DOI: 10.1111/jels.12313

RESEARCH NOTES AND DATASETS

JOURNAL OF EMPIRICAL LEGAL STUDIES

Introducing twin corpora of decisions for the International Court of Justice (ICJ) and the Permanent Court of International Justice (PCIJ)

Seán Fobbe

Lehrstuhl für Völkerrecht und Öffentliches Recht, Ludwig-Maximilians-Universität München, Munich, Germany

Correspondence
Seán Fobbe, Lehrstuhl für Völkerrecht und Öffentliches Recht, Ludwig-Maximilians-Universität München, Professor-Huber-Platz 2, D-80539 Munich, Germany.
Email: fobbe-data@posteo.de

Abstract

In this article I present the first two of a new series of open and high-quality international legal data sets: comprehensive, fully reproducible, human- and machine-readable open access collections covering one hundred years of case law of the primary judicial organs of the United Nations and the League of Nations: the *Corpus of Decisions: International Court of Justice* (*CD-ICJ*) and the *Corpus of Decisions: Permanent Court of International Justice* (*CD-PCIJ*). Each corpus is designed to capture in its entirety the published case law of its eponymous Court, including majority opinions (judgments, advisory opinions and orders), but also the minority opinions annexed to each decision (declarations, separate opinions and dissenting opinions). The corpora are enriched with useful metadata to enhance text-as-data research and enable stand-alone metadata analyses. While each corpus can stand on its own, the twin corpora are designed to be perfectly interoperable for the purposes of analyses that wish to treat the ICJ and PCIJ as a continuous entity. The most recent versions of the corpora will always be available open access at https://doi.org/10.5281/zenodo.3826444 (CD-ICJ) and https://doi.org/10.5281/zenodo.3840479 (CD-PCIJ).

INTRODUCTION

The quantitative analysis of international legal data is still in its infancy, a situation which is exacerbated by the lack of high-quality open access data sets. Particularly *corpora*, that is, data sets consisting primarily of (legal) texts, are in short supply. Most advanced data sets are held in commercial databases or, if held in UN or academic databases, are only available through functionally limited web interfaces that

generally do not permit more complex analyses than keyword or full-text searches with boolean operators. The published literature therefore often relies on creating bespoke data sets for each computational analysis, causing entry into this research field to be prohibitively difficult and time-consuming for most researchers, as well as creating difficulties in comparing and replicating research.

In this article I present the first two of a new series of open and high-quality international legal data sets: comprehensive, fully reproducible, human- and machine-readable open access collections covering one hundred years of case law of the primary judicial organs of the United Nations and the League of Nations: the *Corpus of Decisions: International Court of Justice* (*CD-ICJ*) and the *Corpus of Decisions: Permanent Court of International Justice* (*CD-PCIJ*).

The International Court of Justice (ICJ) is the primary judicial organ of the United Nations and one of the most influential courts in international law, "situated as it is at the apex of international tribunals" (Weeramantry, 1995, p. 345). Called the "World Court" by many, it is currently the only international court with general thematic jurisdiction. It was created by Chapter XIV of the Charter of the United Nations in 1945, commenced its activities in 1946 and celebrated its 75th anniversary in 2021.

While critics note the glacial pace of proceedings, low judicial output and sharply limited access to the Court for anyone but sovereign states and UN bodies, the opinions of the ICJ continue to have an outsize influence on the modern interpretation, codification and wider development of international law (Cassese, 2012, pp. 239–241). Even though its judgments only have *inter partes* binding force (ICJ Statute, art. 59) and advisory opinions none at all, the ICJ's influence is such that its majority opinions are regularly cited as authoritative pronouncements on the international law in force, particularly customary international law. International legal textbooks cover the workings and decisions of the Court *in extenso* (see, e.g., Shaw, 2017) and participation in international moot courts—especially the Philip C. Jessup International Law Moot Court Competition—without regular reference to and citation of the ICJ's decisions is unthinkable. To give but one example, the seminal judgment in the *Military and Paramilitary Activities in and against Nicaragua* case lastingly defined the law on the use of force in a manner that is widely accepted and extensively quoted in encyclopaedias of international law (Dörr, 2019).

