Corpus der Entscheidungen des Bundesarbeitsgerichts (CE-BAG)

Codebook

Version 2020-08-28

DOI: 10.5281/zenodo.4006645

Titel	Corpus der Entscheidungen des Bundesarbeitsgerichts
Abkürzung	CE-BAG
Autor	Sean Fobbe
Version	2020-08-28
Download	`https://doi.org/10.5281/zenodo.4006645`
Lizenz	CC0 1.0 Universal

Zitiervorschlag

Zitieren Sie diesen Datensatz bitte, wenn Sie ihn benutzen. Beispielsweise mit diesem Zitiervorschlag:

Fobbe, Sean, Corpus der Entscheidungen des Bundesarbeitsgerichts (CE-BAG), Version 2020-08-28, Zenodo, DOI: 10.5281/zenodo.4006645

Digital Object Identifiers: Concept DOI und Version DOI

Die in diesem Codebook verwendete DOI ist eine »concept DOI« und bezieht sich auf das Gesamtkonzept des Datensatzes. Sie verlinkt daher immer automatisch die aktuellste Version. Für jede spezifische Version steht eine weitere »version DOI« zur Verfügung, die auf der Zenodo-Seite der jeweiligen Version abgerufen werden kann. Diese Seiten sind auch über Links von der aktuellsten Version erreichbar.

Urheberrecht

Inhaltsverzeichnis

Inhaltsverzeichnis **3**

1 Einführung **4**

2 Datenquelle **4**

3 Hinweis zum Urheberrecht an den Rohdaten **5**

4 Nutzung mit R **5**

5 Metadaten **6**
5.1 Schema 6
5.2 Beispiel 6
5.3 Details zu den Variablen 7

6 Registerzeichen (Bundesarbeitsgericht) **8**

7 Zusätze bei Aktenzeichen (Bundesarbeitsgericht) **8**

8 Inhalt **9**
8.1 Nach Spruchkörper 9
8.2 Nach Registerzeichen 9
8.3 Nach Jahr 10
8.4 Nach ZusatzAZ 10

9 Disclaimer **11**

10 Changelog **11**

1 Einführung

Die quantitative Analyse von juristischen Texten ist in den deutschen Rechtswissenschaften ein noch junges und kaum bearbeitetes Feld. 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*[1] und *openJur*[2] verbieten ausdrücklich das maschinelle Auslesen der Rohdaten.[3] Wissenschaftliche Initiativen wie der Juristische Referenzkorpus (JuReKo) sind nach jahrelanger Arbeit hinter verschlossenen Türen verschwunden.

In einem funktionierenden Rechtsstaat muss die Rechtsprechung öffentlich, transparent und nachvollziehbar sein. Im 21. Jahrhundert bedeutet das, dass sie auch quantitativen Analysen zugänglich sein muss. Der Erstellung und Aufbereitung des Datensatzes liegen daher die Gedanken der Transparenz und wissenschaftlichen Reproduzierbarkeit zugrunde.

Zusätzlich zu den einfach maschinenlesbaren Formaten (TXT und CSV) sind die PDF-Rohdaten enthalten, damit Analysten gegebenenfalls ihre eigene Konvertierung vornehmen können. Die PDF-Rohdaten wurden inhaltlich nicht verändert und die Dateinamen nur geringfügig angepasst um die Lesbarkeit für Mensch und Maschine zu verbessern. Die TXT Dateien wurden mit **pdftotext 0.73.0** erstellt. Diese wurden mit **readtext 0.76** in **R 3.6.3** eingelesen und das CSV file mit dem Standard-Befehl **write.csv()** erstellt.

2 Datenquelle

Die Datenquelle ist: `https://www.bundesarbeitsgericht.de/`

Das Corpus der Entscheidungen des Bundesarbeitsgerichts (CE-BAG) enthält (fast) alle Entscheidungen die zum jeweiligen Stichtag (= Versionsnummer) auf der offizellen Webseite des Bundesarbeitsgerichts in dessen amtlicher Entscheidungsdatenbank abrufbar waren. Für 13 Entscheidungen (Version 2020-08-28) war keine PDF-Datei verfügbar, diese sind daher nicht enthalten.

Die Daten wurden unter Beachtung des Robot Exclusion Standard (RES) gesammelt. Die Entscheidungen sind nach Angaben des Gerichts anonymisiert, aber ungekürzt.

