# Corpus of Decisions

———————

# International Court of Justice

## (CD-ICJ)

CODEBOOK

Version 2023-05-07

| | |
|---|---|
| **Title** | Corpus of Decisions: International Court of Justice |
| **Abbreviation** | CD-ICJ |
| **Author** | Seán Fobbe |
| **Version** | 2023-05-07 |
| **Download** | https://doi.org/10.5281/zenodo.7876286 |
| **License** | CC0 1.0 Universal |

**Citation**

*Seán Fobbe* (2023). Corpus of Decisions: International Court of Justice (CD-ICJ). Version 2023-05-07. Zenodo. DOI: 10.5281/zenodo.7876286.

**Digital Object Identifiers: Concept DOI and Version DOI**

This data set is uniquely identified via the Digital Object Identifier (DOI) system. DOIs are persistent identifiers that are globally unique and can be resolved as a link by entering a DOI into the web service at www.doi.org. The DOI given in this document is a 'Version DOI', which uniquely identifies version 2023-05-07. Academics and others who wish to enable replication analyses are strongly advised to cite the *version DOI* and the precise version of the data used. A 'Concept DOI' is available from the page of the Zenodo record under the heading 'Cite all versions?' and will always resolve to the latest version.

**Public Domain Status**

The full data set and this document are distributed under a **Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication** license. The person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law.

You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. In no way are the patent or trademark rights of any person affected by CC0, nor are the rights that other persons may have in the work or in how the work is used, such as publicity or privacy rights. Unless expressly stated otherwise, the person who associated a work with this deed makes no warranties about the work, and disclaims liability for all uses of the work, to the fullest extent permitted by applicable law.

Please see <https://creativecommons.org/publicdomain/zero/1.0/legalcode> for the full terms of the license.

**Disclaimer**

This data set is a personal academic initiative and is not associated with or endorsed by the International Court of Justice or the United Nations.

# Contents

# 1 Introduction

The **International Court of Justice (ICJ)** is the primary judicial organ of the United Nations and one of the most consequential courts in international law.

Called the 'World Court' by many, it is the only international court with general thematic jurisdiction. While critics occasionally note the lack of compulsory jurisdiction and sharply limited access to the Court,[1] its opinions continue to have an outsize influence on the modern interpretation, codification and wider development of international law. Every international legal textbook covers the workings and decisions of the Court *in extenso* and participation in international moot courts, such as the Philip C. Jessup Moot Court, without regular reference to and citation of the International Court of Justice's decisions, is unthinkable.

The **Corpus of Decisions: International Court of Justice (CD-ICJ)** collects and presents for the first time in human- and machine-readable form all published decisions of the International Court of Justice. Among these are judgments, advisory opinions and orders, as well as their respective appended minority opinions (declarations, separate opinions and dissenting opinions).

This data set is designed to be complementary to and fully compatible with the *Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)*, which is also available open access.[2]

The quantitative analysis of international legal data is still in its infancy, a situation which is exacerbated by the lack of high-quality empirical data. Most advanced data sets are held in commercial databases and are therefore not easily available to academic researchers, journalists and the general public. With this data set I hope to contribute to a more systematic and empirical view of the international legal system. In an international community founded on the rule of law the activities of the judiciary must be public, transparent and defensible. In the 21st century this requires quantitative scientific review of decisions and actions.

Design, construction and compilation of this data set are based on the principles of general availability through freedom from copyright (public domain status), strict transparency and full scientific reproducibility. The *FAIR Guiding Principles for Scientific Data Management and Stewardship* (Findable, Accessible, Interoperable and Reusable) inspire both the design and the manner of publication.[3]

---

[1]   Only States may be party to proceedings in contentious jurisdiction and only certain bodies of international organizations may request advisory opinions.

[2]   Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ). <https://doi.org/10.5281/zenodo.3840480>.

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

# 2 Reading Files

The data are published in open, interoperable and widely used formats (CSV, TXT, PDF). They can be used with all modern programming languages (e.g. Python or R) and graphical interfaces. The PDF collections are intended to facilitate traditional legal research.

**Important:** Missing values are always coded as 'NA'.

## 2.1 CSV Files

Working with the CSV files is recommended. CSV[4] is an open and simple machine-readable tabular data format. In this data set values are separated by commas. Each column is a variable and each row is a document. Variables are explained in detail in section 5.

To read **CSV** files into R I strongly recommend using the fast file reader **fread()** from the **data.table** package (available on CRAN). The file can be read into **R** like so:

```r
library(data.table)
icj.en <- fread("filename.csv")
```

## 2.2 TXT Files

The **TXT** files, including metadata, can be read into **R** with the package **readtext** (available on CRAN) thus:

```r
library(readtext)
icj.en <- readtext("EN_TXT_BEST_FULL/*.txt",
                   docvarsfrom = "filenames",
                   docvarnames = c("court",
                                   "caseno",
                                   "shortname",
                                   "applicant",
                                   "respondent",
                                   "date",
                                   "doctype",
                                   "collision",
                                   "opinion",
                                   "language"),
                   dvsep = "_",
                   encoding = "UTF-8")
```

---

[4]    The CSV format is defined in RFC 4180: <https://tools.ietf.org/html/rfc4180>.

# 3 Data Set Design

## 3.1 Description of Data Set

The **Corpus of Decisions: International Court of Justice (CD-ICJ)** collects and structures in human- and machine-readable form all published decisions of the International Court of Justice. Among these are judgments, advisory opinions and orders, as well as their respective appended minority opinions (declarations, separate opinions and dissenting opinions).

It consists of a CSV file of the full data set, a CSV file with the metadata only, individual TXT files for each document and PDF files with an enhanced text layer generated by the LSTM neural network engine of the optical character recognition software (OCR) *Tesseract*.

Additionally, the raw PDF files and some intermediate stages of refinement are included to allow for easier replication of results and for production use in the event that even higher quality methods of optical character recognition (OCR) can be applied to the documents in the future.

## 3.2 Complementarity

This data set is intended to be complementary to and fully compatible with the *Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)*, which is also available open access.[5]

## 3.3 Table of Sources

| Data Source | Citation |
|---|---|
| Primary Data Source | https://www.icj-cij.org |
| Source Code | https://doi.org/10.5281/zenodo.7876287 |
| Country Codes | https://doi.org/10.5281/zenodo.7876287 |
| Entity Codes | https://doi.org/10.5281/zenodo.7876287 |
| Cases Names and Parties | https://doi.org/10.5281/zenodo.7876287 |

## 3.4 Data Collection

Data were collected with the explicit consent of the Registry of the International Court of Justice. All documents were downloaded via TLS-encrypted connections and cryptographically signed after data processing was complete. The data set collects all decisions and appended opinions issued by the International Court of Justice that were published on the official website of the International Court of Justice on 2023-05-07.

---

5    Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ). <https://doi.org/10.5281/zenodo.3840480>.

## 3.5 Source Code and Compilation Report

The full Source Code for the creation of this data set, the resulting Compilation Report and this Codebook are published open access and permanently archived in the scientific repository of CERN.

With every compilation of the full data set an extensive **Compilation Report** is created in a professionally layouted PDF format (comparable to this Codebook). The Compilation Report includes the Source Code, comments and explanations of design decisions, relevant computational results, exact timestamps and a table of contents with clickable internal hyperlinks to each section. The Compilation Report is published under the same DOI as the Source Code.

For details of the construction and validation of the data set please refer to the Compilation Report.

## 3.6 Limitations

Users should bear in mind certain limitations:

1. The data set contains only those documents which were published by the ICJ and have been made available by the ICJ on its official website (*publication bias*).

2. While Tesseract yields high-quality OCR results, current OCR technology is not perfect and minor errors must be expected (*OCR bias*).

3. Automatic language detection is not foolproof and some bilingual documents marked as monolingual may have gone undetected (*language mismatch*).

4. Lengthy quotations in languages other than the language indicated in the metadata may further confound analyses (*language blurring*).

## 3.7 Public Domain Status

According to written communication between the author and the Registry of the International Court of Justice the original documents are not subject to copyright.

To ensure the widest possible distribution and to promote the international rule of law I waive any copyright to the data set under a **Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication**. For details of the license please refer to the CC0 copyright notice at the beginning of this Codebook or visit the Creative Commons website for the full terms of the license.[6]

---

[6]    https://creativecommons.org/publicdomain/zero/1.0/legalcode

## 3.8 Quality Assurance

Dozens of automated tests were conducted to ensure the quality of the data and metadata, for example:

1. Auto-detection of language via analysis of n-gram patterns with the *textcat* package for R.

2. Strict validation of variable types via *regular expressions*.

3. Construction of frequency tables for (almost) every variable followed by human review to detect anomalies.

4. Creation of visualizations for many common descriptive analyses.

For results of each test and more information on the construction of the data set please refer to the Compilation Report or the 'ANALYSIS' archive included with the data set.

Action Nodes
Data Nodes

www.icj-cij.org

Extract Links from HTML

Links to Raw PDF Files

Detect Unlabelled Files

Download Unlabelled Files

Unlabelled Files

Handcoding of Labels

Labelling Information

Apply Labelling

Strict REGEX Validation: ICJ File Name Schema

Download Module

Labelled PDF Files

File Split Module

Handcoded Case Names

Filename Enhancement Module

Strict REGEX Validation: Codebook File Name Schema

FR_PDF_ORIGINAL_FULL

EN_PDF_ORIGINAL_FULL

Detect Missing Language Counterparts

Tesseract OCR Module

Text Extraction Module

FR_TXT_TESSERACT_max2004

EN_TXT_TESSERACT_max2004

FR_PDF_ENHANCED_Max2004

FR_TXT_EXTRACTED

EN_TXT_EXTRACTED

EN_PDF_ENHANCED_Max2004

FR_TXT_BEST

EN_TXT_BEST

FR_PDF_BEST_FULL

EN_PDF_BEST_FULL

Create Majority Variant

FR_PDF_BEST_MajorityOpinions

EN_PDF_BEST_MajorityOpinions

Figure 1: Workflow Schematic Part 1: Download, Labelling, Conversion and Sorting of Documents
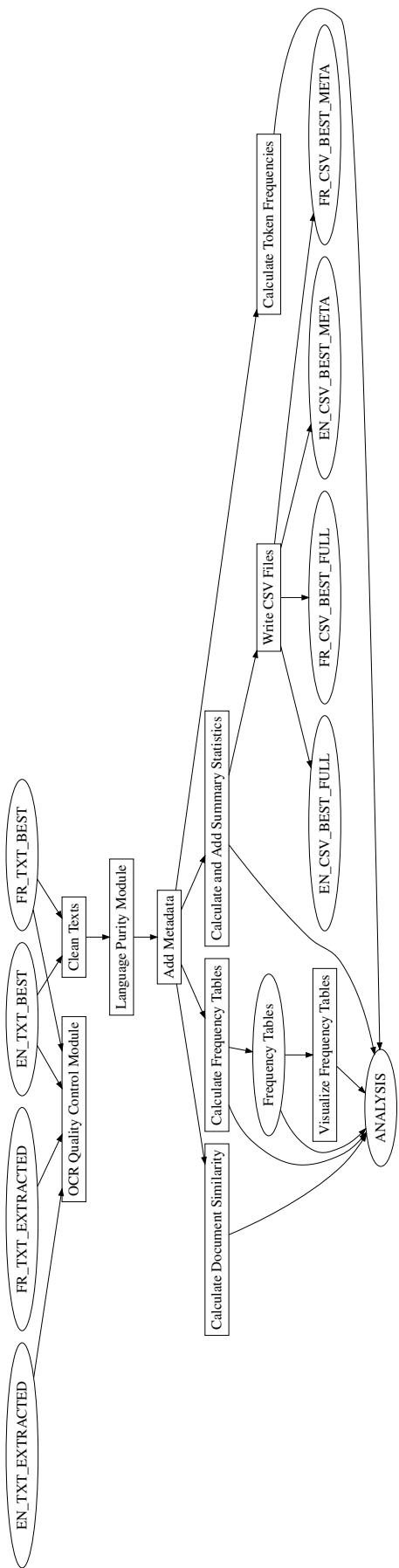
Figure 2: Workflow Schematic Part 2: Ingestion, Pre-Processing, Analysis and Creation of CSV Files

# 4  Variants and Primary Target Audiences

The data set is provided in two language versions (English and French), as well as several differently processed variants geared towards specific target audiences.

A reduced PDF variant of the data set containing only majority opinions is intended to assist practitioners.

| Variant | Target Audience and Description |
| --- | --- |
| PDF_BEST | **Traditional Legal Research (recommended).** A synthesis of all born-digital documents issued by the ICJ combined with older scanned documents (prior to 2005) which were given a new and enhanced text layer created with an advanced LSTM neural network machine learning engine. Its main advantages are vastly improved local searches in individual documents via Ctrl+F and copy/pasting without the need for extensive manual revisions. Researchers with slow internet connections should consider using the 'TXT_BEST' variant, as this still provides a reasonable visual approximation of the original documents, but offers the advantage of drastically reduced file size. A reduced PDF variant of the data set containing only majority opinions is available to assist practitioners. |
| CSV_BEST | **Quantitative Research (recommended).** A structured representation of the full data set within a single comma-delimited file. Includes the full complement of metadata described in the Codebook. The 'FULL' sub-variant includes the full text of the decisions, whereas the sub-variant 'META' only contains the metadata. |
| TXT_BEST | **Quantitative Research.** A synthesis of TXT files created by combining the extracted text of all born-digital documents issued by the ICJ (2005 and later) and the OCR texts from older scanned documents (prior to 2005) generated with an advanced LSTM neural network machine learning engine. R users should strongly consider using the package *readtext* to read them into R with the filename metadata intact. |
| ANALYSIS | **Quantitative Research.** This archive contains almost all of the machine-readable analysis output generated during the data set creation process to facilitate further analysis (CSV for tables, PDF and PNG for plots). Minor analysis results are documented only in the Compilation Report. |

| Variant | Target Audience and Description |
|---|---|
| TXT_EXTRACTED | **Replication Research and Creation of New Data Sets.** TXT files containing the extracted text layer from all original documents as published by the ICJ. The quality of the original OCR text for older documents is poor and this variant should not be used for statistical analysis. Documents dated 2005 or later were born-digital and can be used for all purposes. |
| TXT_TESSERACT | **Replication Research and Creation of New Data Sets.** TXT files containing the OCR text generated with an advanced LSTM neural network machine learning engine for documents predating 2005. Fully included in the BEST variant, but provided separately for reasons of transparency. |
| PDF_ORIGINAL | **Replication Research and Creation of New Data Sets.** The original documents with the original text layer. Only recommended for researchers who wish to replicate the machine-readable files or who wish to create a new and improved data set. Not recommended for traditional research, as the quality of the original OCR text layer is quite poor. |
| PDF_ENHANCED | **Replication Research and Creation of New Data Sets.** Scanned documents of opinions rendered before 2005 which were given a new and enhanced text layer generated with an advanced LSTM neural network machine learning engine. Fully included in the BEST variant, but provided separately for reasons of transparency. |

# 5 Variables

## 5.1 General Remarks

- Missing values are always coded as 'NA'.

- All Strings are encoded in UTF-8.

- A significant part of the metadata was included with the files downloaded from the Court's website.

- The variables 'shortname', 'applicant', 'respondent', 'stage', 'applicant_region', 'applicant_subregion', 'respondent_region' and 'respondent_subregion' were coded manually by the author of the data set and added automatically at compilation time. Country codes conform to the ISO 3166 Alpha-3 standard and geographical classifications to the M49 standard used by the UN Statistics Division.

- The variable 'fullname' is coded according to case headings as published on the ICJ website. Includes the full names of the parties in parentheses. Introductory phrases such as 'Case concerning...' are omitted.

- The variables 'nchars', 'ntokens', 'ntypes', 'nsentences' and 'year' were calculated automatically based on the content and metadata of each document.

- The variables 'version', 'doi_concept', 'doi_version' and 'license' were added automatically during the data set creation process to document provenance and to comply with FAIR Data Principles F1, F3 and R1.1.