The ICJ fulfils a similar function in the international legal system as its predecessor, the Permanent Court of International Justice (PCIJ). The PCIJ was the primary judicial organ of the League of Nations, the ill-fated predecessor of the United Nations, which operated from 1920 to 1946 and was officially dissolved in 1947 (Myers, 1948, p. 320). As the first international court with general thematic jurisdiction the PCIJ influenced international law in profound ways that are still felt today. Principles from the *Lotus* and *Factory at Chorzów* cases remain touchstones of international law, although their impact has been

qualified and reduced over the years. The Statute of the PCIJ formed the almost verbatim template for the later Statute of the ICJ.

Each corpus presented in this article is designed to capture in its entirety the published case law of its eponymous Court, including majority opinions (judgments, advisory opinions and orders), but also the minority opinions annexed to each decision (declarations, separate opinions and dissenting opinions). The corpora are enriched with useful metadata to enhance text-as-data research and enable stand-alone metadata analyses. While each corpus can stand on its own, the twin corpora are designed to be perfectly interoperable for the purposes of analyses that wish to treat the ICJ and PCIJ as a continuous entity. Together they cover the primary judicial output of the two Courts for a period covering one hundred years (1922–2021), although no case law was published in the years 1941–1946 due to the invasion and occupation of the Netherlands by German forces (1940–1945) and the time required for the newly formed ICJ to commence its activities in 1946.

The corpora are presented in English and French monolingual versions of equal quality. Each language version is available in more than half a dozen variants prepared for different use cases and target audiences, among them traditional scholars, legal practitioners and quantitative researchers. Each variant represents a different stage of refinement, beginning with a variant containing all original PDF documents as acquired from the Court's website (albeit with filenames enhanced for human and machine audiences), superseded by an enhanced PDF variant containing a vastly improved text layer created by neural networks, up to a structured and extensively documented CSV file with full texts and metadata ready for immediate use in advanced natural language processing applications.

Design, construction and compilation of both data sets are based on the principles of public access to public goods via 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 (Wilkinson et al., 2016).

THE REPLICATION STANDARD AND RELATED WORK

The replication standard is one of the key foundations of modern scientific research. The standard is met in the case that "sufficient information exists with which to understand, evaluate, and build upon a prior work" so that a "third party could replicate the results without any additional information from the author" (King, 1995, p. 444). This includes, in principle, an obligation to publish the underlying data, source code and information specifying the computational environment (Buckheit & Donoho, 1995; Donoho, 2010; Morin et al., 2012).

Exceptions may apply where copyright, data protection laws and the protection of human subjects take precedence. Unfortunately, full adherence to strict reproducibility is an ideal that most quantitative research does not live up to, particularly in international law.

International legal data are generally locked in databases, with functionally limited web interfaces permitting only certain queries on the information held within, such as keyword or full-text searches with boolean operators. Databases further present technical and legal obstacles to systematically acquiring the underlying data. Executing reproducible research queries is rarely possible, as database owners make no such commitment and may modify data and software without notice. Without access to the raw data it is also impossible to conduct complex computational analyses, such as network analysis or machine learning classification tasks.

This situation forces researchers to construct their own custom data sets for each individual research project. Bespoke corpora present special risks, as the high dimensionality of textual data means that similar analyses on dissimilar corpora may easily result in very different results, independent of theoretical reasons. The ability and willingness of researchers to invest in quality data collection may also be limited, as the final analysis tends to be the public and widely acknowledged end result of the research effort, not the data itself.

