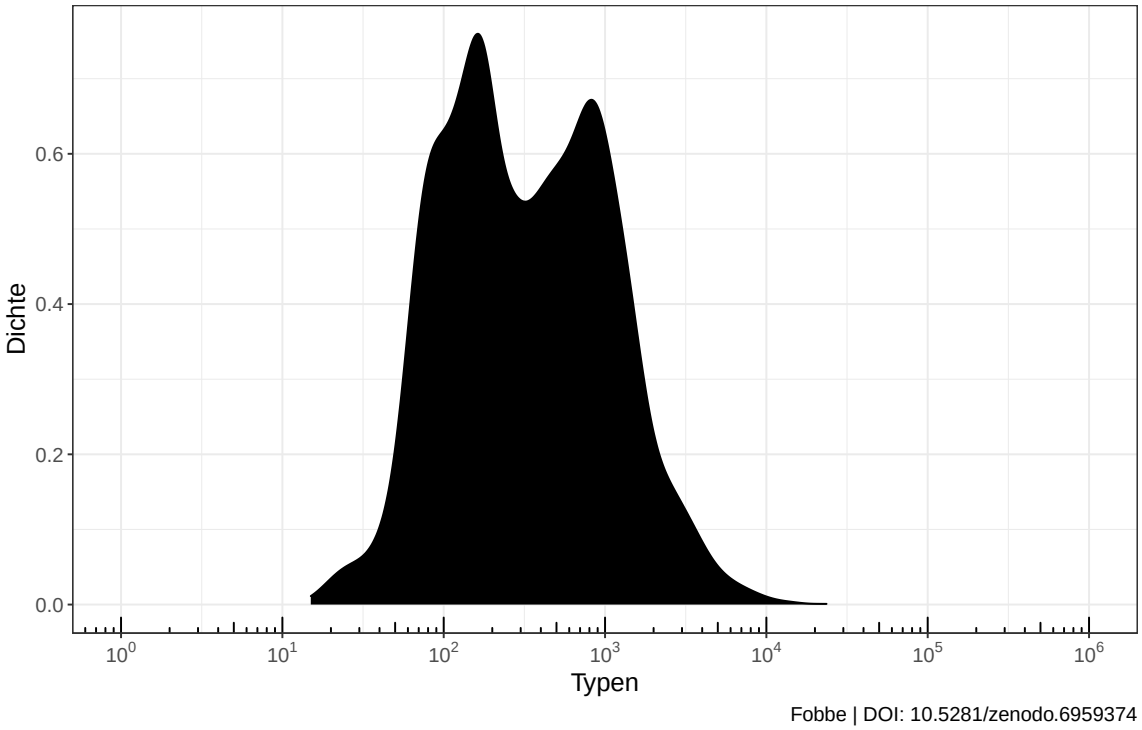


Fobbe | DOI: 10.5281/zenodo.6959374

6.6 Verteilung Typen

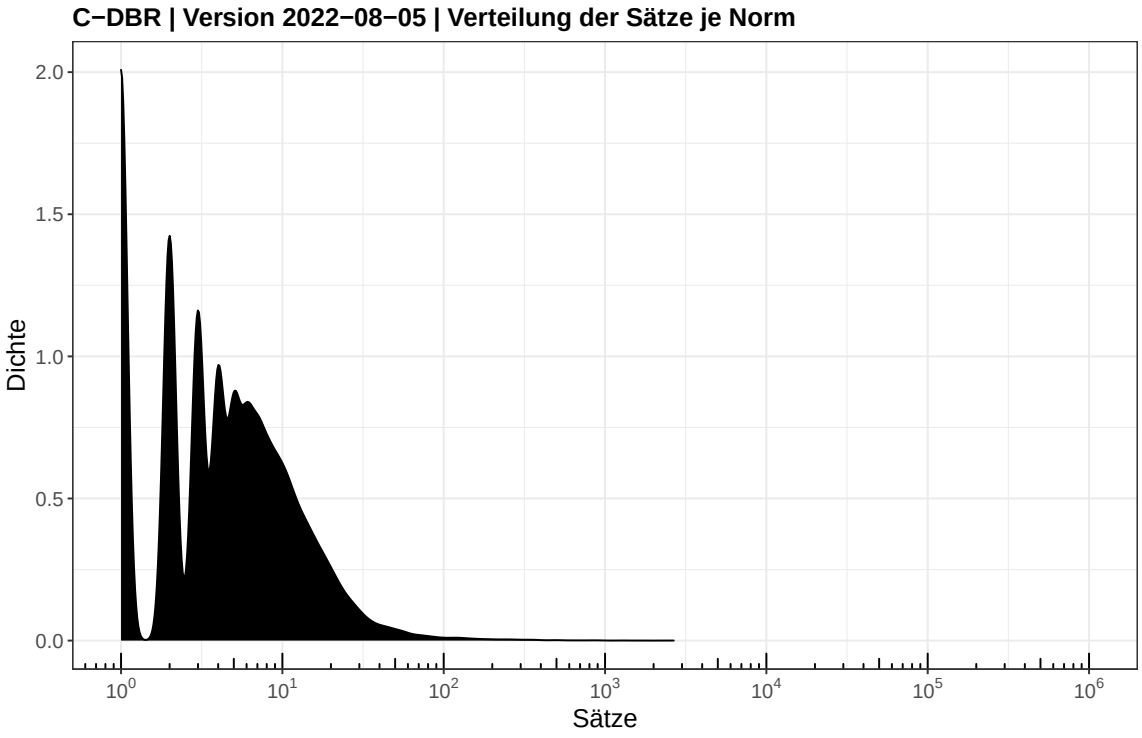


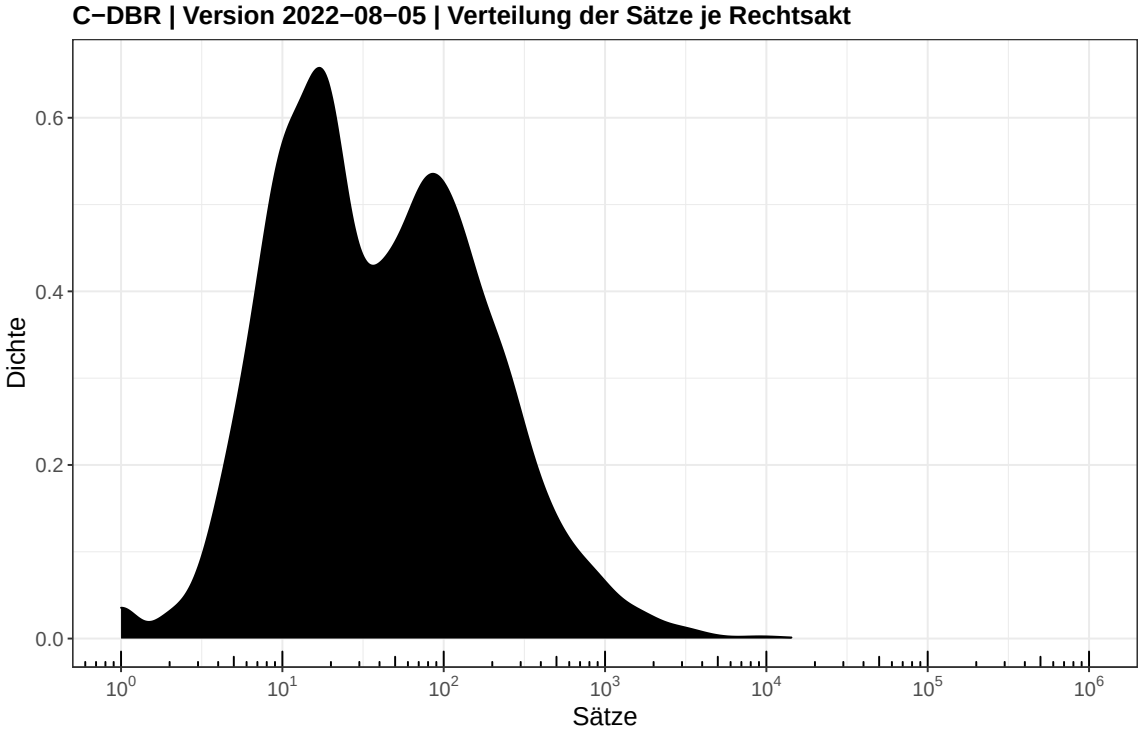
Fobbe | DOI: 10.5281/zenodo.6959374



Fobbe | DOI: 10.5281/zenodo.6959374

6.7 Verteilung Sätze

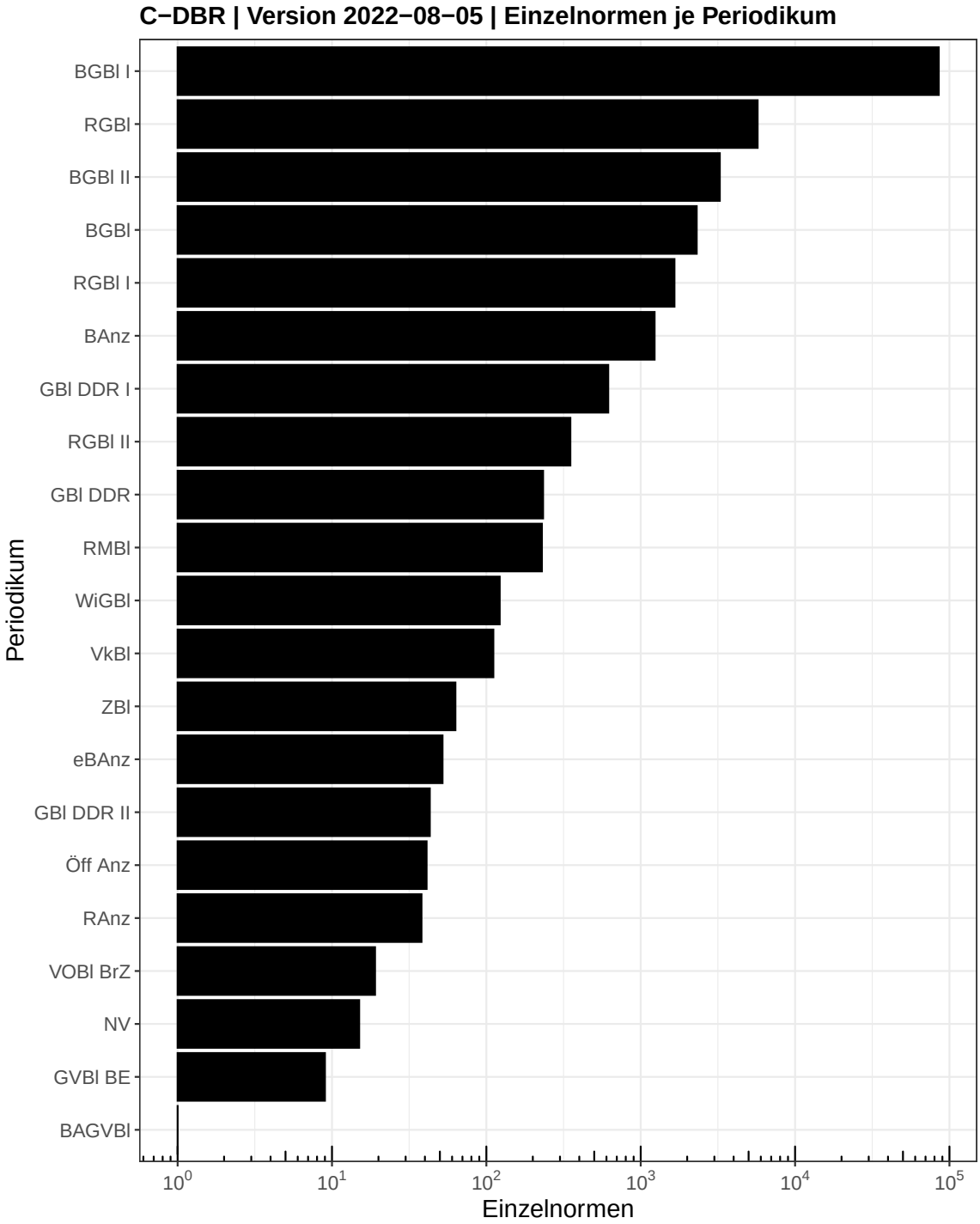




7 Inhalt

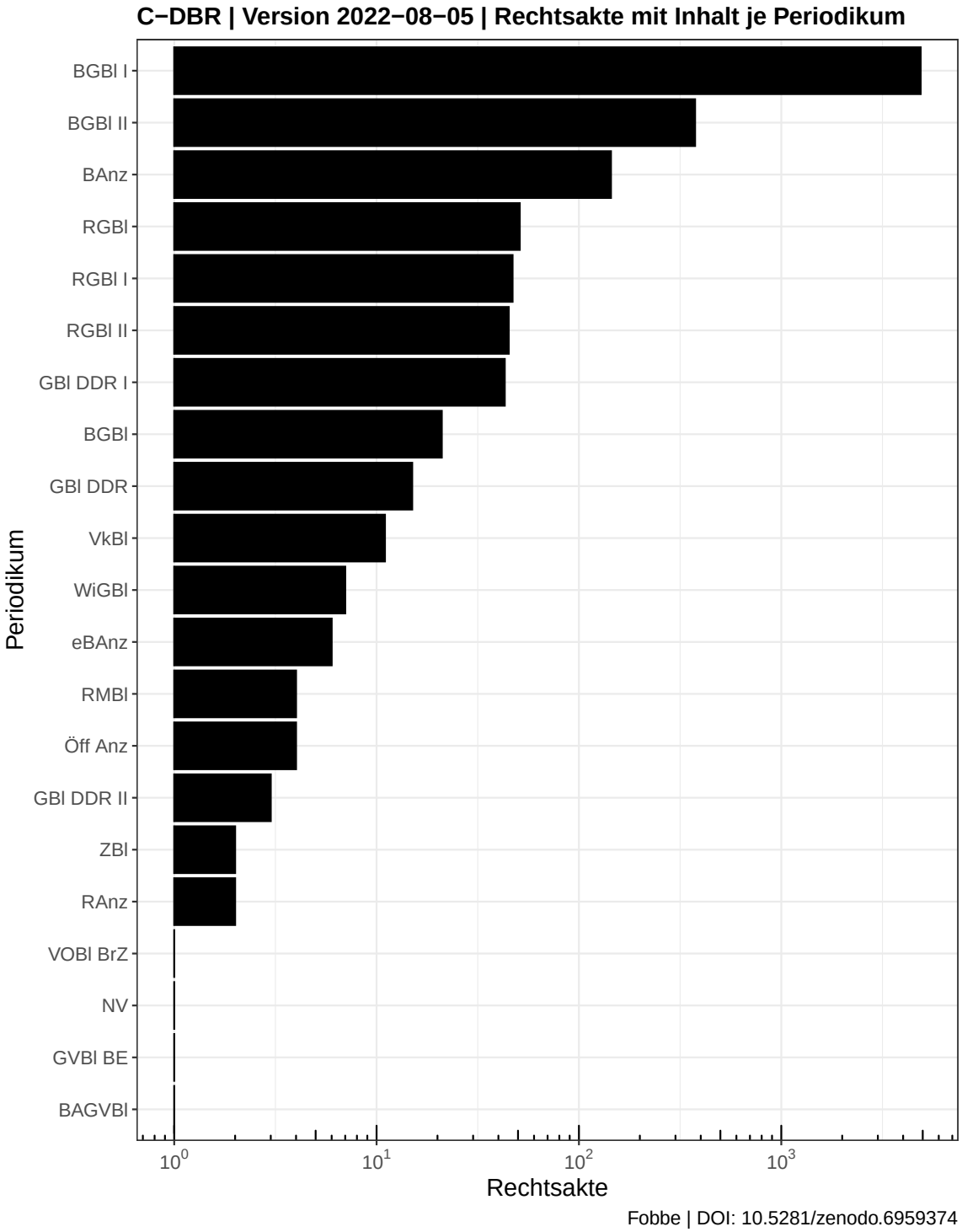
7.1 Nach Periodikum

7.1.1 Einzelnormen



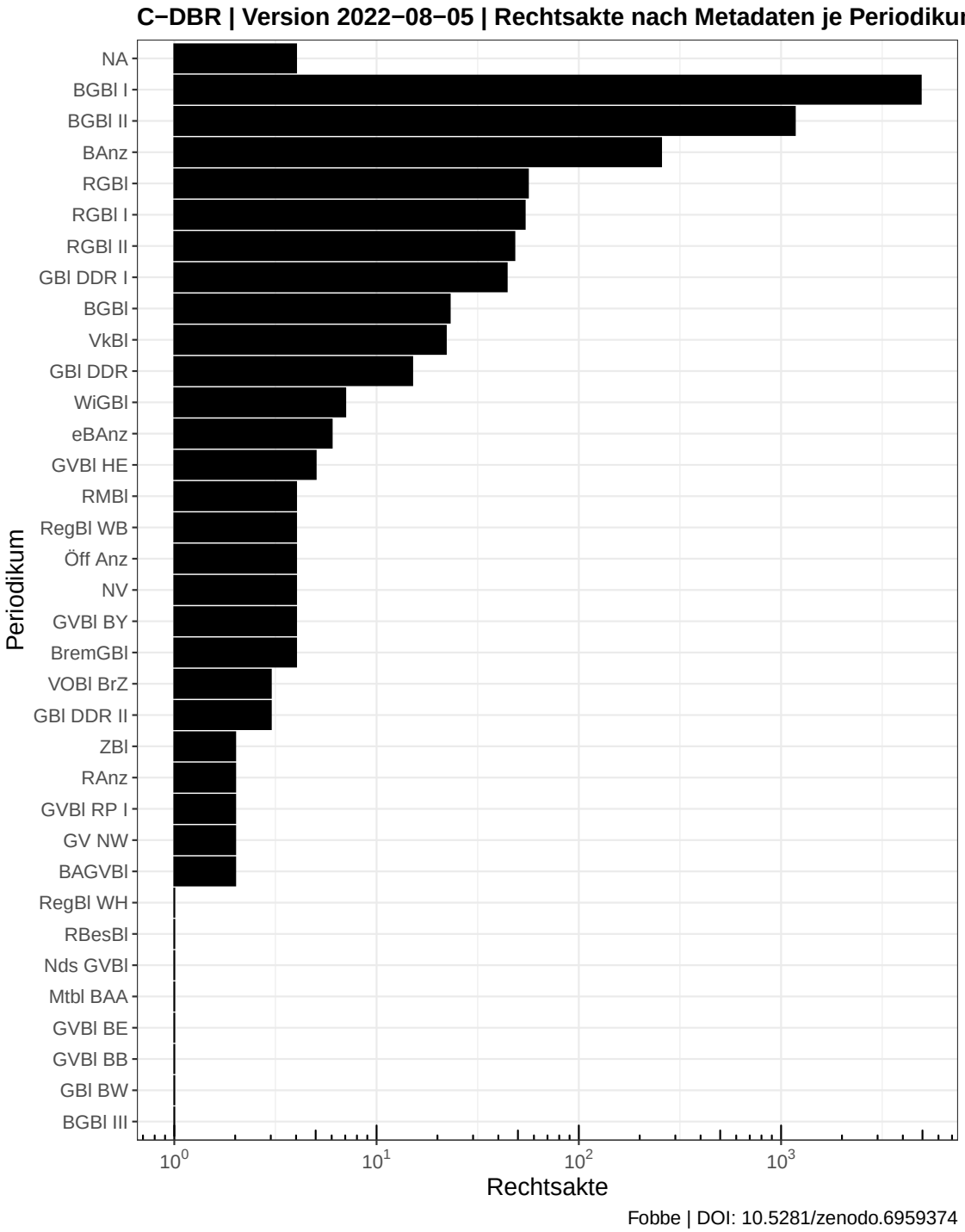
Periodikum	Einzelnormen	% Gesamt	% Kumulativ