## 5.2 Structure of TXT File Names

```
[court]_[caseno]_[shortname]_[applicant]_[respondent]_[date]_[doctype]_
[collision]_[stage]_[opinion]_[language]
```

## 5.3 Example TXT File Name

```
ICJ_001_CorfuChannel_GBR_ALB_1949-04-09_JUD_01_ME_05_EN.txt
```

## 5.4 Structure of CSV Metadata

```
## Classes 'data.table' and 'data.frame':   2264 obs. of  27 variables:
##  $ doc_id             : chr  "ICJ_001_CorfuChannel_GBR_ALB_1947-07-31_ORD_01_
   NA_00_EN.txt" "ICJ_001_CorfuChannel_GBR_ALB_1947-12-10_ORD_01_NA_00_EN.txt" "
   ICJ_001_CorfuChannel_GBR_ALB_1948-03-25_JUD_01_PO_00_EN.txt" "ICJ_001_
   CorfuChannel_GBR_ALB_1948-03-25_JUD_01_PO_01_EN.txt" ...
##  $ court              : chr  "ICJ" "ICJ" "ICJ" "ICJ" ...
##  $ caseno             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ shortname          : chr  "CorfuChannel" "CorfuChannel" "CorfuChannel" "
   CorfuChannel" ...
##  $ fullname           : chr  "Corfu Channel (United Kingdom of Great Britain
   and Northern Ireland v. Albania)" "Corfu Channel (United Kingdom of Great
   Britain and Northern Ireland v. Albania)" "Corfu Channel (United Kingdom of
   Great Britain and Northern Ireland v. Albania)" "Corfu Channel (United
   Kingdom of Great Britain and Northern Ireland v. Albania)" ...
##  $ applicant          : chr  "GBR" "GBR" "GBR" "GBR" ...
##  $ respondent         : chr  "ALB" "ALB" "ALB" "ALB" ...
##  $ applicant_region   : chr  "Europe" "Europe" "Europe" "Europe" ...
##  $ respondent_region  : chr  "Europe" "Europe" "Europe" "Europe" ...
##  $ applicant_subregion : chr  "Northern Europe" "Northern Europe" "Northern
   Europe" "Northern Europe" ...
##  $ respondent_subregion: chr  "Southern Europe" "Southern Europe" "Southern
   Europe" "Southern Europe" ...
##  $ date               : IDate, format: "1947-07-31" "1947-12-10" ...
##  $ doctype            : chr  "ORD" "ORD" "JUD" "JUD" ...
##  $ collision          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ stage              : chr  NA NA "PO" "PO" ...
##  $ opinion            : int  0 0 0 1 2 0 0 0 1 2 ...
##  $ language           : chr  "EN" "EN" "EN" "EN" ...
##  $ year               : int  1947 1947 1948 1948 1948 1948 1948 1949 1949
   1949 ...
##  $ minority           : int  0 0 0 1 1 0 0 0 1 1 ...
##  $ nchars             : int  5782 2435 48544 3883 34300 6148 7846 172995
   25052 23145 ...
##  $ ntokens            : int  1141 464 9046 744 6633 1169 1602 33912 4750 4310
    ...
##  $ ntypes             : int  368 175 1353 284 1037 373 516 3674 1037 1026 ...
##  $ nsentences         : int  45 12 272 28 209 32 53 1254 161 118 ...
##  $ version            : IDate, format: "2023-05-07" "2023-05-07" ...
##  $ doi_concept        : chr  "10.5281/zenodo.3826444" "10.5281/zenodo
   .3826444" "10.5281/zenodo.3826444" "10.5281/zenodo.3826444" ...
##  $ doi_version        : chr  "10.5281/zenodo.7876286" "10.5281/zenodo
   .7876286" "10.5281/zenodo.7876286" "10.5281/zenodo.7876286" ...
##  $ license            : chr  "Creative Commons Zero 1.0 Universal" "Creative
   Commons Zero 1.0 Universal" "Creative Commons Zero 1.0 Universal" "Creative
   Commons Zero 1.0 Universal" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

## 5.5   Detailed Description of Variables

| Variable | Type | Details |
|---|---|---|
| doc_id | String | (CSV only) The name of the imported TXT file. |
| text | String | (CSV only) The full content of the imported TXT file. |
| court | String | The variable only takes the value 'ICJ', which stands for 'International Court of Justice'. It is generally only useful if combined with the CD-PCIJ or other data sets. |
| caseno | Integer | The case number assigned by the ICJ. The same case may span multiple case numbers, i.e. the Interpretation or Revision stages have different case numbers than the original judgment. To analyze all stages of a case I recommend a pattern search on the variable 'shortname'. Note: case number 2 is unassigned and there are no documents for case number 2 available on the ICJ website. |
| shortname | String | Short name of the case. This was custom-created by the author based on the original title. Short names include well-known components (e.g. 'Nicaragua') to facilitate quick local searches and try to be as faithful to the full title as possible. For requests concerning interpretation or revision of a judgment the shortname is followed by 'Interpretation' or 'Revision'. |
| fullname | String | (CSV only) Full name of the case as published on the ICJ website. Includes the full names of the Parties. Introductory phrases such as 'Case concerning...' are omitted. |
| applicant | String | The unique identifier of the applicant. In contentious proceedings this is the three-letter (Alpha-3) country code as per the ISO 3166-1 standard. Table 6.1 contains an explanation of all country codes used in the data set. Please note that reserved country codes are in use for historical entities (e.g. the Soviet Union). For advisory proceedings this variable refers to the entity which requested an advisory opinion. Table 6.2 explains the detailed advisory coding decisions. |

| Variable | Type | Details |
|---|---|---|
| respondent | String | The unique identifier of the respondent. In contentious proceedings this is the three-letter (Alpha-3) country code as per the ISO 3166-1 standard. Table 6.1 contains an explanation of all country codes used in the data set. Please note that reserved country codes are in use for historical entities (e.g. the Soviet Union). Advisory proceedings do not have a respondent and therefore always take the value 'NA'. |
| applicant_region | String | (CSV only) The geographical region of the applicant according to the UN M49 standard. Please refer to table 6.1 for details and exceptions. Geographical information is only available for countries, not for UN bodies or international organizations. |
| respondent_region | String | (CSV only) The geographical region of the respondent according to the UN M49 standard. Please refer to table 6.1 for details and exceptions. Geographical information is only available for countries, not for UN bodies or international organizations. |
| applicant_subregion | String | (CSV only) The geographical subregion of the applicant according to the UN M49 standard. Please refer to table 6.1 for details and exceptions. Geographical information is only available for countries, not for UN bodies or international organizations. |
| respondent_subregion | String | (CSV only) The geographical subregion of the respondent according to the UN M49 standard. Please refer to table 6.1 for details and exceptions. Geographical information is only available for countries, not for UN bodies or international organizations. |
| date | ISO Date | The date of the document in the format YYYY-MM-DD (ISO-8601). |
| doctype | String | A three-letter code indicating the type of document. Possible values are 'JUD' (judgments in contentious jurisdiction), 'ADV' (advisory opinions) and 'ORD' (orders in all types of jurisidiction). |
| collision | Integer | In rare instances the International Court of Justice issued several decisions of the same type in the same proceedings on the same day. Most documents take the value '01'. If documents with otherwise identical metadata would be issued, the value is incremented. |

| Variable | Type | Details |
| --- | --- | --- |