Open access data sets can significantly lighten the load that researchers must shoulder to comply with the replication standard. A replication-standard data set should ideally be available open access and free of charge, archived in a professional repository serviced by a reputable organisation (e.g., Zenodo/CERN or the Harvard Dataverse), assigned a globally unique identifier (e.g., a DOI), easily downloadable for offline use in an open format (e.g., CSV, XML, TXT) and contain a reasonable amount of relevant metadata. The FAIR Data Principles provide more detailed guidelines for all of these aspects (Wilkinson et al., 2016).

In this section I consider only corpora (i.e., data sets consisting primarily of text data) published open access to be comparable work to the CD-ICJ and CD-PCIJ. While there is a dearth of general-purpose international legal corpora, notable examples do exist, such as the *Text of Trade Agreements (TotA)* corpus (Alschner et al., 2017; Alschner et al., 2018) and the *PA-X* peace agreements corpus (Bell & Badanjak 2019; Bell et al., 2021). The *Electronic Database of Investment Treaties (EDIT)* permits the bulk download of its legal texts in different machine-readable formats (Alschner et al., 2021). The *United Nations General Debate Corpus* (Baturo et al., 2017) was developed for political science research, but can also shed light on the link between politics and international law.

Certainly the most impressive collection of international legal texts is available directly from the European Union, which provides full access to the entire set of EU legal documents held in its CELLAR database via the EUR-Lex web service, a SPARQL endpoint (machine-readable metadata) and a RESTful API (machine-readable content streams and metadata) (EU Publications Office, 2018). Unfortunately the intricacies of SPARQL, RDF graph databases and the EU Common

Data Model are such that very few researchers dare make use of the CELLAR database and instead resort to standard web scraping. This was the case even for the EU Horizon 2020-funded TRIGGER project, which constructed the *CEPS EurLex* corpus of EU regulations, directives and decisions via web scraping instead of direct CELLAR access (Borrett & Laurer, 2020). An R package designed to facilitate access to CELLAR was developed recently by Ovádek (2021), but it remains to be seen whether it will find widespread use.

As far as I am aware there exist no open access corpora that cover the judicial output of the ICJ or the PCIJ and are directly comparable to the corpora presented in this article. The most closely related work appears to be a text-as-data analysis of the citation practice of both the ICJ and the PCIJ, published by Alschner and Charlotin (2018) with most of the data drawn from www.worldcourts.com. The authors did not, to my knowledge, publish the corpus underlying their work in any publicly discoverable manner.

AVAILABILITY OF DATA AND SOURCE CODE

This article is the proverbial tip of the iceberg. Readers should take note of the tip, but direct their attention to the iceberg. Buckheit and Donoho famously paraphrased John Claerbout saying that "an article about computational result[s] is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result" (Buckheit & Donoho, 1995; Donoho, 2010, p. 385).[1] The data and source code (including dependency and version information) are the actual scholarship I present in this article. They are published open access and are securely and permanently archived with Zenodo, the scientific repository of CERN. Zenodo guarantees the availability of archived material for the lifetime of CERN, which has a defined research program for at least another 20 years (Zenodo., 2021).

Each data set and collection of source code is uniquely and persistently identified with Digital Object Identifiers (DOI), which may serve as a hyperlink to the data via the resolution services of www.doi.org. Each research item is linked with two DOIs: a *Version DOI*, uniquely identifying each version of the research item and a *Concept DOI*, uniquely identifying the overarching concept of the research item. The *Concept DOI* will redirect to the newest version. This ensures long-term stability of citations and permanent findability of the twin corpora. Researchers who aim to conduct reproducible research based on the CD-ICJ or CD-PCIJ are advised to cite the version number and the Version DOI. Readers who wish to inform others of the availability of these research items should

[1]The wording in Buckheit and Donoho (1995) is different from Donoho (2010). I have cited the version from Donoho (2010), as it is closer to modern standards of reproducibility.

share the Concept DOI, as this will ensure that the link provided will not need to be updated manually.