BAGVBl	1	0.00	0.00
BAnz	1230	1.21	1.21
BGBl	2305	2.27	3.49
BGBl I	85243	84.11	87.60
BGBl II	3253	3.21	90.81
GBl DDR	233	0.23	91.04
GBl DDR I	616	0.61	91.65
GBl DDR II	43	0.04	91.69
GVBl BE	9	0.01	91.70
NV	15	0.01	91.71
RAnz	38	0.04	91.75
RGBl	5718	5.64	97.39
RGBl I	1654	1.63	99.03
RGBl II	350	0.35	99.37
RMBl	229	0.23	99.60
VOBl BrZ	19	0.02	99.62
VkBl	111	0.11	99.73
WiGBl	122	0.12	99.85
ZBl	63	0.06	99.91
eBAnz	52	0.05	99.96
Öff Anz	41	0.04	100.00

7.1.2 Rechtsakte mit veröffentlichtem Normtext



Periodikum	Rechtsakte	% Gesamt	% Kumulativ
BAGVBl	1	0.02	0.02
BAnz	144	2.54	2.56
BGBl	21	0.37	2.93
BGBl I	4885	86.17	89.10
BGBl II	375	6.61	95.71
GBl DDR	15	0.26	95.98
GBl DDR I	43	0.76	96.74
GBl DDR II	3	0.05	96.79
GVBl BE	1	0.02	96.81
NV	1	0.02	96.82
RAnz	2	0.04	96.86
RGBl	51	0.90	97.76
RGBl I	47	0.83	98.59
RGBl II	45	0.79	99.38
RMBl	4	0.07	99.45
VOBl BrZ	1	0.02	99.47
VkBl	11	0.19	99.66
WiGBl	7	0.12	99.79
ZBl	2	0.04	99.82
eBAnz	6	0.11	99.93
Öff Anz	4	0.07	100.00