| stage | String | The stage of proceedings in contentious jurisdiction, coded based on the title page (primary), or a close reading of the findings (secondary). Possible values are 'PO' (preliminary objections), 'ME' (merits), 'IN' (intervention) and 'CO' (compensation). Please note that the ICJ is very inconsistent in how it classifies admissibility; it can occur in the same document either together with a decision on jurisdiction or a decision on the merits. I have chosen to code pure admissibility decisions as 'ME' (e.g. Second Phase of Nottebohm). In general all of the above types of decisions can occur in the same document. I therefore do not recommend this variable for computational analysis unless great care is taken to understand its limitations. Currently only judgments are coded, orders will be added in the future. |
| opinion | Integer | A sequential number assigned to each opinion. Majority opinions are always coded '00'. Minority opinions begin with '01' and ascend to the maximum number of minority opinions. |
| language | String | The language of the document as a two-letter ISO 639-1 code. This data set contains documents in the languages English ('EN') and French ('FR'). |
| year | Integer | (CSV only) The year the document was issued. The format is YYYY. |
| minority | Integer | (CSV only) This variable indicates whether the document is a majority (0) or minority (1) opinion. |
| nchars | Integer | (CSV only) The number of characters in a given document. |
| ntokens | Integer | (CSV only) The number of tokens (an arbitrary character sequence bounded by whitespace) in a given document. This metric can vary significantly depending on tokenizer and parameters used. This count was generated based on plain tokenization with no further pre-processing (e.g. stopword removal, removal of numbers, lowercasing) applied. Analysts should use this number not as an exact figure, but as an estimate of the order of magnitude of a given document's length. If in doubt, perform an independent calculation with the software of your choice. |

| Variable | Type | Details |
|---|---|---|
| ntypes | Integer | (CSV only) The number of *unique* tokens. This metric can vary significantly depending on tokenizer and parameters used. This count was generated based on plain tokenization with no further pre-processing (e.g. stopword removal, removal of numbers, lowercasing) applied. Analysts should use this number not as an exact figure, but as an estimate of the order of magnitude of a given document's length. If in doubt, perform an independent calculation with the software of your choice. |
| nsentences | Integer | (CSV only) The number of sentences in a given document. The rules for detecting sentence boundaries are very complex and are described in 'Unicode Standard Annex No 29'. This metric can vary significantly depending on tokenizer and parameters used. This count was generated based on plain tokenization with no further pre-processing (e.g. stopword removal, removal of numbers, lowercasing) applied. Analysts should use this number not as an exact figure, but as an estimate of the order of magnitude of a given document's length. If in doubt, perform an independent calculation with the software of your choice. |
| version | ISO Date | (CSV only) The version of the data set as a date in long form as per ISO-8601. The version represents the date on which the data set creation process was begun and the data was acquired from the website of the Court. |
| doi_concept | String | (CSV only) The Digital Object Identifier (DOI) for the *concept* of the data set. Resolving this DOI via www.doi.org allows researchers to always acquire the *latest version* of the data set. The DOI is a persistent identifier suitable for stable long-term citation. Principle F1 of the FAIR Data Principles ('data are assigned globally unique and persistent identifiers') recommends the documentation of each data set with a persistent identifier and Principle F3 its inclusion with the meta-data. Even if the CSV data set is transmitted without the accompanying Codebook this allows researchers to establish provenance of the data. |

| Variable | Type | Details |
|---|---|---|
| doi_version | String | (CSV only) The Digital Object Identifier (DOI) for the *specific version* of the data set. Resolving this DOI via www.doi.org allows researchers to always acquire this *specific version* of the data set. The DOI is a persistent identifier suitable for stable long-term citation. Principle F1 of the FAIR Data Principles ('data are assigned globally unique and persistent identifiers') recommends the documentation of each data set with a persistent identifier and Principle F3 its inclusion with the metadata. Even if the CSV data set is transmitted without the accompanying Codebook this allows researchers to establish provenance of the data. |
| license | String | (CSV only) The license of the data set. In this data set the value is always 'Creative Commons Zero 1.0 Universal'. Ensures compliance with FAIR data principle R1.1 ('clear and accessible data usage license'). |

# 6 Applicant and Respondent Codes

## 6.1 Contentious Jurisdiction: States

Applicants and Respondents in contentious jurisdiction are coded according to the uppercase three-letter (Alpha-3) country codes described in the ISO 3166-1 standard. The codes are taken from the version of the standard which was valid on 4 November 2020. The table below only includes those codes which are used in the data set. The regions and subregions assigned to States generally follow the UN Standard Country or Area Codes for Statistics Use, 1999 (Revision 4), also known as the M49 standard.

Please note that where States have ceased to exist (Soviet Union, Yugoslavia, Serbia and Montenegro, Czechoslovakia) their historical three-letter country codes from ISO 3166-1 are used. These are not part of the current ISO 3166-1 standard, but have been transitionally reserved by the ISO 3166 Maintenance Agency to ensure backwards compatibility. The four-letter ISO 3166-3 standard ('Code for formerly used names of countries') is not used in this data set. The regions and subregions for Yugoslavia and Czechoslovakia are taken from M49 revision 2 (1982). The Soviet Union is coded as 'Europe/Eastern Europe' (the M49 standard considers the SUN its own region). Serbia and Montenegro was never included in the M49 standard and has been assigned the same region and subregion as Yugoslavia.

| ISO-3 | Name | Region | Sub-Region |
|-------|------|--------|------------|
| ALB | Albania | Europe | Southern Europe |
| ARE | United Arab Emirates | Asia | Western Asia |
| ARG | Argentina | Americas | Latin America and the Caribbean |
| ARM | Armenia | Asia | Western Asia |
| AUS | Australia | Oceania | Australia and New Zealand |
| AZE | Azerbaijan | Asia | Western Asia |
| BDI | Burundi | Africa | Sub-Saharan Africa |
| BEL | Belgium | Europe | Western Europe |
| BEN | Benin | Africa | Sub-Saharan Africa |
| BFA | Burkina Faso | Africa | Sub-Saharan Africa |
| BGR | Bulgaria | Europe | Eastern Europe |
| BHR | Bahrain | Asia | Western Asia |
| BIH | Bosnia and Herzegovina | Europe | Southern Europe |
| BLZ | Belize | Americas | Latin America and the Caribbean |
| BOL | Bolivia | Americas | Latin America and the Caribbean |
| BRA | Brazil | Americas | Latin America and the Caribbean |
| BWA | Botswana | Africa | Sub-Saharan Africa |
| CAN | Canada | Americas | Northern America |
| CHE | Switzerland | Europe | Western Europe |
| CHL | Chile | Americas | Latin America and the Caribbean |
| CMR | Cameroon | Africa | Sub-Saharan Africa |

*(continued)*

| ISO-3 | Name | Region | Sub-Region |
|-------|------|--------|------------|
| COD | Democratic Republic of the Congo | Africa | Sub-Saharan Africa |
| COL | Colombia | Americas | Latin America and the Caribbean |
| CRI | Costa Rica | Americas | Latin America and the Caribbean |
| CSK | Czechoslovakia | Europe | Eastern Europe |
| DEU | Germany | Europe | Western Europe |
| DJI | Djibouti | Africa | Sub-Saharan Africa |