Data

Tables 1 and 2 provide key information on the public availability of the twin corpora. Each of the primary data sets is accompanied by an extensive codebook documenting all variants, variables and coding decisions. It further contains statistical summaries and visualisations of the data (measures of central tendency, frequency tables, distributions) to facilitate quick orientation.

Source code

The source code used to generate the full corpora, quality assurance test results and figures is published open access as well. Tables 3 and 4 provide key information on its public availability. Software is a critical part of computational research and should be cited like any other item (Smith et al., 2016). Significant parts of the construction process and source code, including the custom function

TABLE 1 Availability of the data set (CD-ICJ)

Attribute	Details
Full name	Corpus of Decisions: International Court of Justice
Acronym	CD-ICJ
Type	Stand-alone data set
Initial version	2021-11-23
Initial release	https://doi.org/10.5281/zenodo.3826445
Newest version	https://doi.org/10.5281/zenodo.3826444

TABLE 2 Availability of the data set (CD-PCIJ)

Attribute	Details
Full name	Corpus of Decisions: Permanent Court of International Justice
Acronym	CD-PCIJ
Type	Stand-alone data set
Initial version	1.0.0
Initial release	https://doi.org/10.5281/zenodo.3840480
Newest version	https://doi.org/10.5281/zenodo.3840479

TABLE 3 Availability of source code (CD-ICJ)

Attribute	Details
Full name	Source code for the "Corpus of Decisions: International Court of Justice"
Acronym	CD-ICJ-Source
Type	Source code and replication data
Initial version	2021-11-23
Initial release	https://doi.org/10.5281/zenodo.3977177
Newest version	https://doi.org/10.5281/zenodo.3977176

TABLE 4 Availability of source code (CD-PCIJ)

Attribute	Details
Full name	Source code for the "Corpus of Decisions: Permanent Court of International Justice"
Acronym	CD-PCIJ-Source
Type	Source code and replication data
Initial version	1.0.0
Initial release	https://doi.org/10.5281/zenodo.4136956
Newest version	https://doi.org/10.5281/zenodo.4136955

library, are revised and enhanced versions of my latest projects concerning German legal documents (Fobbe, 2022).

The source code includes comprehensive "Compilation Reports" generated during the creation run for each data set. Compilation Reports document the creation of each data set from first contact with the ICJ website to the final ZIP archives. These reports include the full source code and relevant computational results, timestamps, diagnostic data on the data sets and cryptographic hashes for each ZIP archive (SHA2-256 and SHA3-512). Hashes are also provided in separate CSV files and verified during compilation of each Codebook. Compilation Reports further specify the computational environment, that is, operating system and software used, including versions and options for R, all R packages and relevant system libraries.

The source code makes heavy use of advanced parallelisation techniques to achieve reasonable runtimes. On a Fedora Linux system utilising all 16 threads of an AMD Ryzen 7 3700X processor with 64 GB DDR4-3200 RAM and a fast SSD the compilation process for the CD-ICJ requires approximately 10.5 h. The compilation of the CD-PCIJ completes in approximately 1.5 h on the same hardware.

All manually coded elements are published alongside the source code in machine-readable CSV files or are hard-coded in the source code itself. They are programmatically integrated during the compilation process to ensure full reproducibility.

Most of the source code is written in the programming language *R* (R Core Team, 2021), with some calls made to system libraries where no suitable R package was available. The code is run on the Fedora distribution of the Linux operating system (Fedora Project, 2021). The package *data.table* (Dowle & Srinivasan, 2021) is used for advanced handling of tabular data. Computationally intensive tasks rely heavily on the parallelisation framework provided by the *doParallel* (Wallig et al., 2020), *foreach* (Wallig et al., 2020) and *iterators* (Wallig et al., 2020) packages.

Information from HTML pages is extracted with *rvest* (Wickham, 2021). HTTP interfacing makes use of *httr* (Wickham, 2020).