7.1.3 Alle Rechtsakte (mit und ohne Normtext)



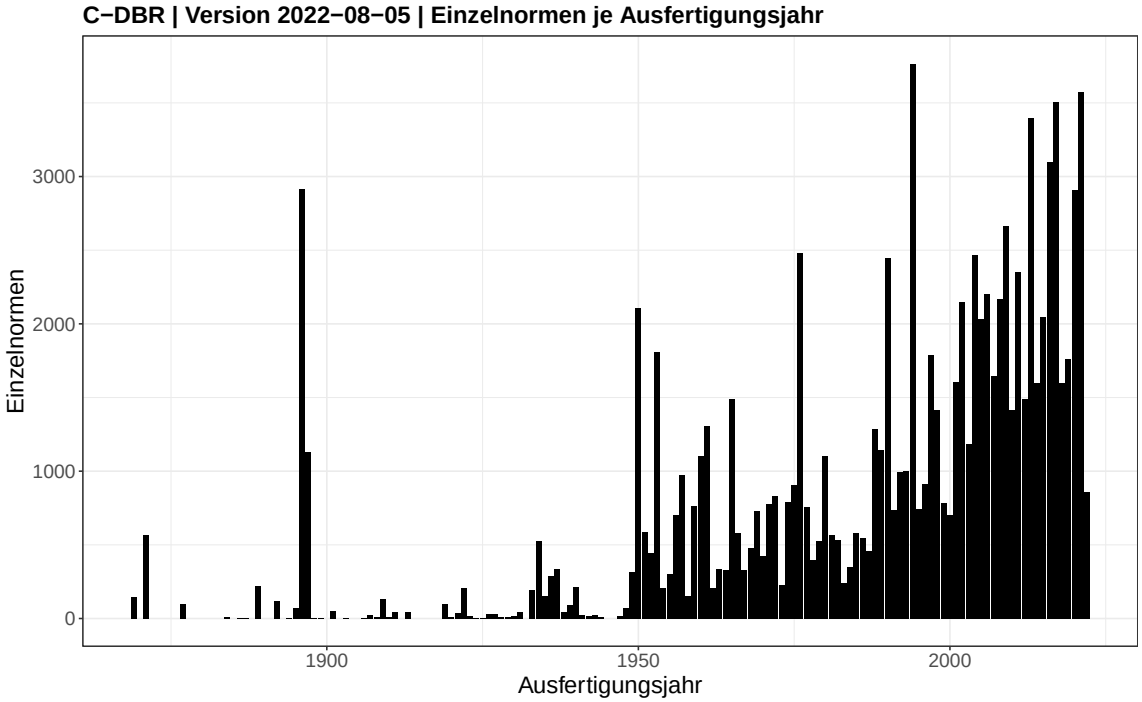
Periodikum	Rechtsakte	% Gesamt	% Kumulativ
NA	4	0.06	0.06
BAGVBl	2	0.03	0.09
BAnz	255	3.83	3.92
BGBl	23	0.35	4.26
BGBl I	4903	73.63	77.89
BGBl II	1169	17.56	95.45
BGBl III	1	0.02	95.46
BremGBl	4	0.06	95.52
GBl BW	1	0.02	95.54
GBl DDR	15	0.23	95.77
GBl DDR I	44	0.66	96.43
GBl DDR II	3	0.05	96.47
GV NW	2	0.03	96.50
GVBl BB	1	0.02	96.52
GVBl BE	1	0.02	96.53
GVBl BY	4	0.06	96.59
GVBl HE	5	0.08	96.67
GVBl RP I	2	0.03	96.70
Mtbl BAA	1	0.02	96.71
NV	4	0.06	96.77
Nds GVBl	1	0.02	96.79
RAnz	2	0.03	96.82
RBesBl	1	0.02	96.83
RGBl	56	0.84	97.67
RGBl I	54	0.81	98.48
RGBl II	48	0.72	99.20
RMBl	4	0.06	99.26
RegBl WB	4	0.06	99.32
RegBl WH	1	0.02	99.34
VOBl BrZ	3	0.05	99.38
VkBl	22	0.33	99.71
WiGBl	7	0.11	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

7.2 Nach Ausfertigungsjahr

7.2.1 Einzelnormen



Jahr	Einzelnormen	% Gesamt	% Kumulativ
1869	149	0.15	0.15
1871	564	0.56	0.70
1877	98	0.10	0.80
1884	10	0.01	0.81
1886	3	0.00	0.81
1887	4	0.00	0.82
1889	218	0.22	1.03
1892	117	0.12	1.15
1894	4	0.00	1.15
1895	74	0.07	1.22
1896	2916	2.88	4.10
1897	1133	1.12	5.22
1898	3	0.00	5.22
1899	2	0.00	5.22
1901	51	0.05	5.28
1903	1	0.00	5.28

(continued)

Jahr	Einzelnormen	% Gesamt	% Kumulativ
1906	1	0.00	5.28
1907	26	0.03	5.30
1908	8	0.01	5.31
1909	129	0.13	5.44
1910	11	0.01	5.45
1911	47	0.05	5.50
1913	46	0.05	5.54
1919	99	0.10	5.64
1920	7	0.01	5.65
1921	40	0.04	5.68
1922	206	0.20	5.89
1923	15	0.01	5.90
1924	2	0.00	5.90
1925	2	0.00	5.91
1926	31	0.03	5.94
1927	32	0.03	5.97
1928	8	0.01	5.98
1929	8	0.01	5.98
1930	14	0.01	6.00
1931	46	0.05	6.04
1933	196	0.19	6.24
1934	526	0.52	6.76
1935	156	0.15	6.91
1936	287	0.28	7.19
1937	335	0.33	7.52
1938	41	0.04	7.56
1939	90	0.09	7.65
1940	215	0.21	7.87
1941	25	0.02	7.89
1942	20	0.02	7.91
1943	24	0.02	7.93

(continued)

Jahr	Einzelnormen	% Gesamt	% Kumulativ
1944	12	0.01	7.95
1947	19	0.02	7.96
1948	73	0.07	8.04
1949	314	0.31	8.35
1950	2106	2.08	10.42
1951	584	0.58	11.00
1952	445	0.44	11.44
1953	1806	1.78	13.22
1954	209	0.21	13.43
1955	303	0.30	13.73
1956	704	0.69	14.42
1957	975	0.96	15.38
1958	150	0.15	15.53
1959	766	0.76	16.29
1960	1106	1.09	17.38
1961	1310	1.29	18.67
1962	210	0.21	18.88
1963	336	0.33	19.21
1964	332	0.33	19.54
1965	1487	1.47	21.00
1966	579	0.57	21.58
1967	330	0.33	21.90
1968	477	0.47	22.37
1969	729	0.72	23.09
1970	425	0.42	23.51
1971	776	0.77	24.28
1972	831	0.82	25.10
1973	228	0.22	25.32
1974	788	0.78	26.10
1975	906	0.89	26.99
1976	2480	2.45	29.44

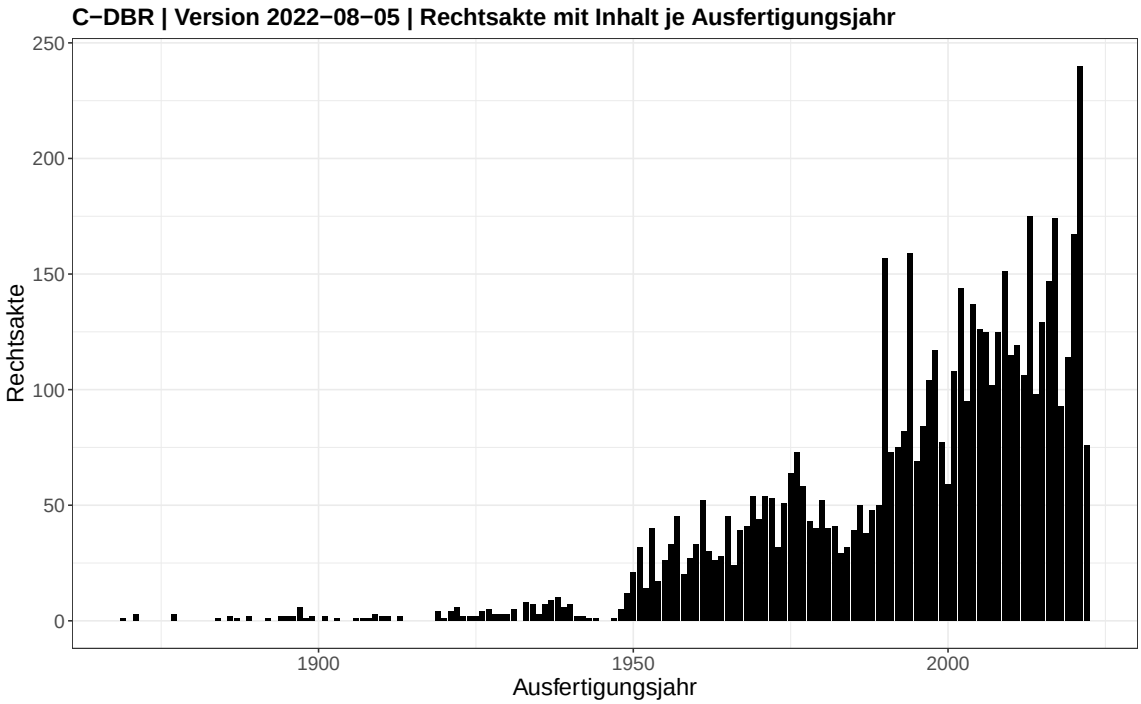
(continued)