| DMA | Dominica | Americas | Latin America and the Caribbean |
| DNK | Denmark | Europe | Northern Europe |
| ECU | Ecuador | Americas | Latin America and the Caribbean |
| EGY | Egypt | Africa | Northern Africa |
| ESP | Spain | Europe | Southern Europe |
| ETH | Ethiopia | Africa | Sub-Saharan Africa |
| FIN | Finland | Europe | Northern Europe |
| FRA | France | Europe | Western Europe |
| GAB | Gabon | Africa | Sub-Saharan Africa |
| GBR | United Kingdom | Europe | Northern Europe |
| GEO | Georgia | Asia | Western Asia |
| GIN | Guinea | Africa | Sub-Saharan Africa |
| GMB | Gambia | Africa | Sub-Saharan Africa |
| GNB | Guinea-Bissau | Africa | Sub-Saharan Africa |
| GNQ | Equatorial Guinea | Africa | Sub-Saharan Africa |
| GRC | Greece | Europe | Southern Europe |
| GTM | Guatemala | Americas | Latin America and the Caribbean |
| GUY | Guyana | Americas | Latin America and the Caribbean |
| HND | Honduras | Americas | Latin America and the Caribbean |
| HRV | Croatia | Europe | Southern Europe |
| HUN | Hungary | Europe | Eastern Europe |
| IDN | Indonesia | Asia | South-eastern Asia |
| IND | India | Asia | Southern Asia |
| IRN | Iran | Asia | Southern Asia |
| ISL | Iceland | Europe | Northern Europe |
| ISR | Israel | Asia | Western Asia |
| ITA | Italy | Europe | Southern Europe |
| JPN | Japan | Asia | Eastern Asia |
| KEN | Kenia | Africa | Sub-Saharan Africa |

| ISO-3 | Name | Region | Sub-Region |
|-------|------|--------|------------|
| KHM | Cambodia | Asia | South-eastern Asia |
| LBN | Lebanon | Asia | Western Asia |
| LBR | Liberia | Africa | Sub-Saharan Africa |
| LBY | Libya | Africa | Northern Africa |
| LIE | Liechtenstein | Europe | Western Europe |
| MEX | Mexico | Americas | Latin America and the Caribbean |
| MHL | Marshall Islands | Oceania | Micronesia |
| MKD | North Macedonia | Europe | Southern Europe |
| MLI | Mali | Africa | Sub-Saharan Africa |
| MLT | Malta | Europe | Southern Europe |
| MMR | Myanmar | Asia | South-eastern Asia |
| MYS | Malaysia | Asia | South-eastern Asia |
| NAM | Namibia | Africa | Sub-Saharan Africa |
| NER | Niger | Africa | Sub-Saharan Africa |
| NGA | Nigeria | Africa | Sub-Saharan Africa |
| NIC | Nicaragua | Americas | Latin America and the Caribbean |
| NLD | Netherlands | Europe | Western Europe |
| NOR | Norway | Europe | Northern Europe |
| NRU | Nauru | Oceania | Micronesia |
| NZL | New Zealand | Oceania | Australia and New Zealand |
| PAK | Pakistan | Asia | Southern Asia |
| PER | Peru | Americas | Latin America and the Caribbean |
| PRT | Portugal | Europe | Southern Europe |
| PRY | Paraguay | Americas | Latin America and the Caribbean |
| PSE | Palestine | Asia | Western Asia |
| QAT | Qatar | Asia | Western Asia |
| ROU | Romania | Europe | Eastern Europe |
| RUS | Russia | Europe | Eastern Europe |
| RWA | Rwanda | Africa | Sub-Saharan Africa |
| SAU | Saudi-Arabia | Asia | Western Asia |
| SCG | Serbia and Montenegro | Europe | Southern Europe |
| SEN | Senegal | Africa | Sub-Saharan Africa |
| SGP | Singapore | Asia | South-eastern Asia |
| SLV | El Salvador | Americas | Latin America and the Caribbean |
| SOM | Somalia | Africa | Sub-Saharan Africa |
| SRB | Serbia | Europe | Southern Europe |

*(continued)*

| ISO-3 | Name | Region | Sub-Region |
|-------|------|--------|------------|
| SUN | Soviet Union | Europe | Eastern Europe |
| SVK | Slovakia | Europe | Eastern Europe |
| SWE | Sweden | Europe | Northern Europe |
| TCD | Chad | Africa | Sub-Saharan Africa |
| THA | Thailand | Asia | South-eastern Asia |
| TLS | Timor Leste | Asia | South-eastern Asia |
| TUN | Tunisia | Africa | Northern Africa |
| TUR | Turkey | Asia | Western Asia |
| UGA | Uganda | Africa | Sub-Saharan Africa |
| UKR | Ukraine | Europe | Eastern Europe |
| URY | Uruguay | Americas | Latin America and the Caribbean |
| USA | United States of America | Americas | Northern America |
| VEN | Venezuela | Americas | Latin America and the Caribbean |
| YUG | Yugoslavia | Europe | Southern Europe |
| ZAF | South Africa | Africa | Sub-Saharan Africa |

## 6.2 Advisory Jurisdiction: Entities

Entities who requested an advisory opinion from the International Court of Justice are not Applicants in the strict sense, but have been coded under this variable to reduce clutter. I have tried to choose widely used codes for each entity.

Note that the *International Maritime Organization (IMO)* was known as the 'Inter-Governmental Maritime Consultative Organization' at the time it requested the advisory opinon. I have coded it with the modern 'IMO', as the organization only underwent a change of name and its legal continuity is not in doubt.

I was unable to discover a well-known acronym for the *Committee on Applications for Review of Administrative Tribunal Judgements* and custom-coded it as 'CARAT'.

| Code | Entity |
| --- | --- |
| CARAT | Committee on Applications for Review of Administrative Tribunal Judgements |
| ECOSOC | UN Economic and Social Council |
| IFAD | International Fund for Agricultural Development |
| IMO | Inter-Governmental Maritime Consultative Organization |
| UNECO | United Nations Educational, Scientific and Cultural Organization |
| UNGA | UN General Assembly |
| UNSC | United Nations Security Council |
| WHO | World Health Organization |

# 7 Linguistic Metrics

## 7.1 Explanation of Metrics

To better communicate the scope of the corpus and its constituent documents I provide a number of classic linguistic metrics and visualize their distributions:

| Metric | Definition |
|---|---|
| Characters | Characters roughly correspond to graphemes, the smallest functional unit in a writing system. The word 'judge' is composed of 5 characters, for example. |
| Tokens | An arbitrary character sequence delimited by whitespace on both sides, e.g. it roughly corresponds to the notion of a 'word'. However, due to its strictly syntactical definition it might also include arbitrary sequences of numbers or special characters. |
| Types | Unique tokens. If, for example, the token 'human' appeared one hundred times in a given document, it would be counted as only one type. |
| Sentences | Corresponds approximately to the colloquial definition of a sentence. The exact rules for determining sentence boundaries are very complex and may be reviewed in 'Unicode Standard: Annex No 29'. |

## 7.2 Summary Statistics

### 7.2.1 English

| Metric | Total | Min | Quart1 | Median | Mean | Quart3 | Max |
|---|---|---|---|---|---|---|---|
| nchars | 88,204,775 | 373 | 4,490.75 | 16,031.0 | 38,959.71 | 44,053.50 | 744,487 |
| ntokens | 15,625,026 | 71 | 761.00 | 2,843.5 | 6,901.51 | 8,043.25 | 142,528 |
| ntypes | 89,108 | 53 | 289.00 | 704.5 | 1,040.32 | 1,392.50 | 9,950 |
| nsentences | 533,680 | 1 | 20.00 | 94.0 | 235.72 | 265.00 | 5,645 |

### 7.2.2 French

| Metric | Total | Min | Quart1 | Median | Mean | Quart3 | Max |
|---|---|---|---|---|---|---|---|
| nchars | 93,498,997 | 396 | 4,572 | 16,898.0 | 41,965.44 | 47,858.00 | 817,862 |
| ntokens | 16,015,389 | 69 | 784 | 2,933.5 | 7,188.24 | 8,412.25 | 148,455 |
| ntypes | 109,626 | 55 | 317 | 825.5 | 1,231.01 | 1,669.50 | 11,867 |
| nsentences | 526,959 | 1 | 23 | 94.5 | 236.52 | 265.50 | 5,577 |

### 7.3   Explanation of Diagrams

#### 7.3.1   Distributions of Document Length

The diagrams in Section 7.4 are combined violin and box plots. They are especially useful in visualizing distributions of quantitative variables. Their interpretation is fairly straightforward: the greater the area under the curve for a given range, the more frequent the values are in this range. The thick center line of the box indicates the median, the outer lines of the box the first and third quartiles. Whiskers extend outwards to 1.5 times the inter-quartile range (IQR). Outliers beyond 1.5 times IQR are shown as individual points.