Conversion from PDF to TIFF to prepare for the OCR step is performed by the Linux system library *Magick* (ImageMagick Development Team, 2021). The generation of OCR text from TIFF image files and conversion to plaintext and PDF files with enhanced OCR files—one of the most critical components of the compilation process—is provided by the system library *Tesseract* (Tesseract Project, 2021). Further image processing capabilities are added by the R packages *DiagrammeR* (Iannone, 2020), *DiagrammeRsvg* (Iannone, 2016), *rsvg* (Ooms, 2021a) and *magick* (Ooms, 2021b). Simple extraction of the text layer from PDF files is performed with *pdftools* (Ooms, 2021c).

Plaintext files and filename metadata are ingested into R with *readtext* (Benoit et al., 2020). Advanced natural language processing capabilities are provided by *quanteda* (Benoit et al., 2018; Benoit, Watanabe, Wang, Nulty, et al., 2021), *quanteda.textstats* (Benoit, Watanabe, Wang, Lua, & Kuha, 2021) and *quanteda.textplots* (Benoit, Watanabe, Wang, Obeng, et al., 2021). Automated language classification testing relies heavily on *textcat* (Hornik et al., 2013; Hornik et al., 2020).

Reproducible reports are generated with *knitr* (Xie, 2014; Xie, 2015; Xie, 2021) and *kableExtra* (Zhu, 2021). Diagrams are built with *ggplot2* (Wickham, 2016; Wickham et al., 2020) and logarithmically scaled with the *scales* package (Wickham & Seidel, 2020). Visualisations utilise the *Color Brewer* (Brewer, 2021; Neuwirth, 2014) and *Viridis* palettes (Garnier et al., 2021).

Adding case short names and country codes uses *mgsub* (Ewing, 2020). For string padding I call *stringr* (Wickham, 2019). Hashes are calculated with *OpenSSL* (Open SSL Software Foundation, 2021). File operations are enhanced with *fs* (Hester & Wickham, 2020).

Public domain status and open licensing

Prior to constructing the data sets I communicated with the ICJ and provided full and detailed information about the scope of my research project (including sample data) and the intent to publish it. I obtained informed consent from the Registrar to publish the data set with Zenodo (CERN) and written clarification that decisions and opinions of the ICJ and PCIJ are not subject to copyright.

It is my sincere belief that the rule of law in the 21st century requires open legal data. To promote this goal I make available all research data described in this article under the most open legal regime possible. The law and the means to determine it should be a public good, available to all irrespective of financial means and social standing.

The two primary data sets (CD-ICJ and CD-PCIJ) I release into the public domain under *Creative Commons Zero 1.0 International Public Domain (CC0 1.0)* waivers. The source code and analysis data are published separately under *MIT No Attribution (MIT-0)* licenses.

The choice of MIT-0 for the source code and replication data is intended to address long-standing concerns in the open source community that the *Creative Commons Zero Public Domain* waiver and its explicit exclusion of software patents might present legal risks to users of free software. *Creative Commons Zero* is therefore currently not recommended for software by the Open Source Initiative (Open Source Initiative, 2021).

Updates and versioning

It is my intention to update the CD-ICJ up to twice per year, ideally every 6 months. If serious errors are discovered, an update will be provided at the earliest opportunity and a highlighted advisory issued on the Zenodo page of the current version. Minor errors will be documented in the GitHub issue tracker and fixed with the next scheduled release. The CD-ICJ is versioned according to the day the data was acquired from the website of the Court, in the ISO format *YYYY-MM-DD*. Its initial release version is 2021-11-23.

The CD-PCIJ will only be updated if errors are discovered, enhancements are developed or in the unlikely event that the Court publishes additional documents within the collection ambit of the data set (PCIJ Series A, B and A/B). The CD-PCIJ is versioned in the traditional software format of *MAJOR.MINOR.PATCH* and its initial release version is 1.0.0.