Jahr	Einzelnormen	% Gesamt	% Kumulativ
1977	760	0.75	30.19
1978	399	0.39	30.58
1979	525	0.52	31.10
1980	1103	1.09	32.19
1981	568	0.56	32.75
1982	535	0.53	33.28
1983	243	0.24	33.52
1984	352	0.35	33.87
1985	580	0.57	34.44
1986	548	0.54	34.98
1987	455	0.45	35.43
1988	1284	1.27	36.69
1989	1143	1.13	37.82
1990	2445	2.41	40.23
1991	734	0.72	40.96
1992	994	0.98	41.94
1993	1000	0.99	42.93
1994	3766	3.72	46.64
1995	743	0.73	47.38
1996	914	0.90	48.28
1997	1789	1.77	50.04
1998	1416	1.40	51.44
1999	782	0.77	52.21
2000	705	0.70	52.91
2001	1604	1.58	54.49
2002	2150	2.12	56.61
2003	1182	1.17	57.78
2004	2465	2.43	60.21
2005	2032	2.01	62.22
2006	2204	2.17	64.39
2007	1648	1.63	66.02

(continued)

Jahr	Einzelnormen	% Gesamt	% Kumulativ
2008	2166	2.14	68.15
2009	2661	2.63	70.78
2010	1416	1.40	72.18
2011	2354	2.32	74.50
2012	1491	1.47	75.97
2013	3396	3.35	79.32
2014	1601	1.58	80.90
2015	2047	2.02	82.92
2016	3102	3.06	85.98
2017	3507	3.46	89.44
2018	1596	1.57	91.02
2019	1763	1.74	92.76
2020	2909	2.87	95.63
2021	3575	3.53	99.15
2022	857	0.85	100.00

7.2.2 Rechtsakte mit veröffentlichtem Normtext



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	0.48
1899	2	0.04	0.51
1901	2	0.04	0.55
1903	1	0.02	0.56
1906	1	0.02	0.58

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1907	1	0.02	0.60
1908	1	0.02	0.62
1909	3	0.05	0.67
1910	2	0.04	0.71
1911	2	0.04	0.74
1913	2	0.04	0.78
1919	4	0.07	0.85
1920	1	0.02	0.86
1921	4	0.07	0.93
1922	6	0.11	1.04
1923	2	0.04	1.08
1924	2	0.04	1.11
1925	2	0.04	1.15
1926	4	0.07	1.22
1927	5	0.09	1.31
1928	3	0.05	1.36
1929	3	0.05	1.41
1930	3	0.05	1.46
1931	5	0.09	1.55
1933	8	0.14	1.69
1934	7	0.12	1.82
1935	3	0.05	1.87
1936	7	0.12	1.99
1937	9	0.16	2.15
1938	10	0.18	2.33
1939	6	0.11	2.43
1940	7	0.12	2.56
1941	2	0.04	2.59
1942	2	0.04	2.63
1943	1	0.02	2.65
1944	1	0.02	2.66

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1947	1	0.02	2.68
1948	5	0.09	2.77
1949	12	0.21	2.98
1950	21	0.37	3.35
1951	32	0.56	3.92
1952	14	0.25	4.16
1953	40	0.71	4.87
1954	17	0.30	5.17
1955	26	0.46	5.63
1956	33	0.58	6.21
1957	45	0.79	7.00
1958	20	0.35	7.36
1959	27	0.48	7.83
1960	33	0.58	8.41
1961	52	0.92	9.33
1962	30	0.53	9.86
1963	26	0.46	10.32
1964	28	0.49	10.81
1965	45	0.79	11.61
1966	24	0.42	12.03
1967	39	0.69	12.72
1968	41	0.72	13.44
1969	54	0.95	14.39
1970	44	0.78	15.17
1971	54	0.95	16.12
1972	53	0.93	17.06
1973	32	0.56	17.62
1974	51	0.90	18.52
1975	64	1.13	19.65
1976	73	1.29	20.94
1977	58	1.02	21.96

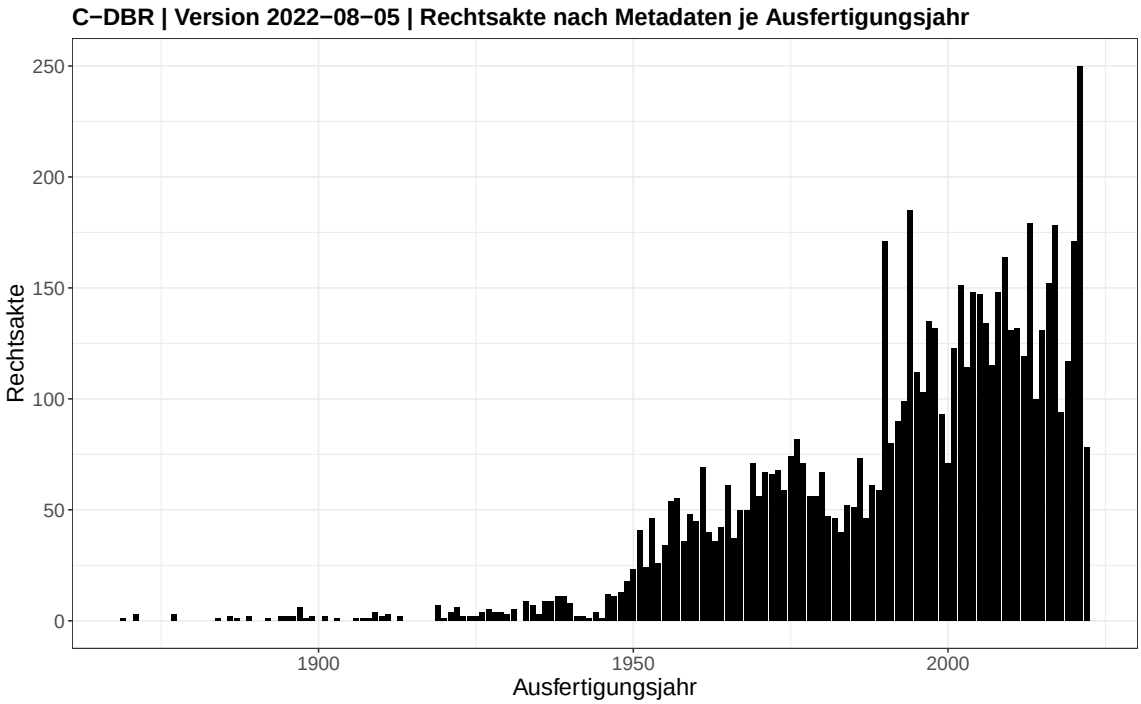
(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1978	43	0.76	22.72
1979	40	0.71	23.43
1980	52	0.92	24.34
1981	40	0.71	25.05
1982	41	0.72	25.77
1983	29	0.51	26.28
1984	32	0.56	26.85
1985	39	0.69	27.54
1986	50	0.88	28.42
1987	38	0.67	29.09
1988	48	0.85	29.93
1989	50	0.88	30.82
1990	157	2.77	33.59
1991	73	1.29	34.87
1992	75	1.32	36.20
1993	82	1.45	37.64
1994	159	2.80	40.45
1995	69	1.22	41.67
1996	84	1.48	43.15
1997	104	1.83	44.98
1998	117	2.06	47.05
1999	77	1.36	48.40
2000	59	1.04	49.44
2001	108	1.91	51.35
2002	144	2.54	53.89
2003	95	1.68	55.57
2004	137	2.42	57.98
2005	126	2.22	60.20
2006	125	2.20	62.41
2007	102	1.80	64.21
2008	125	2.20	66.41