---

[1] `https://opinioiuris.de/`

[2] `https://openjur.de/`

[3] Openjur beabsichtigt eine API anzubieten, diese war aber im August 2020 immernoch nicht verfügbar. Openjur ist seit 2008 in Betrieb.

Nutzer sollten zwei wichtige Einschränkungen beachten:

1. Der Datensatz enthält nur das, was das Gericht auch tatsächlich veröffentlicht (*publication bias*)
2. Erst ab dem Jahr 2010 sind Daten vorhanden

3 Hinweis zum Urheberrecht an den Rohdaten

An den Entscheidungstexten und amtlichen Leitsätzen besteht gem. § 5 I 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) stelle ich gemäß einer *CC0 1.0 Universal Public Domain Lizenz* vollständig urheberrechtsfrei.

4 Nutzung mit R

Die **TXT** files inklusive Metadaten können zum Beispiel mit **R** und dem package **readtext** (auf CRAN verfügbar) folgendermaßen eingelesen werden:

```
txt.bverfg <- readtext("./CE-BAG/*.txt",
                       docvarsfrom = "filenames",
                       docvarnames = c("Gericht",
                                       "Datum",
                                       "SpruchkoerperAZ",
                                       "Registerzeichen",
                                       "Ordinalzahl",
                                       "Eingangsjahr",
                                       "ZusatzAZ"),
                       dvsep = "_",
                       encoding = "UTF-8")
```

Das **CSV** file ist mit **R** folgendermaßen einlesbar:

```
csv.bverfg <- read.csv("./Dateiname.csv", header = TRUE)
```

Selbstverständlich lassen sich die Daten auch mit anderen Programmiersprachen (z.B. Python) und Programmen nutzen.

5 Metadaten

Die Metadaten in den Dateinamen sind größtenteils unverändert von den Hyperlinks und dem Datenbankeintrag zur jeweiligen Datei übernommen. Hinzugefügt wurden von mir nur der Gerichtsname, sowie Unter- und Trennstriche um die Maschinenlesbarkeit zu erleichten.

Die Dateinamen enthalten Gerichtsname, Datum (nach ISO-8601, d.h. YYYY-MM-DD), sowie das offizielle Aktenzeichen (ggf. mit Zusatz).

Die Typen der Variablen wurden mit *regular expressions* strikt validiert. Die möglichen Werte der jeweiligen Variablen wurden zudem durch Frequenztabellen auf ihre Plausibilität geprüft.

5.1 Schema

`[Gericht]_[Datum]_[SpruchkoerperAZ]_[Registerzeichen]_[Ordinalzahl]_[Eingangsjahr]_[ZusatzAZ]`

5.2 Beispiel

`BAG_2020-06-03_3_AZR_255_20_F`

5.3 Details zu den Variablen

Variable	Typ	Erläuterung
Gericht	Alphabetisch	In diesem Datensatz ist nur der Wert »BAG« vergeben. Dies ist der ECLI-Code für »Bundesarbeitsgericht«. Diese Variable dient vor allem zur einfachen und transparenten Verbindung der Daten mit anderen Datensätzen.
Datum	Datum	Das Datum der Entscheidung im Format YYYY-MM-DD (ISO-8601).
SpruchkoerperAZ	Natürliche Zahl	Der im Aktenzeichen angegebene Spruchkörper. Es sind Werte von »1« bis »10« vergeben, die jeweils für den 1. bis 10. Senat stehen.
Registerzeichen	Alphabetisch	Das amtliche Registerzeichen. Es gibt die Verfahrensart an, in der die Entscheidung ergangen ist. Eine Erläuterung der Abkürzungen findet sich unter Punkt 6.
Ordinalzahl	Natürliche Zahl	Verfahren des gleichen Eingangsjahres erhalten vom Gericht eine Ordinalzahl in der Reihenfolge ihres Eingangs, um sie voneinander abzugrenzen.
Eingangsjahr	Natürliche Zahl	Das Jahr in dem das Verfahren beim Gericht anhängig wurde. Das Format ist eine zweistellige Jahreszahl (YY).
ZusatzAZ	Alphabetisch	Wenn ein bereits rechtskräftig abgeschlossenes Verfahren fortgesetzt wird, erhält es ein neues Aktenzeichen mit dem Zusatz »F« (meist Anhörungsrügen). Werden neben der verfahrensbeendenden Entscheidung bzw. dem verfahrensbeendenden Vergleich weitere dokumentationswürdige Entscheidungen veröffentlicht, sind diese durch einen Großbuchstaben (beginnend mit »A«) ausdifferenziert. Diese Zusätze sind jeweils Bestandteil des vom BAG vergebenen Aktenzeichens.