CORPUS OF DECISIONS: INTERNATIONAL COURT OF JUSTICE (CD-ICJ)

Description

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 and opinions of the ICJ for the years 1947 through 2021, up until 23 November 2021. It contains 2169 unique documents in English and 2160 content-equivalent documents in French, for a total of 4329 documents. Among these are judgments,

advisory opinions and orders, as well as their respective appended minority opinions (declarations, separate opinions and dissenting opinions). The CD-ICJ is designed to be complementary to and fully compatible with the *Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)*.

Table 5 contains summary statistics of key linguistic metrics for the CD-ICJ (without lowercasing or removal of any features). *Tokens* are, in principle, defined as arbitrary strings of characters bounded by whitespace and *types* are defined as unique tokens. Sentences are calculated according to *Unicode Standard: Annex No 29*. Figure 1 shows distributions of document length, calculated in characters, tokens, types and sentences. The bounded area indicates the relative prevalence of values within a certain range (violin plot), the overlaid boxplot shows the median (centre line), first and third quartiles (box boundaries) and 1.5 times the interquartile range (whiskers). Outliers are shown as transparent dots. The interquartile range and outliers are calculated prior to statistical transformation to log scale. Diagrams in this article only analyse the English version of the corpus. Equivalent diagrams for the French version are published as part of the data set.

A schematic representation of the compilation workflow for the CD-ICJ is shown in Figures 2 and 3. Readers are advised to refer to the Compilation Report or directly to the source code for a full accounting of the process, as the schematic and a narrative summary cannot convey the full complexity of the calculations performed.

Data collection

The scope of data collection for the CD-ICJ is defined by reference to a set of case numbers drawn from the General List of the Court and defined as a parameter in the configuration file. For version 2021-11-23 these are numbers 1 through 181, excluding case number 2, which appears to be unassigned. Materials for further case numbers have not been published to the best of my knowledge. Based on these case numbers the hyperlinks to all monolingual documents relating to judgments, advisory opinions and orders are extracted from the HTML webpages on www.icj-cij.org. The case pages of the ICJ follow a clear

TABLE 5 Linguistic metrics, English version (CD-ICJ)

Metric	Sum	Min	Quart1	Median	Mean	Quart3	Max
Characters	84,637,041	376	4436	16,373	39,021.23	44,409	744,471
Tokens	15,108,060	71	754	2895	6965.45	8068	142,584
Types	89,901	53	290	720	1050.94	1404	9995
Sentences	512,598	1	20	94	236.33	269	5642