Please note that the x-axis is logarithmically scaled, i.e. in powers of 10. It therefore increases in a non-linear fashion. Additional sub-markings are included to assist with interpretation.

#### 7.3.2   Most Frequent Tokens

A token is defined as any character sequence delimited by whitespace on both sides, e.g. it roughly corresponds to the notion of a 'word'. However, due to the strictly syntactical definition tokens might also include arbitrary sequences of numbers or special characters.

The charts in Sections 7.5 and 7.6 show the 50 most frequent tokens for each language, weighted by both term frequency (TF) and term frequency/inverse document frequency (TF-IDF). Sequences of numbers, special symbols and a general list of frequent words for English and French ('stopwords') were removed prior to constructing the list. For details of the calculations, please refer to the Compilation Report and/or the Source Code.

The term frequency $\mathrm{tf}_{td}$ is calculated as the raw count of the number of times a term $t$ appears in a document $d$.

The term frequency/inverse document frequency $\mathrm{tf\text{-}idf}_{td}$ for a term $t$ in a document $d$ is calculated as follows, with $N$ the total number of documents in a corpus and $\mathrm{df}_t$ being the number of documents in the corpus in which the term $t$ appears:

$$\mathrm{tf\text{-}idf}_{td} = \mathrm{tf}_{td} \times \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

#### 7.3.3   Tokens over Time

The charts in Section 7.7 show the total output of the International Court of Justice for each year as the sum total of the tokens of all published decisions (judgments, advisory opinions, orders, appended opinions). These charts may give a rough estimate of the activity of the International Court of Justice, although they should be interpreted with caution, as duplicate and highly similar opinions were not removed for this simple analysis. Please refer to Section 8 for the scope of identical and near-identical documents in the corpus.

## 7.4 Distributions of Document Length

### 7.4.1 English

**CD–ICJ | EN | Version 2023–05–07 | Distributions of Document Length**

### 7.4.2 French
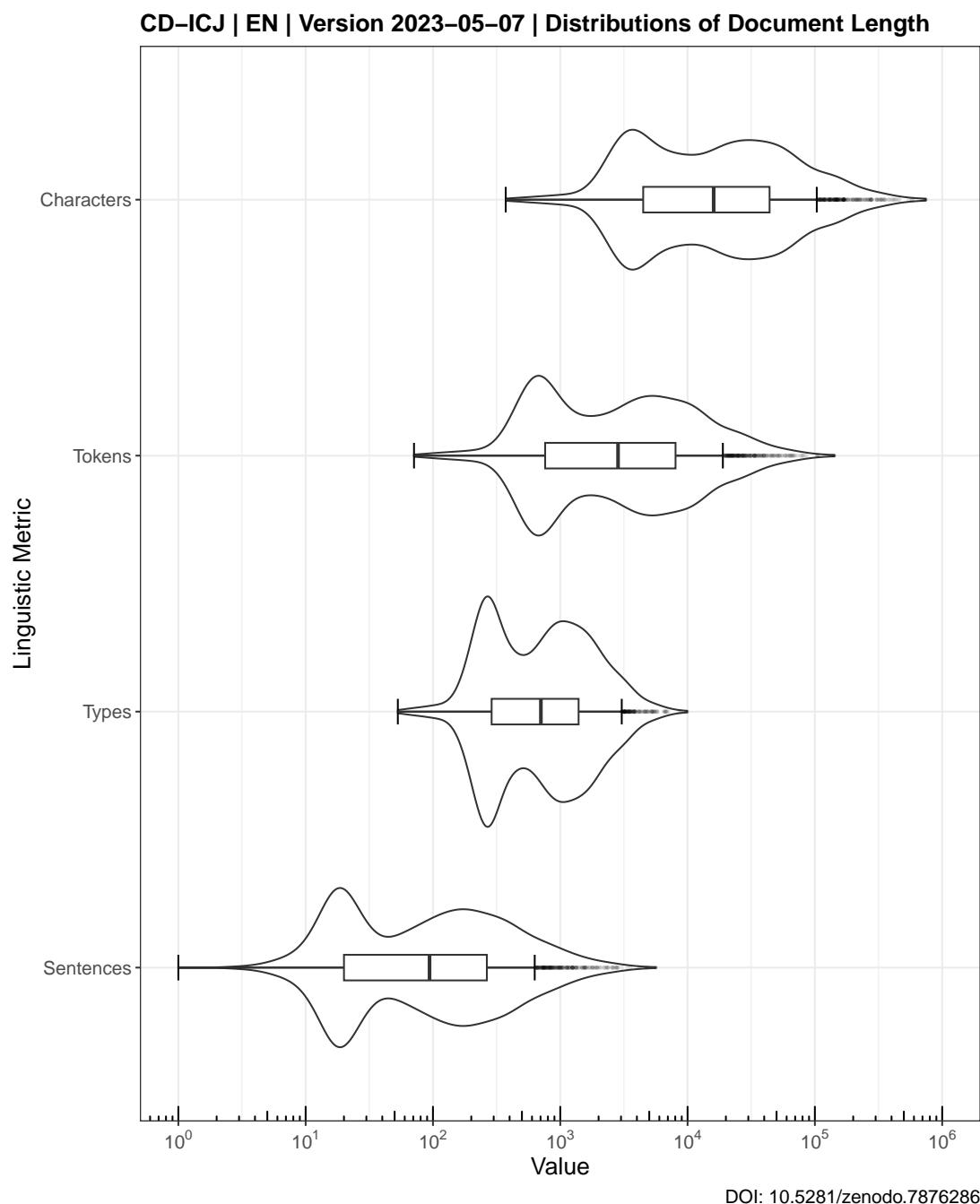
**CD–ICJ | FR | Version 2023–05–07 | Distributions of Document Length**
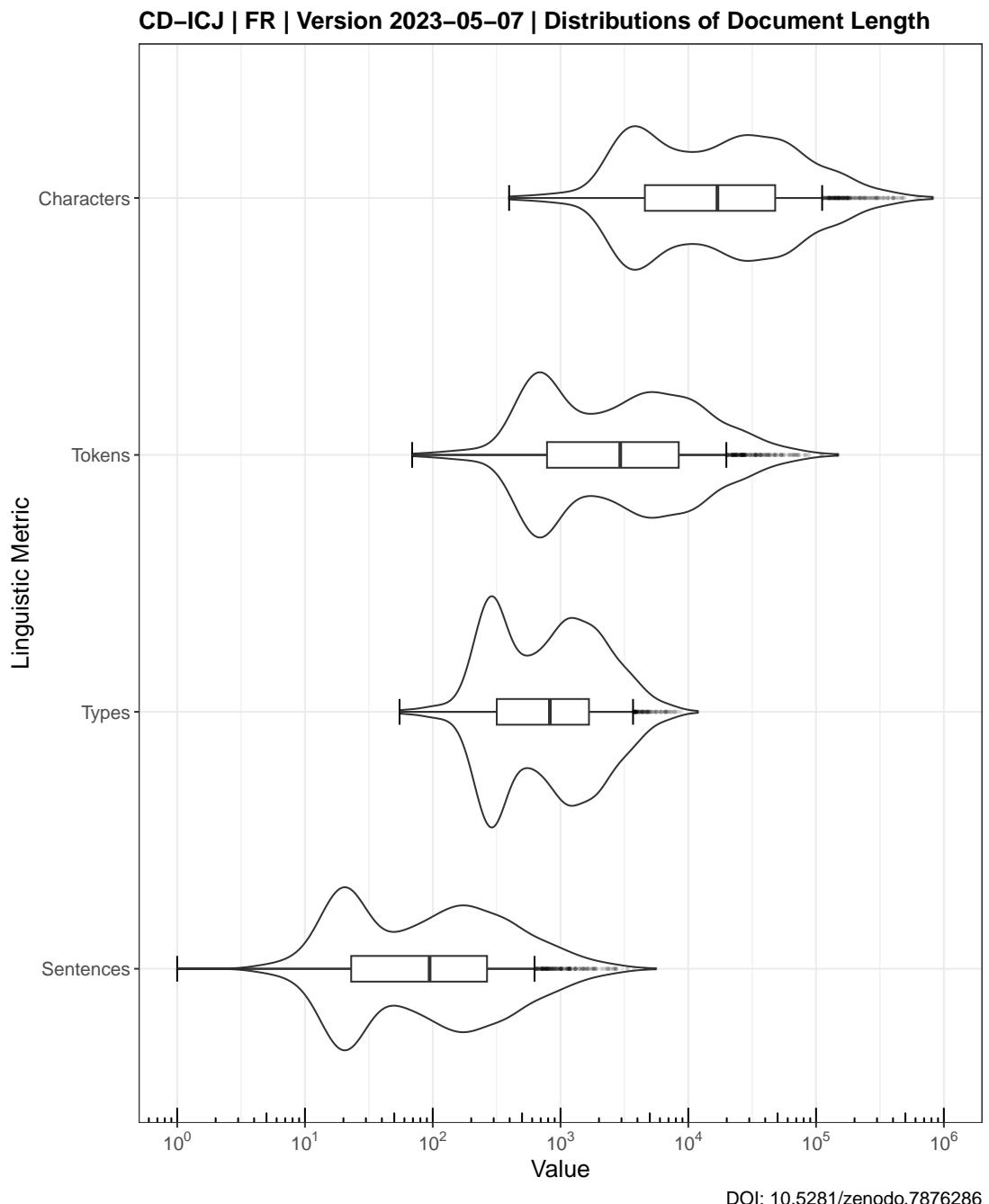
## 7.5 Most Frequent Tokens (English)

### 7.5.1 Term Frequency Weighting (TF)

**CD–ICJ | EN | Version 2023–05–07 | Top 50 Tokens | Term Frequency**
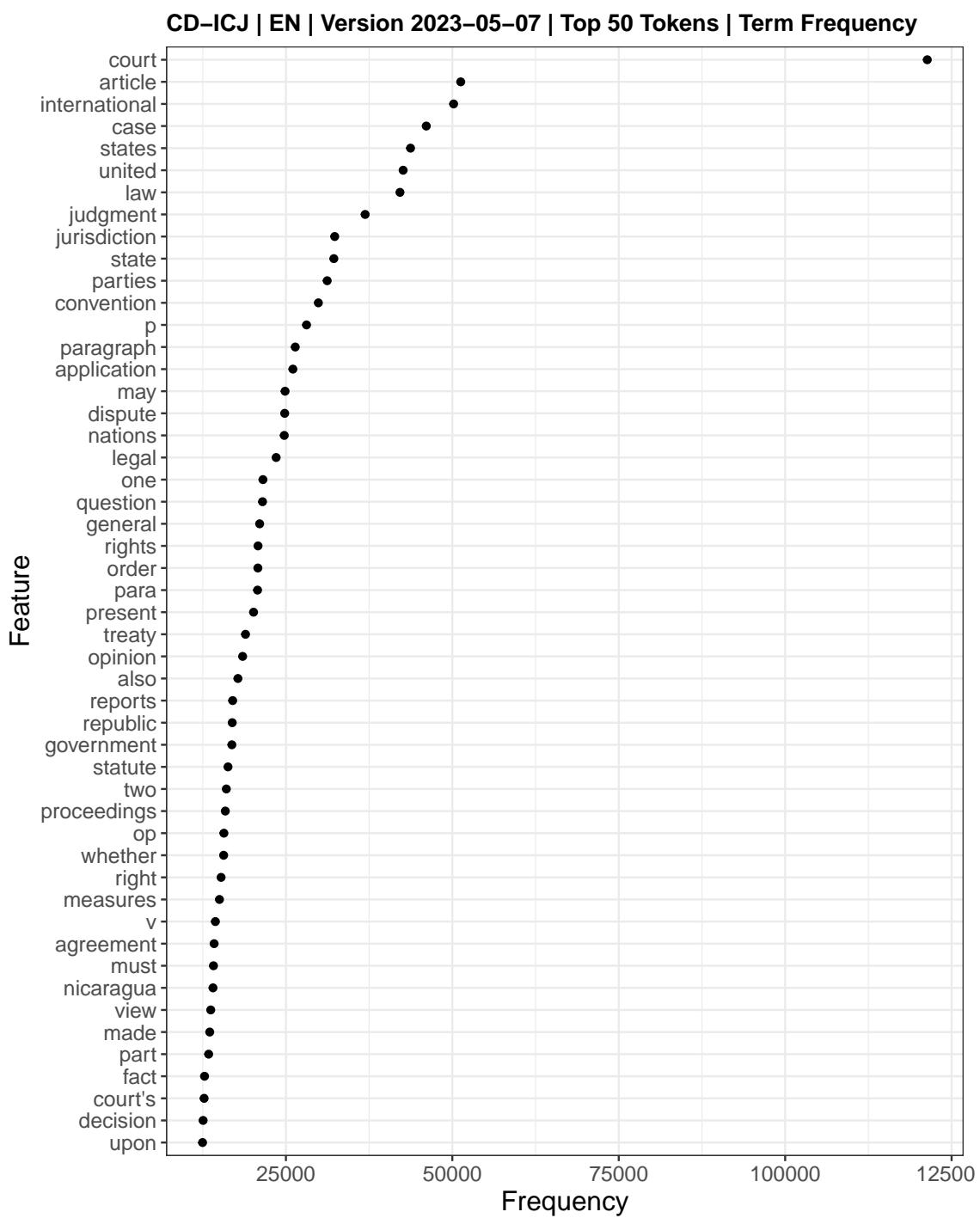
## 7.5.2  Term Frequency/Inverse Document Frequency Weighting (TF-IDF)
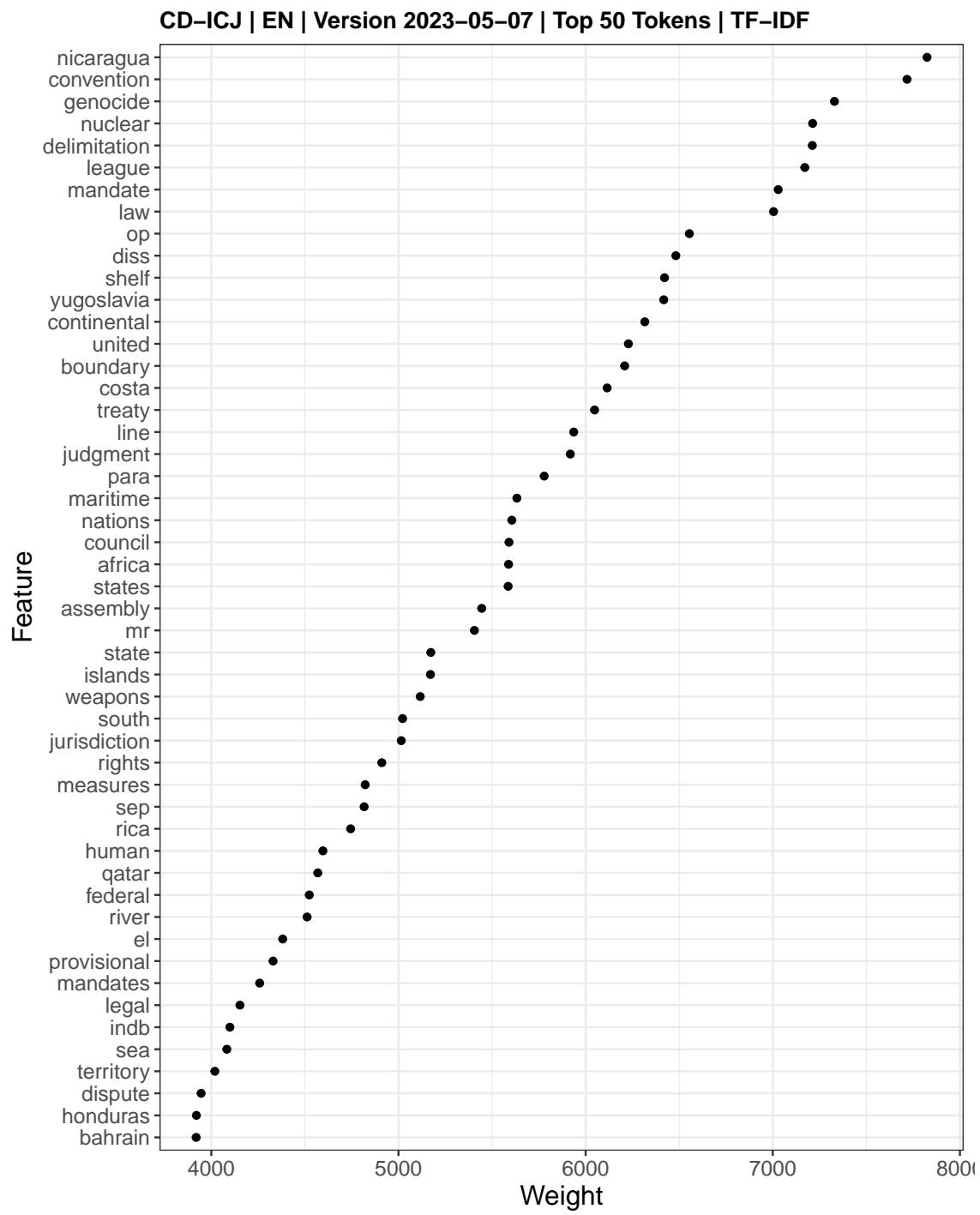
**CD–ICJ | EN | Version 2023–05–07 | Top 50 Tokens | TF–IDF**

## 7.6 Most Frequent Tokens (French)

### 7.6.1 Term Frequency Weighting (TF)
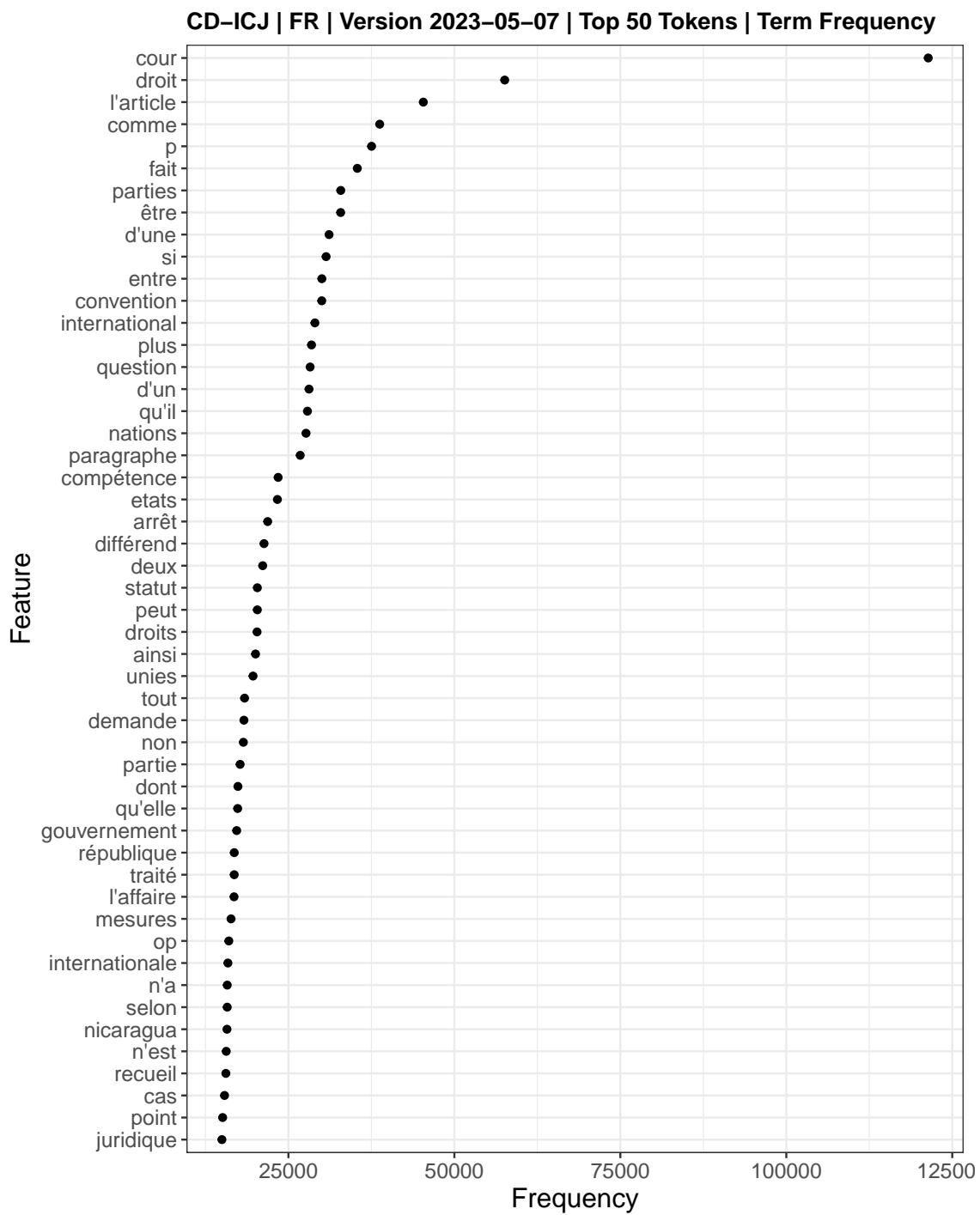
### 7.6.2 Term Frequency/Inverse Document Frequency Weighting (TF-IDF)

**CD–ICJ | FR | Version 2023–05–07 | Top 50 Tokens | TF–IDF**
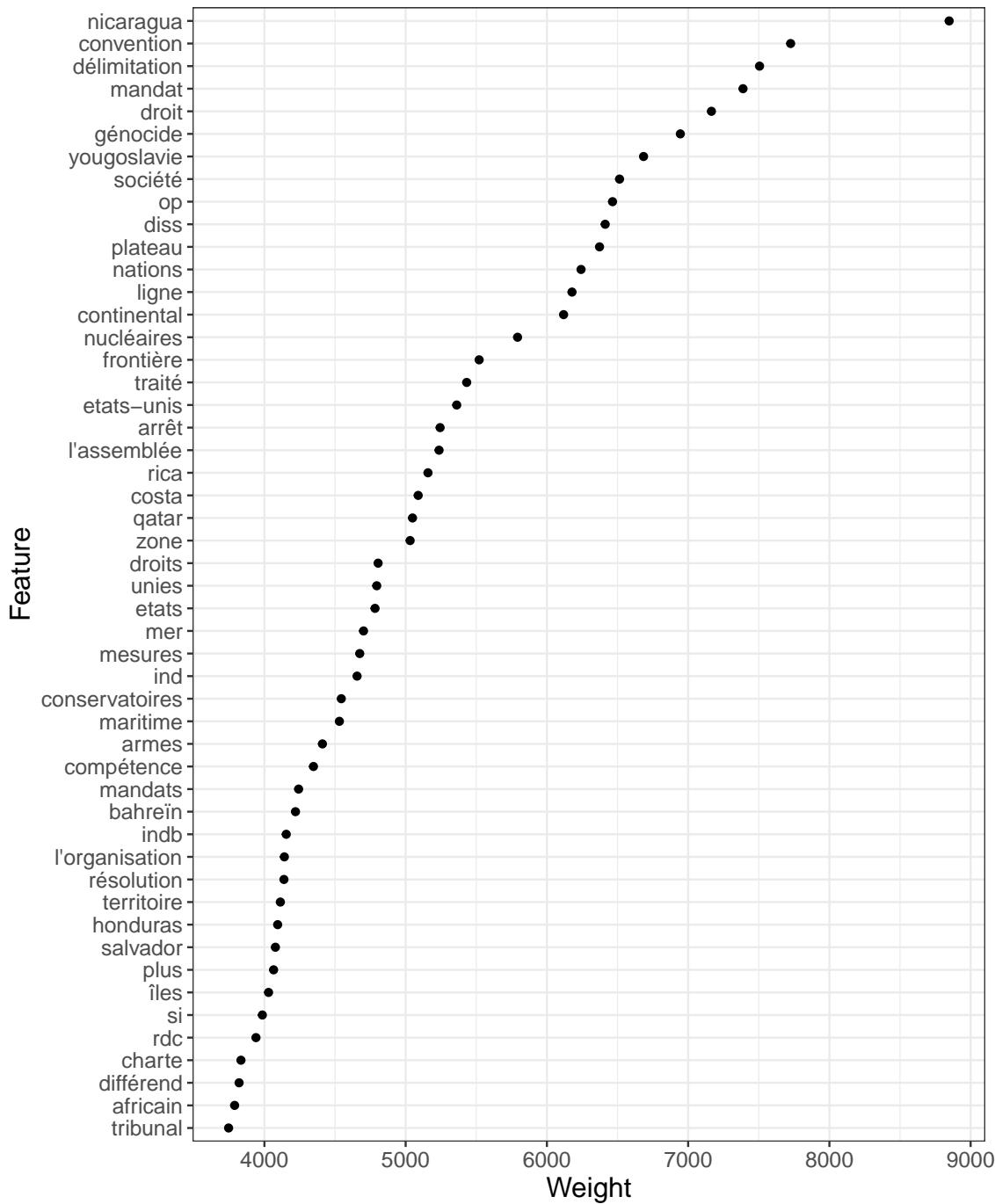
## 7.7 Tokens over Time

### 7.7.1 English

**CD–ICJ | EN | Version 2023–05–07 | Number of Tokens per Year**

### 7.7.2 French

**CD–ICJ | FR | Version 2023–05–07 | Number of Tokens per Year**
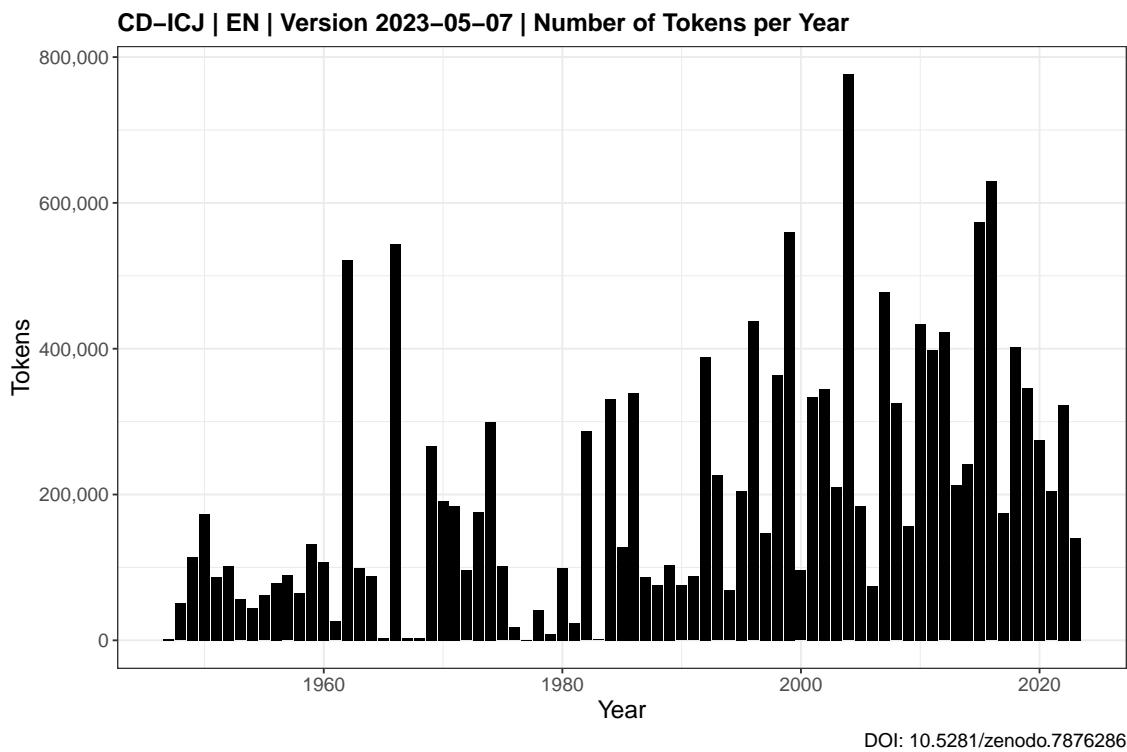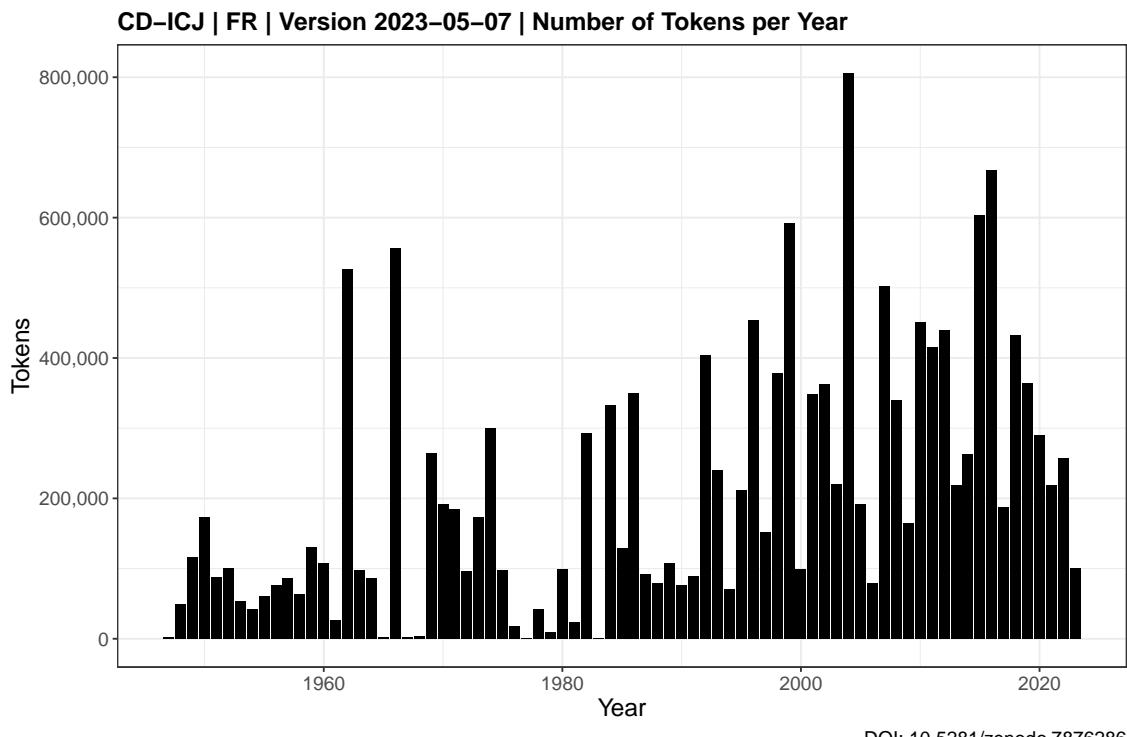
# 8 Document Similarity

## 8.1 English

**CD–ICJ | EN | Version 2023–05–07 | Document Similarity (Correlation)**

## 8.2 French

**CD–ICJ | FR | Version 2023–05–07 | Document Similarity (Correlation)**
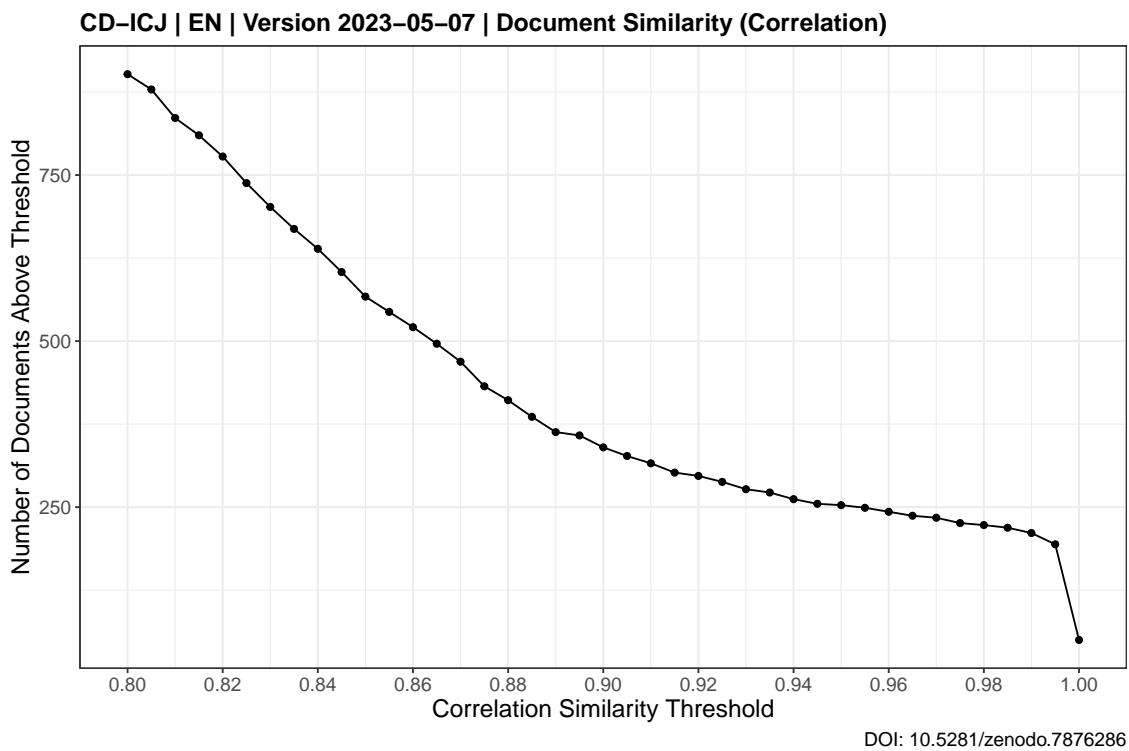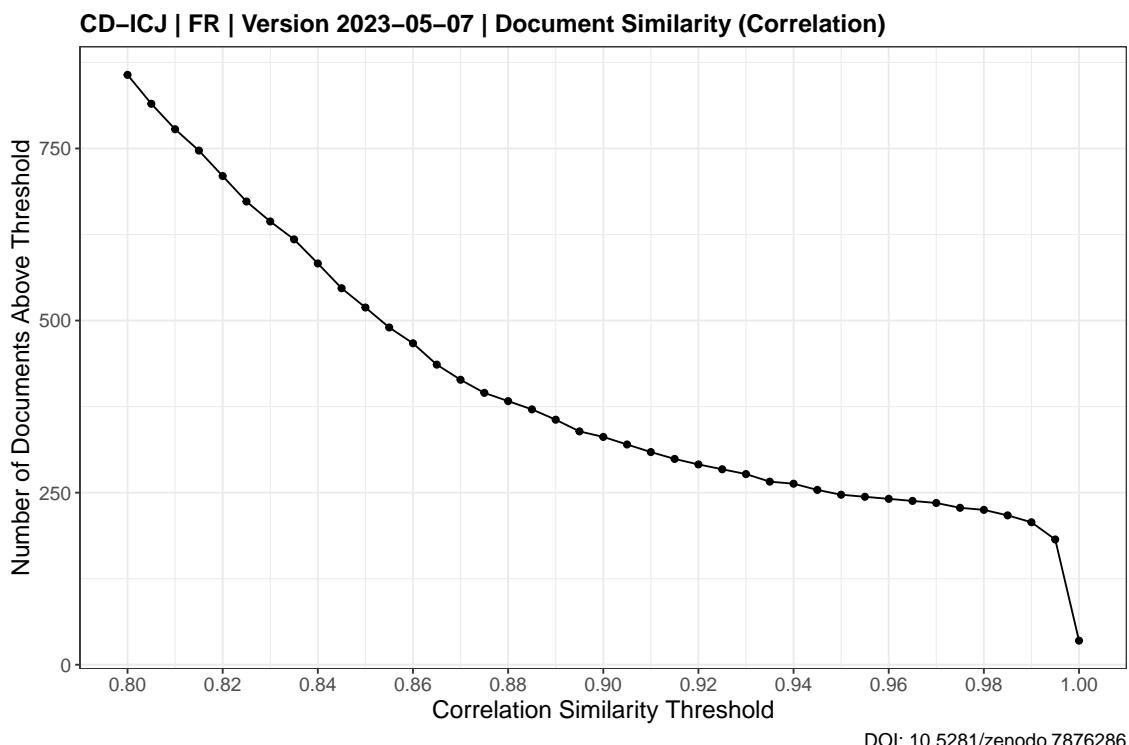
## 8.3   Comment

Analysts are advised that the CD-ICJ contains a non-negligible number of highly similar to near-identical documents. This is due to the Court's long-standing practice of issuing formally different decisions for each Applicant-Respondent pair in the course of the same proceedings. A prime example of such proceedings are the *Use of Force* cases, for which the judgments are identical in content, but differ only in the names of the Parties across more than half a dozen different judgments.

The above figures plot the number of files to be excluded as a function of correlation similarity based on a document-unigram matrix (with the removal of numbers, special symbols and stopwords, as well as lowercasing). Analysts who wish to qualitatively review this computational approach will find the IDs of presumed duplicates, together with the relevant value of correlation similarity, stored as CSV files in the 'ANALYSIS' archive published with the data set (item 17). These document IDs can also easily be read into statistical software and excluded directly from analyses without having to perform one's own similarity analysis. I do, however, recommend double-checking the IDs for false positives. The document pairings and similarity scores are included in a different CSV file (also item 17).

The choice of similarity algorithm, the threshold for marking a document as duplicate and the question of whether duplicate documents should be removed at all should be decided with respect to individual analyses. My goal is to document the Court's output as faithfully as possible and provide analysts with fair warning, as well as the opportunity to make their own choices. Please note that the manner of de-duplication will substantially affect analytical results and should be made after careful consideration of both methodology and the data.

# 9 Metadata Frequency Tables

## 9.1 By Year

### 9.1.1 English

**CD–ICJ | EN | Version 2023–05–07 | Documents per Year**
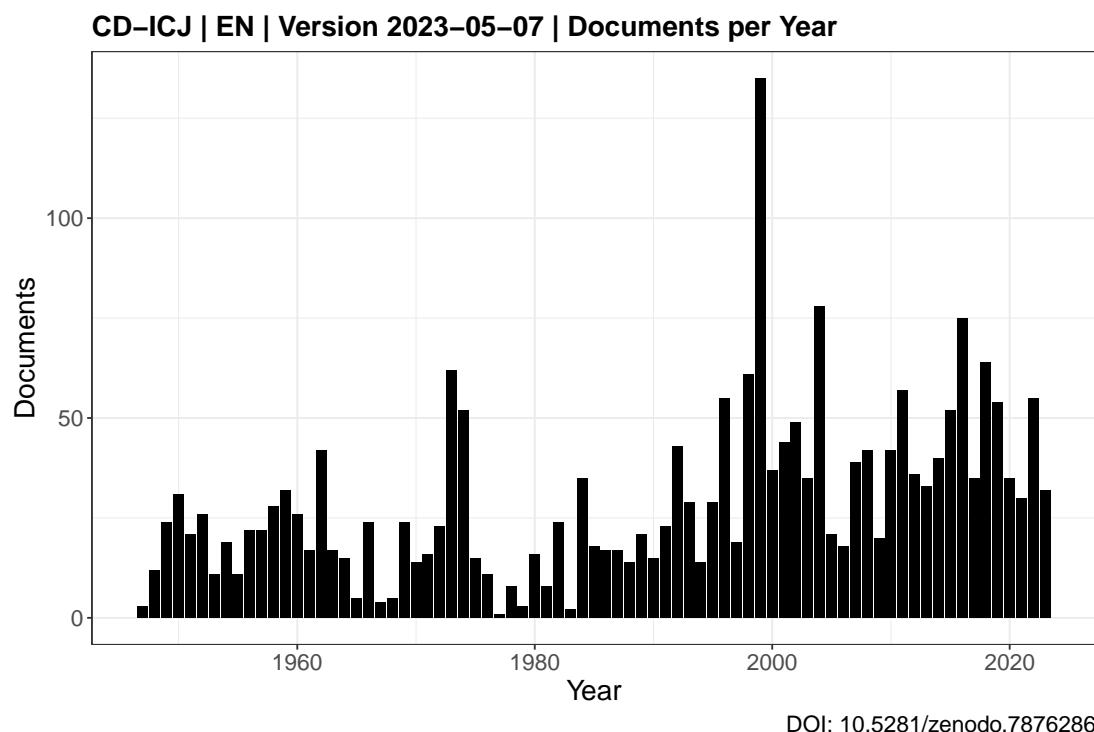
| Year | Documents | % Total | % Cumulative |
| --- | --- | --- | --- |
| 1947 | 3 | 0.13 | 0.13 |
| 1948 | 12 | 0.53 | 0.66 |
| 1949 | 24 | 1.06 | 1.72 |
| 1950 | 31 | 1.37 | 3.09 |
| 1951 | 21 | 0.93 | 4.02 |
| 1952 | 26 | 1.15 | 5.17 |
| 1953 | 11 | 0.49 | 5.65 |
| 1954 | 19 | 0.84 | 6.49 |
| 1955 | 11 | 0.49 | 6.98 |
| 1956 | 22 | 0.97 | 7.95 |
| 1957 | 22 | 0.97 | 8.92 |
| 1958 | 28 | 1.24 | 10.16 |
| 1959 | 32 | 1.41 | 11.57 |

| Year | Documents | % Total | % Cumulative |
| --- | --- | --- | --- |
| 1960 | 26 | 1.15 | 12.72 |
| 1961 | 17 | 0.75 | 13.47 |
| 1962 | 42 | 1.86 | 15.33 |
| 1963 | 17 | 0.75 | 16.08 |
| 1964 | 15 | 0.66 | 16.74 |
| 1965 | 5 | 0.22 | 16.96 |
| 1966 | 24 | 1.06 | 18.02 |
| 1967 | 4 | 0.18 | 18.20 |
| 1968 | 5 | 0.22 | 18.42 |
| 1969 | 24 | 1.06 | 19.48 |
| 1970 | 14 | 0.62 | 20.10 |
| 1971 | 16 | 0.71 | 20.80 |
| 1972 | 23 | 1.02 | 21.82 |
| 1973 | 62 | 2.74 | 24.56 |
| 1974 | 52 | 2.30 | 26.86 |
| 1975 | 15 | 0.66 | 27.52 |
| 1976 | 11 | 0.49 | 28.00 |
| 1977 | 1 | 0.04 | 28.05 |
| 1978 | 8 | 0.35 | 28.40 |
| 1979 | 3 | 0.13 | 28.53 |
| 1980 | 16 | 0.71 | 29.24 |
| 1981 | 8 | 0.35 | 29.59 |
| 1982 | 24 | 1.06 | 30.65 |
| 1983 | 2 | 0.09 | 30.74 |
| 1984 | 35 | 1.55 | 32.29 |
| 1985 | 18 | 0.80 | 33.08 |
| 1986 | 17 | 0.75 | 33.83 |
| 1987 | 17 | 0.75 | 34.58 |
| 1988 | 14 | 0.62 | 35.20 |
| 1989 | 21 | 0.93 | 36.13 |
| 1990 | 15 | 0.66 | 36.79 |
| 1991 | 23 | 1.02 | 37.81 |
| 1992 | 43 | 1.90 | 39.71 |

| Year | Documents | % Total | % Cumulative |
|---|---|---|---|
| 1993 | 29 | 1.28 | 40.99 |
| 1994 | 14 | 0.62 | 41.61 |
| 1995 | 29 | 1.28 | 42.89 |
| 1996 | 55 | 2.43 | 45.32 |
| 1997 | 19 | 0.84 | 46.16 |
| 1998 | 61 | 2.69 | 48.85 |
| 1999 | 135 | 5.96 | 54.81 |
| 2000 | 37 | 1.63 | 56.45 |
| 2001 | 44 | 1.94 | 58.39 |
| 2002 | 49 | 2.16 | 60.56 |
| 2003 | 35 | 1.55 | 62.10 |
| 2004 | 78 | 3.45 | 65.55 |
| 2005 | 21 | 0.93 | 66.48 |
| 2006 | 18 | 0.80 | 67.27 |
| 2007 | 39 | 1.72 | 68.99 |
| 2008 | 42 | 1.86 | 70.85 |
| 2009 | 20 | 0.88 | 71.73 |
| 2010 | 42 | 1.86 | 73.59 |