6 Registerzeichen (Bundesarbeitsgericht)

Register-zeichen	Erläuterung
ABN	Nichtzulassung der Rechtsbeschwerde, § 92a ArbGG
ABR	Rechtsbeschwerde, §§ 92 ff ArbGG
ABV	Erstinstanzliche Beschlussverfahren aufgrund SGB IX im Geschäftsbereich des Bundesnachrichtendienstes
ACA	Erstinstanzliche Urteilsverfahren aufgrund SGB IX im Geschäftsbereich des Bundesnachrichtendienstes
AS	Sonderverfahren
AZA	Prozesskostenhilfe außerhalb anhängiger Verfahren
AZB	Sofortige Beschwerde wegen verspäteter Absetzung des Berufungsurteils, § 72b ArbGG; Revisionsbeschwerde, § 77 ArbGG; Rechtsbeschwerde, § 78 ArbGG
AZM	Nichtzulassung der Revisionsbeschwerde, §§ 77 Satz 2 ArbGG
AZN	Nichtzulassung der Revision, § 72a ArbGG
AZR	Revisionen in Urteilsverfahren, § 72 ArbGG
E	Klagen gegen den Bund nach § 201 GVG iVm. § 9 Abs. 2 ArbGG
GS	Großer Senat

7 Zusätze bei Aktenzeichen (Bundesarbeitsgericht)

ZusatzAZ	Erläuterung
A, B etc.	Werden neben der verfahrensbeendenden Entscheidung bzw. dem verfahrensbeendenden Vergleich weitere dokumentationswürdige Entscheidungen veröffentlicht, sind diese durch einen Großbuchstaben (beginnend mit »A«) ausdifferenziert.
F	Wenn ein bereits rechtskräftig abgeschlossenes Verfahren fortgesetzt wird, erhält es ein neues Aktenzeichen mit dem Zusatz »F« (meist Anhörungsrügen).

8 Inhalt

8.1 Nach Spruchkörper

Senat	Entscheidungen	% Gesamt	% Kumulativ
1	544	9.67	9.67
2	478	8.5	18.17
3	803	14.28	32.44
4	761	13.53	45.97
5	577	10.26	56.23
6	541	9.62	65.85
7	512	9.1	74.95
8	402	7.15	82.1
9	446	7.93	90.03
10	561	9.97	100
<NA>	0	0	100
Total	5625	100	100

8.2 Nach Registerzeichen

Register-zeichen	Entscheidungen	% Gesamt	% Kumulativ
ABN	8	0.14	0.14
ABR	503	8.94	9.08
AS	11	0.2	9.28
AZA	4	0.07	9.35
AZB	98	1.74	11.09
AZM	4	0.07	11.16
AZN	87	1.55	12.71
AZR	4910	87.29	100
<NA>	0	0	100
Total	5625	100	100

8.3 Nach Jahr

Jahr	Entscheidungen	% Gesamt	% Kumulativ
2010	698	12.41	12.41
2011	603	10.72	23.13
2012	642	11.41	34.54
2013	593	10.54	45.08
2014	556	9.88	54.97
2015	550	9.78	64.75
2016	498	8.85	73.6
2017	540	9.6	83.2
2018	411	7.31	90.51
2019	428	7.61	98.12
2020	106	1.88	100
<NA>	0	0	100
Total	5625	100	100

8.4 Nach ZusatzAZ

ZusatzAZ	Entscheidungen	% Gesamt	% Kumulativ
A	58	1.03	1.03
B	3	0.05	1.08
F	49	0.87	1.96
<NA>	5515	98.04	100
Total	5625	100	100

9 Disclaimer

Dieser Datensatz ist eine private wissenschaftliche Initiative und steht nicht mit dem Bundesarbeitsgericht in Verbindung.

10 Changelog

Version	Details
2020-08-28	Erstveröffentlichung