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
2009	151	2.66	69.08
2010	115	2.03	71.11
2011	119	2.10	73.21
2012	106	1.87	75.07
2013	175	3.09	78.16
2014	98	1.73	79.89
2015	129	2.28	82.17
2016	147	2.59	84.76
2017	174	3.07	87.83
2018	93	1.64	89.47
2019	114	2.01	91.48
2020	167	2.95	94.43
2021	240	4.23	98.66
2022	76	1.34	100.00

7.2.3 Alle Rechtsakte (mit und ohne Normtext)



Jahr	Rechtsakte	% Gesamt	% Kumulativ
NA	1	0.02	0.02
1869	1	0.02	0.03
1871	3	0.05	0.08
1877	3	0.05	0.12
1884	1	0.02	0.14
1886	2	0.03	0.17
1887	1	0.02	0.18
1889	2	0.03	0.21
1892	1	0.02	0.23
1894	2	0.03	0.26
1895	2	0.03	0.29
1896	2	0.03	0.32
1897	6	0.09	0.41
1898	1	0.02	0.42
1899	2	0.03	0.45
1901	2	0.03	0.48
1903	1	0.02	0.50

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1906	1	0.02	0.51
1907	1	0.02	0.53
1908	1	0.02	0.54
1909	4	0.06	0.60
1910	2	0.03	0.63
1911	3	0.05	0.68
1913	2	0.03	0.71
1919	7	0.11	0.81
1920	1	0.02	0.83
1921	4	0.06	0.89
1922	6	0.09	0.98
1923	2	0.03	1.01
1924	2	0.03	1.04
1925	2	0.03	1.07
1926	4	0.06	1.13
1927	5	0.08	1.20
1928	4	0.06	1.26
1929	4	0.06	1.32
1930	3	0.05	1.37
1931	5	0.08	1.44
1933	9	0.14	1.58
1934	7	0.11	1.68
1935	3	0.05	1.73
1936	9	0.14	1.86
1937	9	0.14	2.00
1938	11	0.17	2.16
1939	11	0.17	2.33
1940	8	0.12	2.45
1941	2	0.03	2.48
1942	2	0.03	2.51
1943	1	0.02	2.52

(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
1944	4	0.06	2.58
1945	1	0.02	2.60
1946	12	0.18	2.78
1947	11	0.17	2.94
1948	13	0.20	3.14
1949	18	0.27	3.41
1950	23	0.35	3.75
1951	41	0.62	4.37
1952	24	0.36	4.73
1953	46	0.69	5.42
1954	26	0.39	5.81
1955	34	0.51	6.32
1956	54	0.81	7.13
1957	55	0.83	7.96
1958	36	0.54	8.50
1959	48	0.72	9.22
1960	45	0.68	9.90
1961	69	1.04	10.93
1962	40	0.60	11.53
1963	36	0.54	12.07
1964	42	0.63	12.70
1965	61	0.92	13.62
1966	37	0.56	14.18
1967	50	0.75	14.93
1968	50	0.75	15.68
1969	71	1.07	16.74
1970	56	0.84	17.59
1971	67	1.01	18.59
1972	66	0.99	19.58
1973	68	1.02	20.60
1974	59	0.89	21.49

(continued)

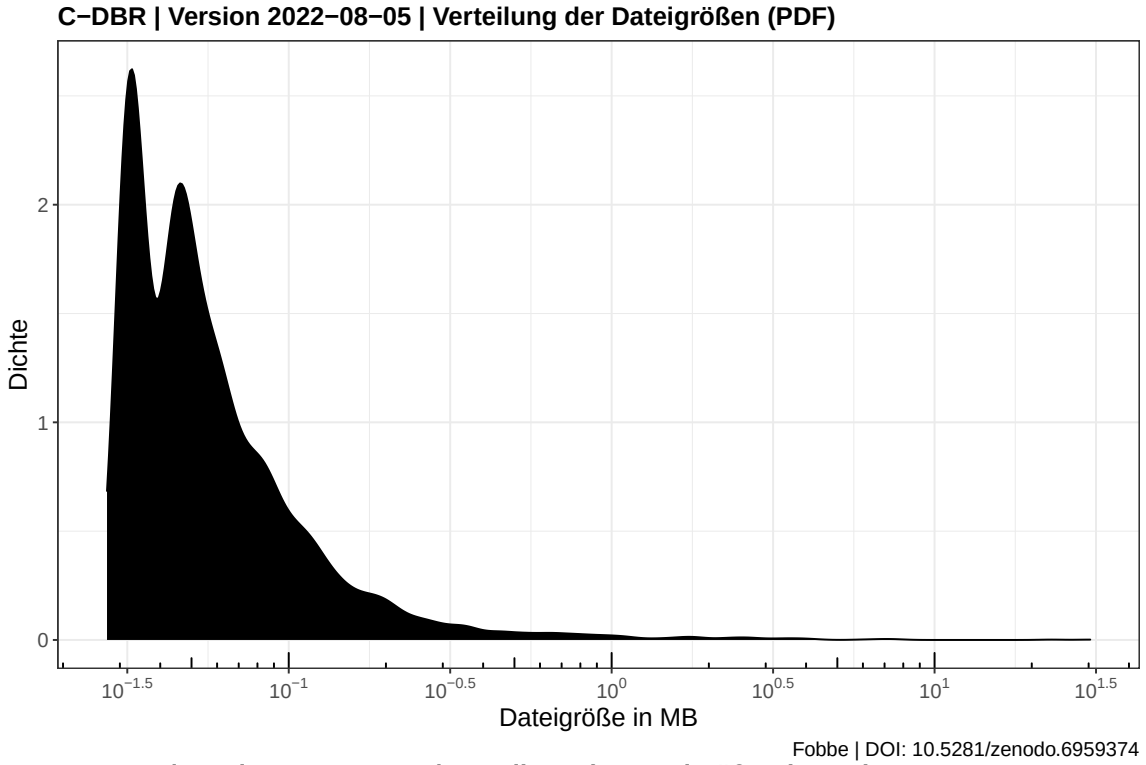
Jahr	Rechtsakte	% Gesamt	% Kumulativ
1975	74	1.11	22.60
1976	82	1.23	23.83
1977	71	1.07	24.90
1978	56	0.84	25.74
1979	56	0.84	26.58
1980	67	1.01	27.59
1981	47	0.71	28.29
1982	46	0.69	28.98
1983	40	0.60	29.58
1984	52	0.78	30.36
1985	51	0.77	31.13
1986	73	1.10	32.23
1987	46	0.69	32.92
1988	61	0.92	33.83
1989	59	0.89	34.72
1990	171	2.57	37.29
1991	80	1.20	38.49
1992	90	1.35	39.84
1993	99	1.49	41.33
1994	185	2.78	44.11
1995	112	1.68	45.79
1996	103	1.55	47.33
1997	135	2.03	49.36
1998	132	1.98	51.34
1999	93	1.40	52.74
2000	71	1.07	53.81
2001	123	1.85	55.65
2002	151	2.27	57.92
2003	114	1.71	59.63
2004	148	2.22	61.86
2005	147	2.21	64.06