| 2011 | 57 | 2.52 | 76.10 |
| 2012 | 36 | 1.59 | 77.69 |
| 2013 | 33 | 1.46 | 79.15 |
| 2014 | 40 | 1.77 | 80.92 |
| 2015 | 52 | 2.30 | 83.22 |
| 2016 | 75 | 3.31 | 86.53 |
| 2017 | 35 | 1.55 | 88.07 |
| 2018 | 64 | 2.83 | 90.90 |
| 2019 | 54 | 2.39 | 93.29 |
| 2020 | 35 | 1.55 | 94.83 |
| 2021 | 30 | 1.33 | 96.16 |
| 2022 | 55 | 2.43 | 98.59 |
| 2023 | 32 | 1.41 | 100.00 |
| Total | 2264 | 100.00 | 100.00 |

### 9.1.2 French

**CD–ICJ | FR | Version 2023–05–07 | Documents per Year**
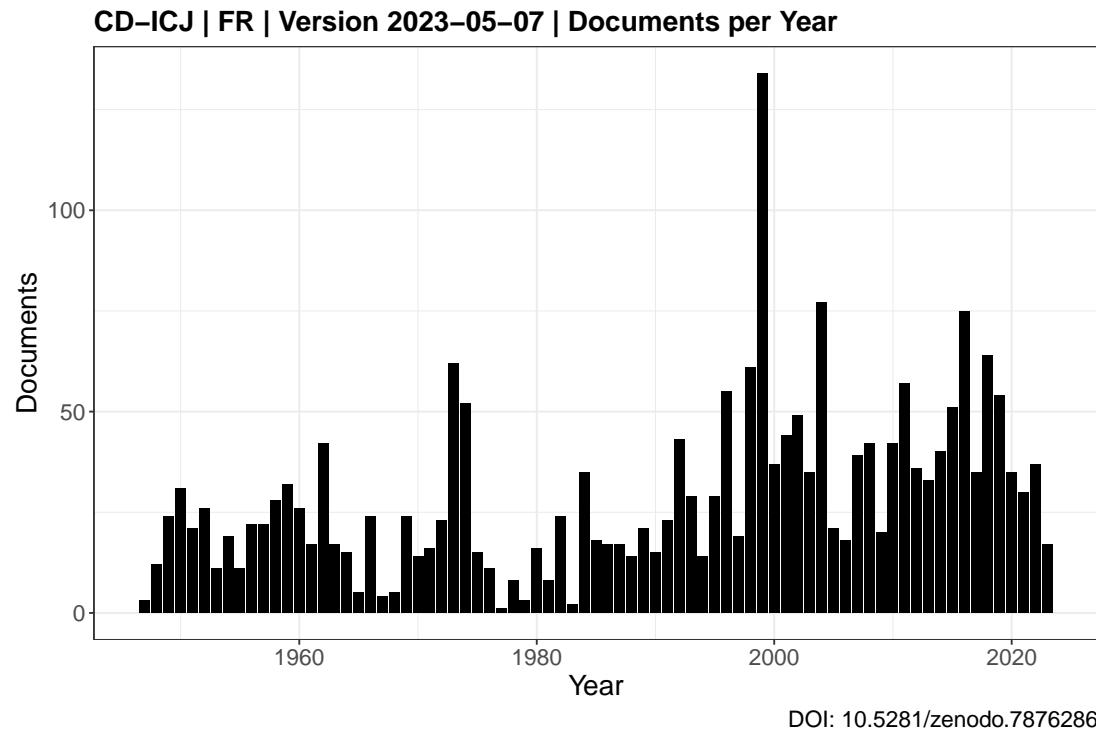
| Year | Documents | % Total | % Cumulative |
| --- | --- | --- | --- |
| 1947 | 3 | 0.13 | 0.13 |
| 1948 | 12 | 0.54 | 0.67 |
| 1949 | 24 | 1.08 | 1.75 |
| 1950 | 31 | 1.39 | 3.14 |
| 1951 | 21 | 0.94 | 4.08 |
| 1952 | 26 | 1.17 | 5.25 |
| 1953 | 11 | 0.49 | 5.75 |
| 1954 | 19 | 0.85 | 6.60 |
| 1955 | 11 | 0.49 | 7.09 |
| 1956 | 22 | 0.99 | 8.08 |
| 1957 | 22 | 0.99 | 9.07 |
| 1958 | 28 | 1.26 | 10.32 |
| 1959 | 32 | 1.44 | 11.76 |
| 1960 | 26 | 1.17 | 12.93 |
| 1961 | 17 | 0.76 | 13.69 |
| 1962 | 42 | 1.89 | 15.57 |

| Year | Documents | % Total | % Cumulative |
|------|-----------|---------|--------------|
| 1963 | 17 | 0.76 | 16.34 |
| 1964 | 15 | 0.67 | 17.01 |
| 1965 | 5 | 0.22 | 17.24 |
| 1966 | 24 | 1.08 | 18.31 |
| 1967 | 4 | 0.18 | 18.49 |
| 1968 | 5 | 0.22 | 18.72 |
| 1969 | 24 | 1.08 | 19.79 |
| 1970 | 14 | 0.63 | 20.42 |
| 1971 | 16 | 0.72 | 21.14 |
| 1972 | 23 | 1.03 | 22.17 |
| 1973 | 62 | 2.78 | 24.96 |
| 1974 | 52 | 2.33 | 27.29 |
| 1975 | 15 | 0.67 | 27.96 |
| 1976 | 11 | 0.49 | 28.46 |
| 1977 | 1 | 0.04 | 28.50 |
| 1978 | 8 | 0.36 | 28.86 |
| 1979 | 3 | 0.13 | 28.99 |
| 1980 | 16 | 0.72 | 29.71 |
| 1981 | 8 | 0.36 | 30.07 |
| 1982 | 24 | 1.08 | 31.15 |
| 1983 | 2 | 0.09 | 31.24 |
| 1984 | 35 | 1.57 | 32.81 |
| 1985 | 18 | 0.81 | 33.62 |
| 1986 | 17 | 0.76 | 34.38 |
| 1987 | 17 | 0.76 | 35.14 |
| 1988 | 14 | 0.63 | 35.77 |
| 1989 | 21 | 0.94 | 36.71 |
| 1990 | 15 | 0.67 | 37.39 |
| 1991 | 23 | 1.03 | 38.42 |
| 1992 | 43 | 1.93 | 40.35 |
| 1993 | 29 | 1.30 | 41.65 |
| 1994 | 14 | 0.63 | 42.28 |
| 1995 | 29 | 1.30 | 43.58 |
| 1996 | 55 | 2.47 | 46.05 |

| Year | Documents | % Total | % Cumulative |
|---|---|---|---|
| 1997 | 19 | 0.85 | 46.90 |
| 1998 | 61 | 2.74 | 49.64 |
| 1999 | 134 | 6.01 | 55.66 |
| 2000 | 37 | 1.66 | 57.32 |
| 2001 | 44 | 1.97 | 59.29 |
| 2002 | 49 | 2.20 | 61.49 |
| 2003 | 35 | 1.57 | 63.06 |
| 2004 | 77 | 3.46 | 66.52 |
| 2005 | 21 | 0.94 | 67.46 |
| 2006 | 18 | 0.81 | 68.27 |
| 2007 | 39 | 1.75 | 70.02 |
| 2008 | 42 | 1.89 | 71.90 |
| 2009 | 20 | 0.90 | 72.80 |
| 2010 | 42 | 1.89 | 74.69 |
| 2011 | 57 | 2.56 | 77.24 |
| 2012 | 36 | 1.62 | 78.86 |
| 2013 | 33 | 1.48 | 80.34 |
| 2014 | 40 | 1.80 | 82.14 |
| 2015 | 51 | 2.29 | 84.43 |
| 2016 | 75 | 3.37 | 87.79 |
| 2017 | 35 | 1.57 | 89.36 |
| 2018 | 64 | 2.87 | 92.24 |
| 2019 | 54 | 2.42 | 94.66 |
| 2020 | 35 | 1.57 | 96.23 |
| 2021 | 30 | 1.35 | 97.58 |
| 2022 | 37 | 1.66 | 99.24 |
| 2023 | 17 | 0.76 | 100.00 |
| Total | 2228 | 100.00 | 100.00 |

## 9.2 By Document Type

### 9.2.1 English

**CD–ICJ | EN | Version 2023–05–07 | Documents per Document Type**
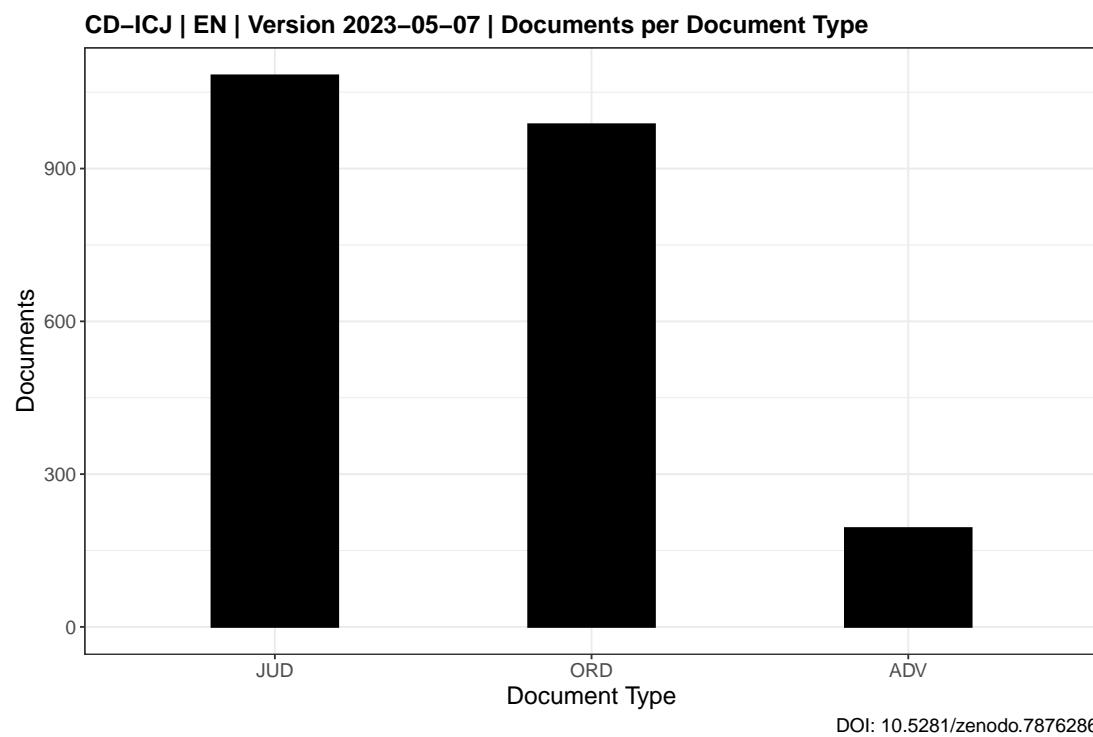
| DocType | Documents | % Total | % Cumulative |
|---------|-----------|---------|--------------|
| ADV | 194 | 8.57 | 8.57 |
| JUD | 1083 | 47.84 | 56.40 |
| ORD | 987 | 43.60 | 100.00 |
| Total | 2264 | 100.00 | 100.00 |

### 9.2.2 French

**CD–ICJ | FR | Version 2023–05–07 | Documents per Document Type**
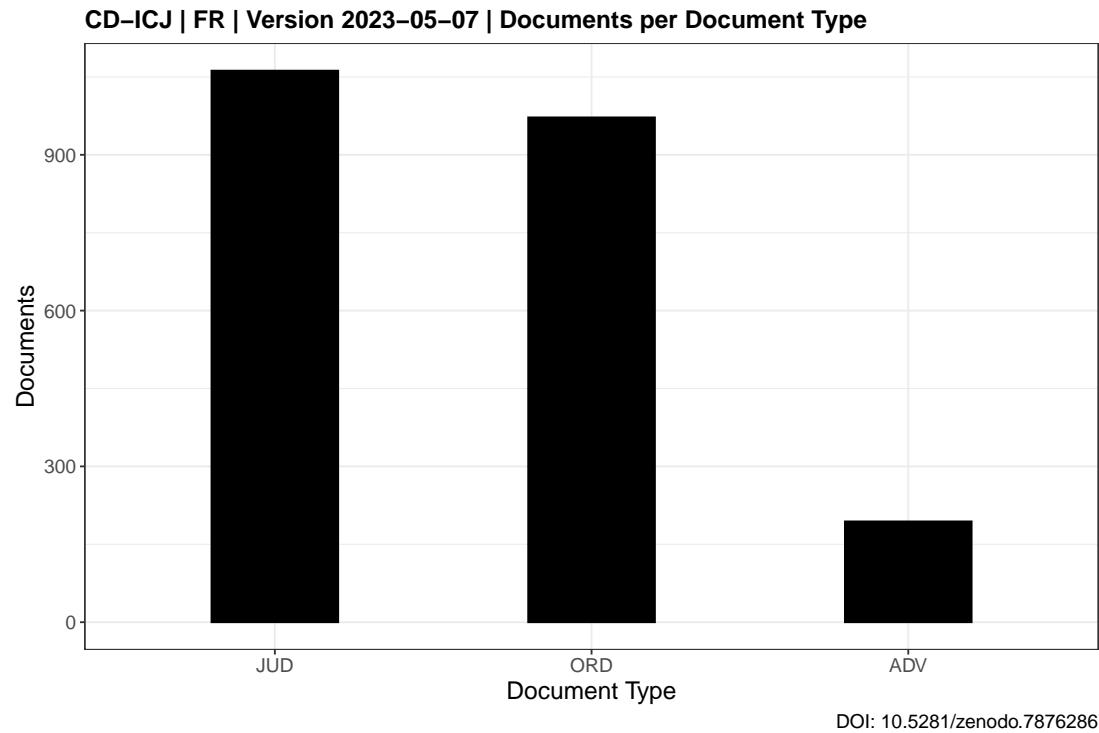
| DocType | Documents | % Total | % Cumulative |
|---------|-----------|---------|--------------|
| ADV     | 194       | 8.71    | 8.71         |
| JUD     | 1062      | 47.67   | 56.37        |
| ORD     | 972       | 43.63   | 100.00       |
| Total   | 2228      | 100.00  | 100.00       |

## 9.3 By Opinion Number

### 9.3.1 English

**CD–ICJ | EN | Version 2023–05–07 | Documents per Opinion Number**
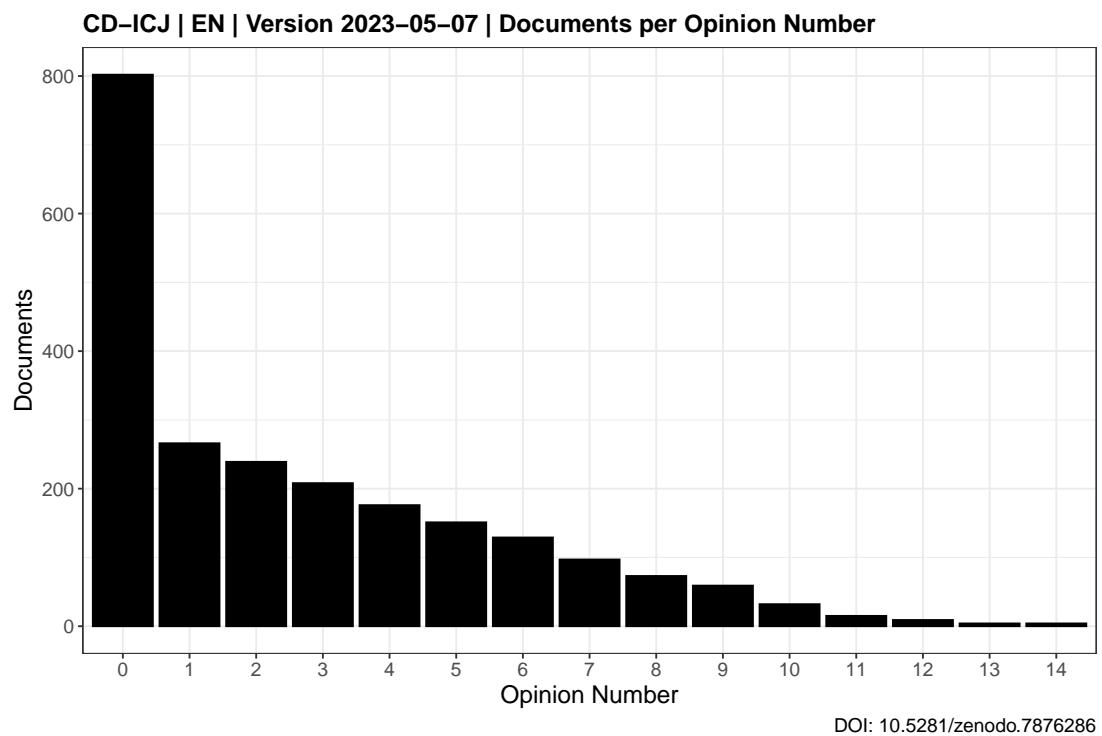
| Opinion Number | Documents | % Total | % Cumulative |
| --- | --- | --- | --- |
| 0 | 802 | 35.42 | 35.42 |
| 1 | 266 | 11.75 | 47.17 |
| 2 | 239 | 10.56 | 57.73 |
| 3 | 208 | 9.19 | 66.92 |
| 4 | 176 | 7.77 | 74.69 |
| 5 | 151 | 6.67 | 81.36 |
| 6 | 129 | 5.70 | 87.06 |
| 7 | 97 | 4.28 | 91.34 |
| 8 | 73 | 3.22 | 94.57 |
| 9 | 59 | 2.61 | 97.17 |
| 10 | 32 | 1.41 | 98.59 |
| 11 | 15 | 0.66 | 99.25 |
| 12 | 9 | 0.40 | 99.65 |
| 13 | 4 | 0.18 | 99.82 |
| 14 | 4 | 0.18 | 100.00 |
| Total | 2264 | 100.00 | 100.00 |

### 9.3.2 French

**CD–ICJ | FR | Version 2023–05–07 | Documents per Opinion Number**
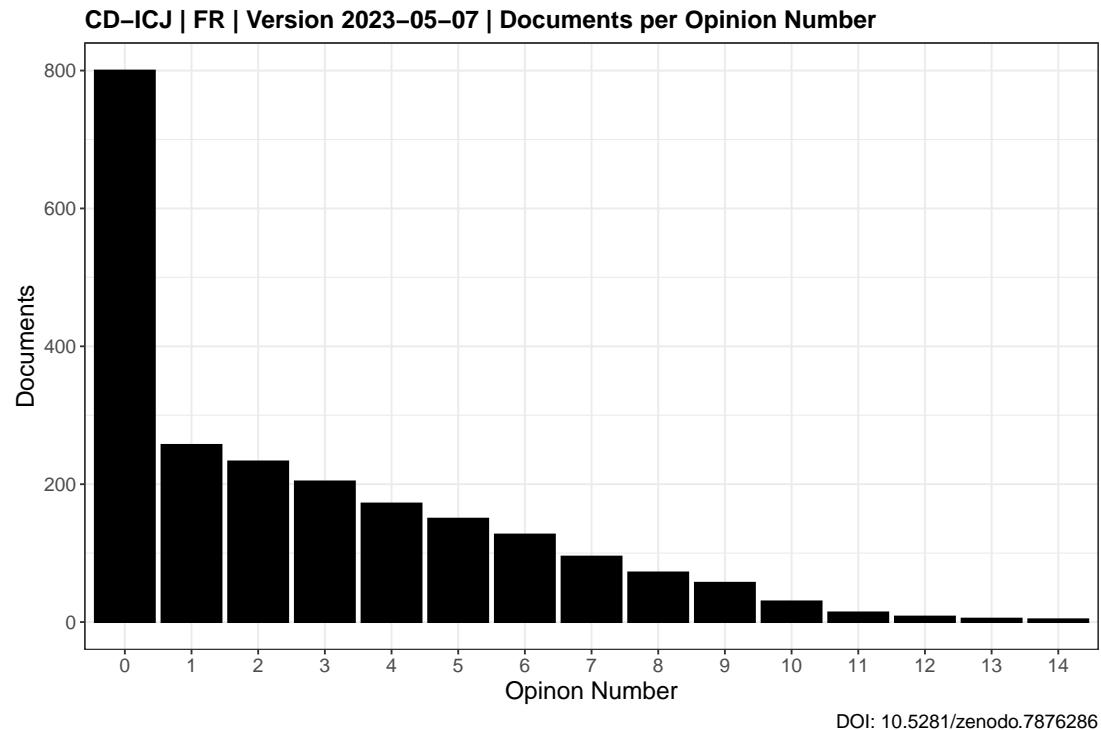
| Opinion Number | Documents | % Total | % Cumulative |
|:---:|:---:|:---:|:---:|
| 0 | 800 | 35.91 | 35.91 |
| 1 | 257 | 11.54 | 47.44 |
| 2 | 233 | 10.46 | 57.90 |
| 3 | 204 | 9.16 | 67.06 |
| 4 | 172 | 7.72 | 74.78 |
| 5 | 150 | 6.73 | 81.51 |
| 6 | 127 | 5.70 | 87.21 |
| 7 | 95 | 4.26 | 91.47 |
| 8 | 72 | 3.23 | 94.70 |
| 9 | 57 | 2.56 | 97.26 |
| 10 | 30 | 1.35 | 98.61 |
| 11 | 14 | 0.63 | 99.24 |
| 12 | 8 | 0.36 | 99.60 |
| 13 | 5 | 0.22 | 99.82 |
| 14 | 4 | 0.18 | 100.00 |
| Total | 2228 | 100.00 | 100.00 |

## 9.4 By Applicant

### 9.4.1 English

| Applicant | Documents | % Total | % Cumulative |
| --- | --- | --- | --- |
| ARG | 20 | 0.88 | 0.88 |
| ARM | 14 | 0.62 | 1.50 |
| AUS | 48 | 2.12 | 3.62 |
| AZE | 8 | 0.35 | 3.98 |
| BEL | 64 | 2.83 | 6.80 |
| BEN | 7 | 0.31 | 7.11 |
| BFA | 16 | 0.71 | 7.82 |
| BHR-EGY-ARE | 6 | 0.27 | 8.08 |
| BHR-EGY-SAU-ARE | 6 | 0.27 | 8.35 |
| BIH | 49 | 2.16 | 10.51 |
| BLZ | 1 | 0.04 | 10.56 |
| BOL | 14 | 0.62 | 11.17 |
| BWA | 12 | 0.53 | 11.70 |
| CAN | 12 | 0.53 | 12.23 |
| CARAT | 32 | 1.41 | 13.65 |
| CHE | 16 | 0.71 | 14.36 |
| CHL | 8 | 0.35 | 14.71 |
| CMR | 56 | 2.47 | 17.18 |
| COD | 90 | 3.98 | 21.16 |
| COL | 12 | 0.53 | 21.69 |
| CRI | 69 | 3.05 | 24.73 |
| DEU | 83 | 3.67 | 28.40 |
| DJI | 11 | 0.49 | 28.89 |
| DMA | 1 | 0.04 | 28.93 |
| DNK | 13 | 0.57 | 29.51 |
| ECOSOC | 11 | 0.49 | 29.99 |
| ECU | 4 | 0.18 | 30.17 |
| ESP | 13 | 0.57 | 30.74 |
| ETH | 30 | 1.33 | 32.07 |
| FIN | 7 | 0.31 | 32.38 |
| FRA | 32 | 1.41 | 33.79 |

| Applicant | Documents | % Total | % Cumulative |
|-----------|-----------|---------|--------------|
| GAB | 2 | 0.09 | 33.88 |
| GBR | 77 | 3.40 | 37.28 |
| GEO | 16 | 0.71 | 37.99 |
| GIN | 24 | 1.06 | 39.05 |
| GMB | 12 | 0.53 | 39.58 |
| GNB | 16 | 0.71 | 40.28 |
| GNQ | 28 | 1.24 | 41.52 |
| GRC | 36 | 1.59 | 43.11 |
| GTM | 3 | 0.13 | 43.24 |
| GUY | 16 | 0.71 | 43.95 |
| HND | 7 | 0.31 | 44.26 |
| HRV | 31 | 1.37 | 45.63 |
| HUN | 16 | 0.71 | 46.33 |
| IDN | 14 | 0.62 | 46.95 |
| IFAD | 5 | 0.22 | 47.17 |
| IMO | 4 | 0.18 | 47.35 |
| IND | 25 | 1.10 | 48.45 |
| IRN | 78 | 3.45 | 51.90 |
| ISR | 8 | 0.35 | 52.25 |
| ITA | 7 | 0.31 | 52.56 |
| KHM | 30 | 1.33 | 53.89 |
| LBR | 30 | 1.33 | 55.21 |
| LBY | 84 | 3.71 | 58.92 |
| LIE | 19 | 0.84 | 59.76 |
| MEX | 19 | 0.84 | 60.60 |
| MHL | 51 | 2.25 | 62.85 |
| MKD | 8 | 0.35 | 63.21 |
| MYS | 11 | 0.49 | 63.69 |
| NIC | 146 | 6.45 | 70.14 |
| NLD | 11 | 0.49 | 70.63 |
| NRU | 11 | 0.49 | 71.11 |
| NZL | 37 | 1.63 | 72.75 |
| PAK | 12 | 0.53 | 73.28 |
| PER | 13 | 0.57 | 73.85 |

| Applicant | Documents | % Total | % Cumulative |
|---|---|---|---|
| PRT | 35 | 1.55 | 75.40 |
| PRY | 7 | 0.31 | 75.71 |
| PSE | 1 | 0.04 | 75.75 |
| QAT | 51 | 2.25 | 78.00 |
| ROU | 4 | 0.18 | 78.18 |
| SCG | 163 | 7.20 | 85.38 |
| SLV | 24 | 1.06 | 86.44 |
| SOM | 17 | 0.75 | 87.19 |
| TLS | 11 | 0.49 | 87.68 |
| TUN | 19 | 0.84 | 88.52 |
| UKR | 35 | 1.55 | 90.06 |
| UNESCO | 9 | 0.40 | 90.46 |
| UNGA | 143 | 6.32 | 96.78 |
| UNSC | 16 | 0.71 | 97.48 |
| USA | 21 | 0.93 | 98.41 |
| WHO | 20 | 0.88 | 99.29 |
| YUG | 16 | 0.71 | 100.00 |
| Total | 2264 | 100.00 | 100.00 |

### 9.4.2 French

| Applicant | Documents | % Total | % Cumulative |
|---|---|---|---|
| ARG | 20 | 0.90 | 0.90 |
| ARM | 9 | 0.40 | 1.30 |
| AUS | 48 | 2.15 | 3.46 |
| AZE | 5 | 0.22 | 3.68 |
| BEL | 64 | 2.87 | 6.55 |
| BEN | 6 | 0.27 | 6.82 |
| BFA | 16 | 0.72 | 7.54 |
| BHR-EGY-ARE | 6 | 0.27 | 7.81 |
| BHR-EGY-SAU-ARE | 6 | 0.27 | 8.08 |
| BIH | 49 | 2.20 | 10.28 |
| BLZ | 1 | 0.04 | 10.32 |
| BOL | 14 | 0.63 | 10.95 |
| BWA | 12 | 0.54 | 11.49 |
| CAN | 12 | 0.54 | 12.03 |
| CARAT | 32 | 1.44 | 13.46 |
| CHE | 16 | 0.72 | 14.18 |
| CHL | 6 | 0.27 | 14.45 |
| CMR | 56 | 2.51 | 16.97 |
| COD | 90 | 4.04 | 21.01 |
| COL | 12 | 0.54 | 21.54 |
| CRI | 69 | 3.10 | 24.64 |
| DEU | 83 | 3.73 | 28.37 |
| DJI | 11 | 0.49 | 28.86 |
| DMA | 1 | 0.04 | 28.90 |
| DNK | 13 | 0.58 | 29.49 |
| ECOSOC | 11 | 0.49 | 29.98 |
| ECU | 4 | 0.18 | 30.16 |
| ESP | 13 | 0.58 | 30.75 |
| ETH | 30 | 1.35 | 32.09 |
| FIN | 7 | 0.31 | 32.41 |
| FRA | 32 | 1.44 | 33.84 |
| GAB | 2 | 0.09 | 33.93 |

| Applicant | Documents | % Total | % Cumulative |
| --- | --- | --- | --- |
| GBR | 77 | 3.46 | 37.39 |
| GEO | 16 | 0.72 | 38.11 |
| GIN | 24 | 1.08 | 39.18 |
| GMB | 10 | 0.45 | 39.63 |
| GNB | 16 | 0.72 | 40.35 |
| GNQ | 28 | 1.26 | 41.61 |
| GRC | 36 | 1.62 | 43.22 |
| GTM | 3 | 0.13 | 43.36 |
| GUY | 13 | 0.58 | 43.94 |
| HND | 7 | 0.31 | 44.25 |
| HRV | 31 | 1.39 | 45.65 |
| HUN | 16 | 0.72 | 46.36 |
| IDN | 14 | 0.63 | 46.99 |
| IFAD | 5 | 0.22 | 47.22 |
| IMO | 4 | 0.18 | 47.40 |
| IND | 25 | 1.12 | 48.52 |
| IRN | 71 | 3.19 | 51.71 |
| ISR | 8 | 0.36 | 52.06 |
| ITA | 7 | 0.31 | 52.38 |
| KHM | 30 | 1.35 | 53.73 |
| LBR | 30 | 1.35 | 55.07 |
| LBY | 83 | 3.73 | 58.80 |
| LIE | 19 | 0.85 | 59.65 |
| MEX | 19 | 0.85 | 60.50 |
| MHL | 51 | 2.29 | 62.79 |
| MKD | 8 | 0.36 | 63.15 |
| MYS | 11 | 0.49 | 63.64 |
| NIC | 139 | 6.24 | 69.88 |
| NLD | 11 | 0.49 | 70.38 |
| NRU | 11 | 0.49 | 70.87 |
| NZL | 37 | 1.66 | 72.53 |
| PAK | 12 | 0.54 | 73.07 |
| PER | 13 | 0.58 | 73.65 |

| Applicant | Documents | % Total | % Cumulative |
|-----------|-----------|---------|--------------|
| PRT | 35 | 1.57 | 75.22 |
| PRY | 7 | 0.31 | 75.54 |
| PSE | 1 | 0.04 | 75.58 |
| QAT | 51 | 2.29 | 77.87 |
| ROU | 4 | 0.18 | 78.05 |
| SCG | 163 | 7.32 | 85.37 |
| SLV | 24 | 1.08 | 86.45 |
| SOM | 17 | 0.76 | 87.21 |
| TLS | 10 | 0.45 | 87.66 |
| TUN | 19 | 0.85 | 88.51 |
| UKR | 31 | 1.39 | 89.90 |
| UNESCO | 9 | 0.40 | 90.31 |
| UNGA | 143 | 6.42 | 96.72 |
| UNSC | 16 | 0.72 | 97.44 |
| USA | 21 | 0.94 | 98.38 |
| WHO | 20 | 0.90 | 99.28 |
| YUG | 16 | 0.72 | 100.00 |
| Total | 2228 | 100.00 | 100.00 |

## 9.5 By Respondent

### 9.5.1 English

| Respondent | Documents | % Total | % Cumulative |
|:---:|:---:|:---:|:---:|
| NA | 240 | 10.60 | 10.60 |
| ALB | 19 | 0.84 | 11.44 |
| ARE | 22 | 0.97 | 12.41 |
| ARG | 1 | 0.04 | 12.46 |
| ARM | 8 | 0.35 | 12.81 |
| AUS | 32 | 1.41 | 14.22 |
| AZE | 14 | 0.62 | 14.84 |
| BDI | 3 | 0.13 | 14.97 |
| BEL | 43 | 1.90 | 16.87 |
| BGR | 21 | 0.93 | 17.80 |
| BHR | 29 | 1.28 | 19.08 |
| BLZ | 3 | 0.13 | 19.21 |
| BOL | 8 | 0.35 | 19.57 |
| BRA | 1 | 0.04 | 19.61 |
| CAN | 34 | 1.50 | 21.11 |
| CHE | 4 | 0.18 | 21.29 |
| CHL | 28 | 1.24 | 22.53 |
| COD | 24 | 1.06 | 23.59 |
| COL | 80 | 3.53 | 27.12 |
| CRI | 19 | 0.84 | 27.96 |
| CSK | 1 | 0.04 | 28.00 |
| DEU | 27 | 1.19 | 29.20 |
| DNK | 22 | 0.97 | 30.17 |
| EGY | 1 | 0.04 | 30.21 |
| ESP | 52 | 2.30 | 32.51 |
| FRA | 137 | 6.05 | 38.56 |
| FRA-GBR-USA | 7 | 0.31 | 38.87 |
| GBR | 107 | 4.73 | 43.60 |
| GNQ | 2 | 0.09 | 43.68 |
| GRC | 8 | 0.35 | 44.04 |
| GTM | 11 | 0.49 | 44.52 |
| HND | 44 | 1.94 | 46.47 |

| Respondent | Documents | % Total | % Cumulative |
| --- | --- | --- | --- |
| HUN | 1 | 0.04 | 46.51 |
| IND | 54 | 2.39 | 48.90 |
| IRN | 19 | 0.84 | 49.73 |
| ISL | 49 | 2.16 | 51.90 |
| ITA | 43 | 1.90 | 53.80 |
| JPN | 17 | 0.75 | 54.55 |
| KEN | 17 | 0.75 | 55.30 |
| LBN | 7 | 0.31 | 55.61 |
| LBY | 19 | 0.84 | 56.45 |
| MLI | 8 | 0.35 | 56.80 |
| MLT | 22 | 0.97 | 57.77 |
| MMR | 12 | 0.53 | 58.30 |
| MYS | 14 | 0.62 | 58.92 |
| NAM | 12 | 0.53 | 59.45 |
| NER | 15 | 0.66 | 60.11 |
| NGA | 38 | 1.68 | 61.79 |
| NIC | 75 | 3.31 | 65.11 |
| NLD | 44 | 1.94 | 67.05 |
| NOR | 33 | 1.46 | 68.51 |
| PAK | 42 | 1.86 | 70.36 |
| PER | 12 | 0.53 | 70.89 |
| PRT | 21 | 0.93 | 71.82 |
| QAT | 12 | 0.53 | 72.35 |
| RUS | 51 | 2.25 | 74.60 |
| RWA | 19 | 0.84 | 75.44 |
| SCG | 43 | 1.90 | 77.34 |
| SEN | 33 | 1.46 | 78.80 |
| SGP | 11 | 0.49 | 79.28 |
| SRB | 31 | 1.37 | 80.65 |
| SUN | 4 | 0.18 | 80.83 |
| SVK | 16 | 0.71 | 81.54 |
| SWE | 11 | 0.49 | 82.02 |
| TCD | 8 | 0.35 | 82.38 |

| Respondent | Documents | % Total | % Cumulative |
|------------|-----------|---------|--------------|
| THA | 30 | 1.33 | 83.70 |
| TUR | 20 | 0.88 | 84.58 |
| UGA | 35 | 1.55 | 86.13 |
| UKR | 4 | 0.18 | 86.31 |
| URY | 20 | 0.88 | 87.19 |
| USA | 208 | 9.19 | 96.38 |
| VEN | 16 | 0.71 | 97.08 |
| YUG | 6 | 0.27 | 97.35 |
| ZAF | 60 | 2.65 | 100.00 |
| Total | 2264 | 100.00 | 100.00 |

### 9.5.2 French

| Respondent | Documents | % Total | % Cumulative |
|---|---|---|---|
| NA | 240 | 10.77 | 10.77 |
| ALB | 19 | 0.85 | 11.62 |
| ARE | 22 | 0.99 | 12.61 |
| ARG | 1 | 0.04 | 12.66 |
| ARM | 5 | 0.22 | 12.88 |
| AUS | 31 | 1.39 | 14.27 |
| AZE | 9 | 0.40 | 14.68 |
| BDI | 3 | 0.13 | 14.81 |
| BEL | 43 | 1.93 | 16.74 |
| BGR | 21 | 0.94 | 17.68 |
| BHR | 29 | 1.30 | 18.99 |
| BLZ | 3 | 0.13 | 19.12 |
| BOL | 6 | 0.27 | 19.39 |
| BRA | 1 | 0.04 | 19.43 |
| CAN | 34 | 1.53 | 20.96 |
| CHE | 4 | 0.18 | 21.14 |
| CHL | 28 | 1.26 | 22.40 |
| COD | 24 | 1.08 | 23.47 |
| COL | 73 | 3.28 | 26.75 |
| CRI | 19 | 0.85 | 27.60 |
| CSK | 1 | 0.04 | 27.65 |
| DEU | 27 | 1.21 | 28.86 |
| DNK | 22 | 0.99 | 29.85 |
| EGY | 1 | 0.04 | 29.89 |
| ESP | 52 | 2.33 | 32.23 |
| FRA | 137 | 6.15 | 38.38 |
| FRA-GBR-USA | 7 | 0.31 | 38.69 |
| GBR | 107 | 4.80 | 43.49 |
| GNQ | 2 | 0.09 | 43.58 |
| GRC | 8 | 0.36 | 43.94 |
| GTM | 11 | 0.49 | 44.43 |
| HND | 44 | 1.97 | 46.41 |
| HUN | 1 | 0.04 | 46.45 |

| Respondent | Documents | % Total | % Cumulative |
| --- | --- | --- | --- |
| IND | 54 | 2.42 | 48.88 |
| IRN | 19 | 0.85 | 49.73 |
| ISL | 49 | 2.20 | 51.93 |
| ITA | 43 | 1.93 | 53.86 |
| JPN | 17 | 0.76 | 54.62 |
| KEN | 17 | 0.76 | 55.39 |
| LBN | 7 | 0.31 | 55.70 |
| LBY | 19 | 0.85 | 56.55 |
| MLI | 8 | 0.36 | 56.91 |
| MLT | 22 | 0.99 | 57.90 |
| MMR | 10 | 0.45 | 58.35 |
| MYS | 14 | 0.63 | 58.98 |
| NAM | 12 | 0.54 | 59.52 |
| NER | 14 | 0.63 | 60.14 |
| NGA | 38 | 1.71 | 61.85 |
| NIC | 75 | 3.37 | 65.22 |
| NLD | 44 | 1.97 | 67.19 |
| NOR | 33 | 1.48 | 68.67 |
| PAK | 42 | 1.89 | 70.56 |
| PER | 12 | 0.54 | 71.10 |
| PRT | 21 | 0.94 | 72.04 |
| QAT | 12 | 0.54 | 72.58 |
| RUS | 47 | 2.11 | 74.69 |
| RWA | 19 | 0.85 | 75.54 |
| SCG | 43 | 1.93 | 77.47 |
| SEN | 33 | 1.48 | 78.95 |
| SGP | 11 | 0.49 | 79.44 |
| SRB | 31 | 1.39 | 80.83 |
| SUN | 4 | 0.18 | 81.01 |
| SVK | 16 | 0.72 | 81.73 |
| SWE | 11 | 0.49 | 82.23 |
| TCD | 8 | 0.36 | 82.59 |
| THA | 30 | 1.35 | 83.93 |

| Respondent | Documents | % Total | % Cumulative |
|---|---|---|---|
| TUR | 20 | 0.90 | 84.83 |
| UGA | 35 | 1.57 | 86.40 |
| UKR | 4 | 0.18 | 86.58 |
| URY | 20 | 0.90 | 87.48 |
| USA | 200 | 8.98 | 96.45 |
| VEN | 13 | 0.58 | 97.04 |
| YUG | 6 | 0.27 | 97.31 |
| ZAF | 60 | 2.69 | 100.00 |
| Total | 2228 | 100.00 | 100.00 |

# 10 Verification of Cryptographic Signatures

This Codebook automatically verifies the SHA3-512 cryptographic signatures ('hashes') of all ZIP archives during its compilation. SHA3-512 hashes are calculated via system call to the OpenSSL library on Linux systems.

A successful check is indicated by 'Signature verified!'. A failed check will print the line 'ERROR!'