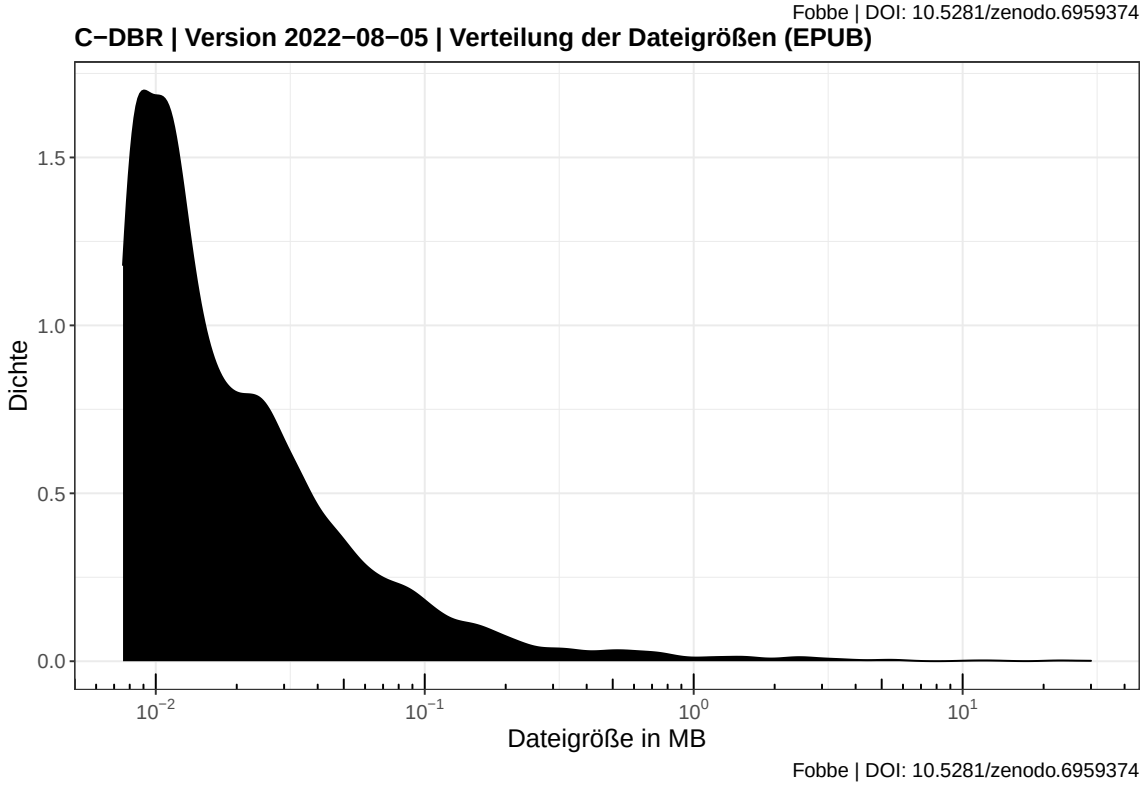
(continued)

Jahr	Rechtsakte	% Gesamt	% Kumulativ
2006	134	2.01	66.08
2007	115	1.73	67.80
2008	148	2.22	70.03
2009	164	2.46	72.49
2010	131	1.97	74.46
2011	132	1.98	76.44
2012	119	1.79	78.22
2013	179	2.69	80.91
2014	100	1.50	82.41
2015	131	1.97	84.38
2016	152	2.28	86.66
2017	178	2.67	89.34
2018	94	1.41	90.75
2019	117	1.76	92.51
2020	171	2.57	95.07
2021	250	3.75	98.83
2022	78	1.17	100.00

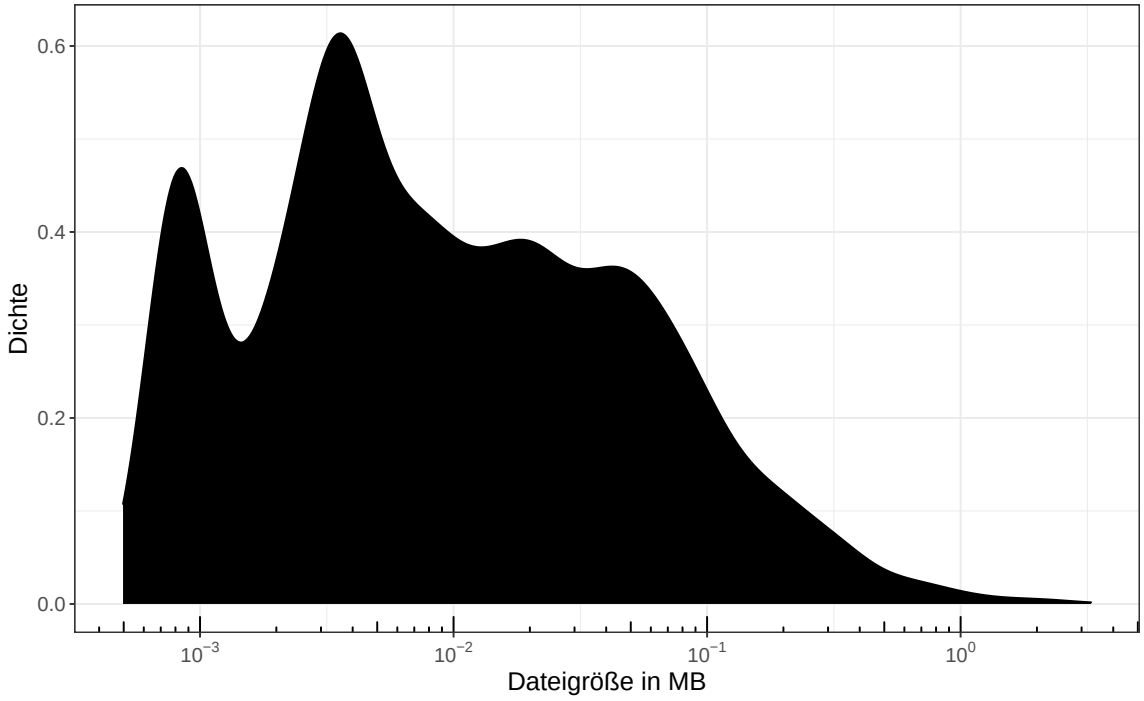
8 Dateigrößen: Summen und Verteilungen

Datei	Größe in MB
C-DBR_2022-08-05_DE_ANALYSE.zip	1.71
C-DBR_2022-08-05_DE_CSV_Einzelnormen_Datensatz.zip	39.91
C-DBR_2022-08-05_DE_CSV_Einzelnormen_Metadaten.zip	3.07
C-DBR_2022-08-05_DE_CSV_MetadatenXML.zip	0.57
C-DBR_2022-08-05_DE_CSV_Rechtsakte_Datensatz.zip	35.44
C-DBR_2022-08-05_DE_CSV_Rechtsakte_Metadaten.zip	0.50
C-DBR_2022-08-05_DE_EPUB_Datensatz.zip	435.78
C-DBR_2022-08-05_DE_Netzwerke.zip	16.44
C-DBR_2022-08-05_DE_PDF_Datensatz.zip	583.76
C-DBR_2022-08-05_DE_TXT_Datensatz.zip	48.01
C-DBR_2022-08-05_DE_XML_Anlagen.zip	267.10
C-DBR_2022-08-05_DE_XML_Datensatz.zip	49.90
C-DBR_2022-08-05_Source_Code.zip	0.10


C−DBR | Version 2022−08−05 | Verteilung der Dateigrößen (PDF)
Dichte
2
1
0
$10^{-1.5}$
10^{-1}
$10^{-0.5}$
10^{0}
$10^{0.5}$
10^{1}
$10^{1.5}$
Dateigröße in MB
Fobbe | DOI: 10.5281/zenodo.6959374