```r
# Function: Test SHA3-Hashes
sha3test <- function(filename, sig){
    sig.new <- system2("openssl",
                       paste("sha3-512", filename),
                       stdout = TRUE)
    sig.new <- gsub("^.*\\= ", "", sig.new)
    if (sig == sig.new){
        return("Signature verified!")
    }else{
        return("ERROR!")
    }
}


# Import Original Signatures
input <- fread(hashfile)
filename <- input$filename
sha3.512 <- input$sha3.512


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


# Print Results
testresult <- data.table(filename, sha3.512.result)

kable(testresult,
      format = "latex",
      align = c("l", "r"),
      booktabs = TRUE,
      col.names = c("File",
                    "Result"))
```

| File | Result |
| --- | --- |
| CD-ICJ_2023-05-07_EN_CSV_BEST_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN_CSV_BEST_META.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN_PDF_BEST_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN_PDF_BEST_MajorityOpinions.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN_PDF_ENHANCED_max2004.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN_PDF_ORIGINAL_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN_TXT_BEST_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN_TXT_EXTRACTED_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN_TXT_TESSERACT_max2004.zip | Signature verified! |
| CD-ICJ_2023-05-07_EN-FR_ANALYSIS.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_CSV_BEST_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_CSV_BEST_META.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_PDF_BEST_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_PDF_BEST_MajorityOpinions.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_PDF_ENHANCED_max2004.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_PDF_ORIGINAL_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_TXT_BEST_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_TXT_EXTRACTED_FULL.zip | Signature verified! |
| CD-ICJ_2023-05-07_FR_TXT_TESSERACT_max2004.zip | Signature verified! |
| CD-ICJ_2023-05-07_Source_Files.zip | Signature verified! |
| CD-ICJ_2023-05-07_UnlabelledFiles.zip | Signature verified! |

# 11  Changelog

The Changelog documents changes made to the data set. Versions are named according to the day on which the data creation process began.

## 11.1  Version 2023-05-07

- Full recompilation of data set
- Entire computational environment now version-controlled with Docker
- Scope extends up to case number 187: *Obligations of States in respect of climate change* (Advisory Opinion)
- Upgrade Tesseract OCR to version 5.3.1
- Upgrade OCR training data to "tesseract_best"
- Simplified config file
- Simplified function loading
- Ensure that debug mode only processes cases once
- Fix download manifest
- Update download function
- Contents of source ZIP file linked to Git manifest

## 11.2  Version 2022-09-07

- Full recompilation of data set
- Scope extends up to case number 183: *Jurisdictional Immunities* (Germany v Italy)
- Upgraded OCR to Tesseract 5.0.1
- CHANGELOG and README converted to external markdown files
- The ZIP archive of source files includes the TEX files
- Config file converted to TOML format
- All R packages are version-controlled with {renv}
- Data set creation process cleans up all files from previous runs before a new data set is created
- Removed redundant color from violin plots
- Added custom split instructions for the 2021-07-21 Order in the Amity Treaty case

## 11.3  Version 2021-11-23

- Initial Release

# 12 Strict Replication Parameters

```
## [1] "OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022)"
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.2 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] doParallel_1.0.17        iterators_1.0.14
##  [3] foreach_1.5.2            data.table_1.14.8
##  [5] textcat_1.0-8            quanteda.textplots_0.94.2
##  [7] quanteda.textstats_0.96.1 quanteda_3.2.4
##  [9] readtext_0.81            RColorBrewer_1.1-3
## [11] viridis_0.6.2            viridisLite_0.4.1
## [13] scales_1.2.1             ggplot2_3.4.1
## [15] rsvg_2.4.0               DiagrammeRsvg_0.1
## [17] DiagrammeR_1.0.9         magick_2.7.4
## [19] kableExtra_1.3.4         knitr_1.42
## [21] fs_1.6.1                 pdftools_3.3.3
## [23] stringr_1.5.0            mgsub_1.7.3
## [25] rvest_1.0.3              httr_1.4.5
## [27] RcppTOML_0.2.2           rmarkdown_2.20
##
## loaded via a namespace (and not attached):
##  [1] jsonlite_1.8.4    RcppParallel_5.1.7  askpass_1.1
##  [4] selectr_0.4-2     yaml_2.3.7          slam_0.1-50
##  [7] qpdf_1.3.0        pillar_1.8.1        lattice_0.20-45
## [10] glue_1.6.2        digest_0.6.31       tau_0.0-24
## [13] colorspace_2.1-0  htmltools_0.5.4     Matrix_1.5-1
## [16] pkgconfig_2.0.3   ISOcodes_2022.09.29 webshot_0.5.4
## [19] svglite_2.1.1     nsyllable_1.0.1     tibble_3.2.0
## [22] farver_2.1.1      generics_0.1.3      withr_2.5.0
## [25] cli_3.6.0         magrittr_2.0.3      evaluate_0.20
## [28] stopwords_2.3     fansi_1.0.4         xml2_1.3.3
## [31] tools_4.2.2       lifecycle_1.0.3     V8_4.2.2
## [34] munsell_0.5.0     compiler_4.2.2      proxyC_0.3.3
```

```
## [37] tinytex_0.44       systemfonts_1.0.4   rlang_1.0.6
## [40] grid_4.2.2         rstudioapi_0.14     htmlwidgets_1.6.1
## [43] visNetwork_2.1.2   labeling_0.4.2      gtable_0.3.1
## [46] codetools_0.2-18   curl_5.0.0          R6_2.5.1
## [49] gridExtra_2.3      dplyr_1.1.0         fastmap_1.1.1
## [52] utf8_1.2.3         fastmatch_1.1-3     stringi_1.7.12
## [55] Rcpp_1.0.10        vctrs_0.5.2         tidyselect_1.2.0
## [58] xfun_0.37
```

# References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2023. *Rmarkdown: Dynamic Documents for R.*

Analytics, Revolution, and Steve Weston. 2022. *Iterators: Provides Iterator Construct.* https://github.com/RevolutionAnalytics/iterators.

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, Jiong Wei Lua, and Jouni Kuha. 2023. *Quanteda.textstats: Textual Statistics for the Quantitative Analysis of Textual Data.* https://quanteda.io.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018a. "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.

———. 2018b. "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.

———. 2018c. "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.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2022. *Quanteda.textplots: Plots for the Quantitative Analysis of Textual Data.*

Corporation, Microsoft, and Steve Weston. 2022. *DoParallel: Foreach Parallel Adaptor for the Parallel Package.* https://github.com/RevolutionAnalytics/doparallel.

Dowle, Matt, and Arun Srinivasan. 2023. *Data.table: Extension of 'Data.frame'.*

Eddelbuettel, Dirk. 2023. *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.*

Garnier, Simon. 2021. *Viridis: Colorblind-Friendly Color Maps for R.*

———. 2022. *ViridisLite: Colorblind-Friendly Color Maps (Lite Version).*

Hester, Jim, Hadley Wickham, and Gábor Csárdi. 2023. *Fs: Cross-Platform File System Operations Based on Libuv.*

Hornik, Kurt, Patrick Mair, Johannes Rauch, Wilhelm Geiger, Christian Buchta, and Ingo Feinerer. 2013. "The textcat Package for *n*-Gram Based Text Categorization in R." *Journal of Statistical Software* 52 (6): 1–17. https://doi.org/10.18637/jss.v052.i06.

Hornik, Kurt, Johannes Rauch, Christian Buchta, and Ingo Feinerer. 2023. *Textcat: N-Gram Based Text Categorization.*

Iannone, Richard. 2016. *DiagrammeRsvg: Export Diagrammer Graphviz Graphs as Svg.* https://github.com/rich-iannone/DiagrammeRsvg.

———. 2022. *DiagrammeR: Graph/Network Visualization.* https://github.com/rich-iannone/DiagrammeR.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes.*

Ooms, Jeroen. 2022. *Rsvg: Render Svg Images into Pdf, Png, (Encapsulated) Postscript, or Bitmap Arrays.*

———. 2023a. *Magick: Advanced Graphics and Image-Processing in R.*

———. 2023b. *Pdftools: Text Extraction, Rendering and Converting of Pdf Documents.*

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

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

———. 2022a. *Rvest: Easily Harvest (Scrape) Web Pages.*

———. 2022b. *Stringr: Simple, Consistent Wrappers for Common String Operations.*

———. 2023. *Httr: Tools for Working with Urls and Http.*

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.*

Wickham, Hadley, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization.*

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.

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

———. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 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.*