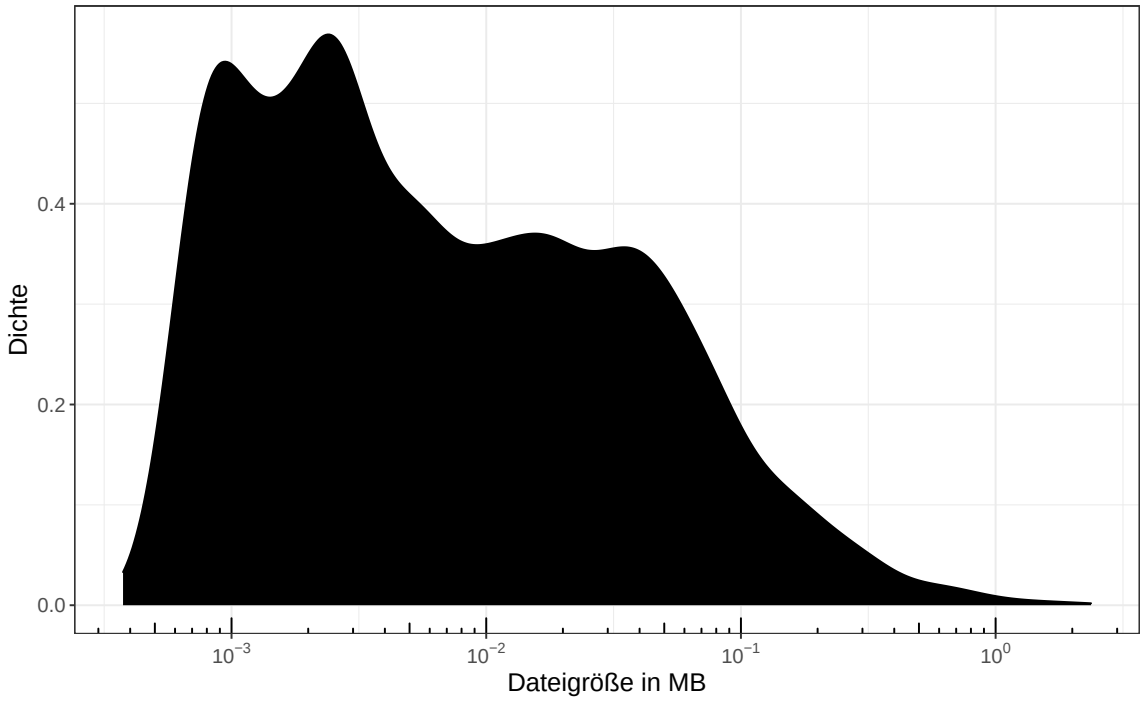

C−DBR | Version 2022−08−05 | Verteilung der Dateigrößen (EPUB)
Dichte
1.5
1.0
0.5
0.0
10^{-2}
10^{-1}
10^{0}
10^{1}
Dateigröße in MB
Fobbe | DOI: 10.5281/zenodo.6959374

C−DBR | Version 2022−08−05 | Verteilung der Dateigrößen (XML)



Fobbe | DOI: 10.5281/zenodo.6959374

C−DBR | Version 2022−08−05 | Verteilung der Dateigrößen (TXT)



Fobbe | DOI: 10.5281/zenodo.6959374

9 Signaturprüfung

9.1 Allgemeines

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

In **Phase I** werden während der Kompilierung für jedes ZIP-Archiv Hash-Werte in zwei verschiedenen Verfahren 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!«

9.2 Persönliche GPG-Signatur

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

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

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

9.3 Import: Public Key

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

9.4 Prüfung: GPG-Signatur der Hash-Datei

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

```
## [1] "output/C-DBR_2022-08-05_KryptographischeHashes.csv"
```

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

```
## [1] "output/C-DBR_2022-08-05_FobbeSignaturGPG_Hashes.gpg"
```

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

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

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

Ergebnis
gpg: Signature made Fri 05 Aug 2022 10:19:42 PM CEST
gpg: using RSA key FE6FB888F0E5656C1D253B9A50C41384F44A4E42
gpg: Good signature from Sean Fobbe <fobbe-data@posteo.de> [unknown]
gpg: WARNING: This key is not certified with a trusted signature!
gpg: There is no indication that the signature belongs to the owner.
Primary key fingerprint: FE6F B888 F0E5 656C 1D25 3B9A 50C4 1384 F44A 4E42

9.5 Prüfung: SHA3-512 Hashes der ZIP-Archive

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

# Ursprüngliche Signaturen importieren
table.hashes <- fread(hashfile)
filename <- 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
C-DBR_2022-08-05_DE_ANALYSE.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_CSV_Einzelnormen_Datensatz.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_CSV_Einzelnormen_Metadaten.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_CSV_MetadatenXML.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_CSV_Rechtsakte_Datensatz.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_CSV_Rechtsakte_Metadaten.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_EPUB_Datensatz.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_Netzwerke.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_PDF_Datensatz.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_TXT_Datensatz.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_XML_Anlagen.zip	Signatur verifiziert!
C-DBR_2022-08-05_DE_XML_Datensatz.zip	Signatur verifiziert!
C-DBR_2022-08-05_Source_Code.zip	Signatur verifiziert!

10 Changelog

10.1 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

10.2 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

10.3 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 deaktivert, 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

10.4 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

10.5 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

10.6 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

10.7 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.

10.8 Version 2020-07-08

- Vollständige Aktualisierung der Daten

10.9 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] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] future.apply_1.9.0 future_1.27.0      ggraph_2.0.5       igraph_1.3.4
##  [5] openssl_2.0.2      scales_1.2.0       quanteda_3.2.1     data.table
    _1.14.2
##  [9] ggplot2_3.3.6      pdftools_3.3.0     magick_2.7.3       kableExtra_1.3.4
## [13] knitr_1.39         RcppTOML_0.1.7     xml2_1.3.3         rvest_1.0.2
## [17] zip_2.2.0
##
## loaded via a namespace (and not attached):
##  [1] viridis_0.6.2      httr_1.4.3         tidyr_1.2.0        bit64_4.0.5
##  [5] tidygraph_1.2.1    viridisLite_0.4.0  RcppParallel_5.1.5 askpass_1.1
##  [9] highr_0.9          selectr_0.4-2      renv_0.15.5        yaml_2.3.5
## [13] ggrepel_0.9.1      globals_0.16.0     qpdf_1.2.0         pillar_1.8.0
## [17] lattice_0.20-45    glue_1.6.2         digest_0.6.29      polyclip_1.10-0
## [21] colorspace_2.0-3   htmltools_0.5.3    Matrix_1.4-1       pkgconfig_2.0.3
## [25] listenv_0.8.0      purrr_0.3.4        webshot_0.5.3      svglite_2.1.0
## [29] tweenr_1.0.2       ggforce_0.3.3      tibble_3.1.8       generics_0.1.3
## [33] farver_2.1.1       withr_2.5.0        cli_3.3.0          magrittr_2.0.3
## [37] evaluate_0.15      stopwords_2.3      fansi_1.0.3        parallelly
    _1.32.1
## [41] MASS_7.3-58.1      tools_4.1.3        lifecycle_1.0.1    stringr_1.4.0
## [45] munsell_0.5.0      compiler_4.1.3     tinytex_0.40       systemfonts
    _1.0.4
## [49] rlang_1.0.4        grid_4.1.3         rstudioapi_0.13    labeling_0.4.2
## [53] rmarkdown_2.14     gtable_0.3.0       codetools_0.2-18   curl_4.3.2
## [57] graphlayouts_0.8.0 R6_2.5.1           gridExtra_2.3      dplyr_1.0.9
## [61] bit_4.0.4          fastmap_1.1.0      utf8_1.2.2         fastmatch_1.1-3
## [65] stringi_1.7.8      Rcpp_1.0.9         vctrs_0.4.1        tidyselect_1.1.2
## [69] xfun_0.31
```

Literaturverzeichnis

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, 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.

Csardi, Gabor, and Tamas Nepusz. 2006. “The Igraph Software Package for Complex Network Research.” *InterJournal* Complex Systems: 1695. https://igraph.org.

Csárdi, Gábor, Kuba Podgórski, and Rich Geldreich. 2021. *Zip: Cross-Platform Zip Compression*. https://github.com/r-lib/zip#readme.

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.

file., See AUTHORS. 2022. *Igraph: Network Analysis and Visualization*. https://CRAN.R-project.org/package=igraph.

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=pdftools.

Pedersen, Thomas Lin. 2021. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. https://CRAN.R-project.org/package=ggraph.

R Core Team. 2022. *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.

———. 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. 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/i

sbn/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/.

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