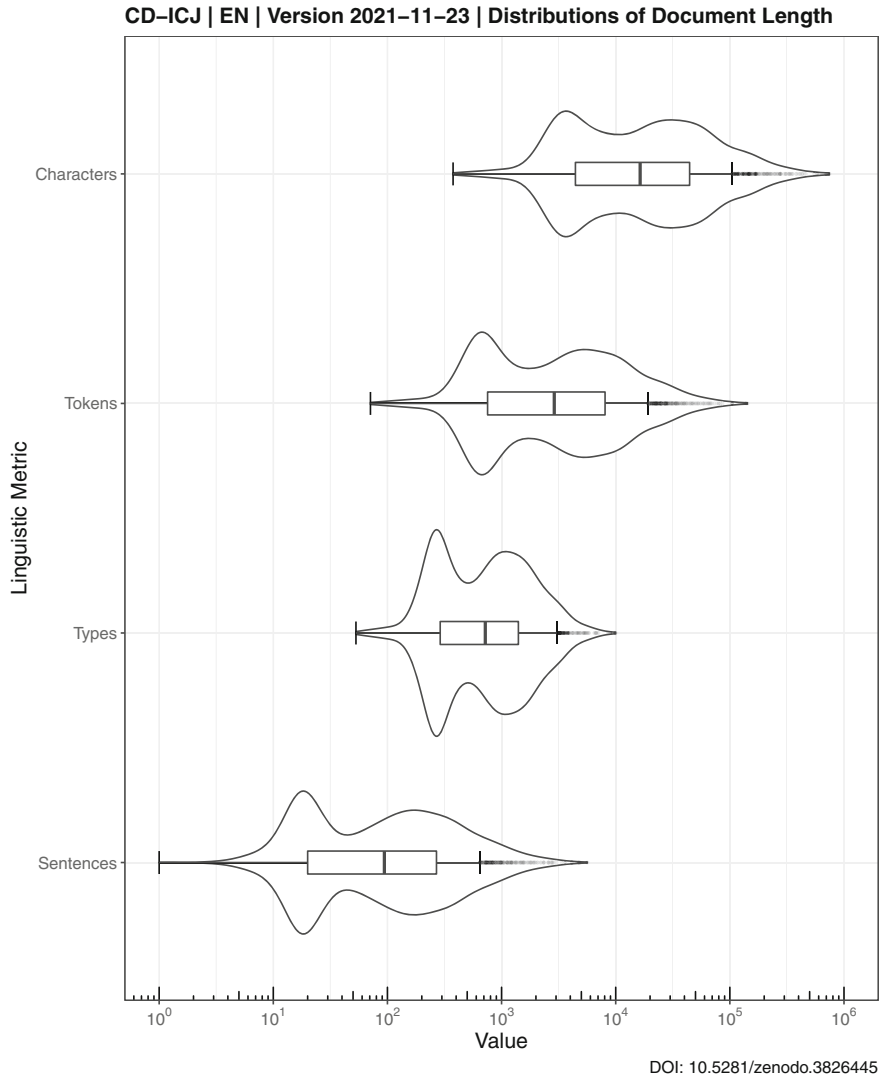

FIGURE 1 Distributions of document length in characters, tokens, types and sentences (CD-ICJ)

and easily iterated URL scheme, with little opportunity for error. Certain links to specific documents that were found to result in duplicate or error-prone documents during later automatic testing are removed during this step and, if possible, replaced by different links to functional bilingual documents that require further treatment. A debugging mode exists to limit the scope of data collection

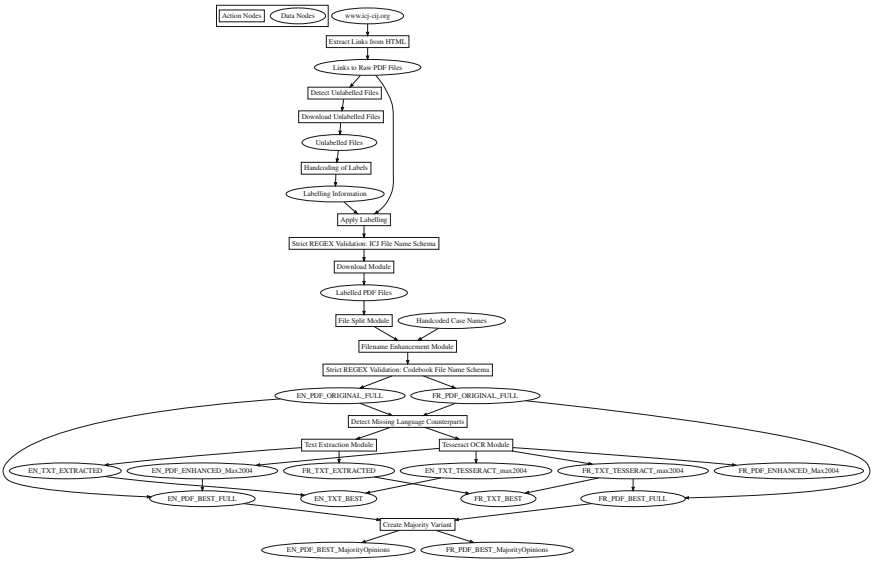

FIGURE 2 Workflow schematic, part 1 (CD-ICJ)

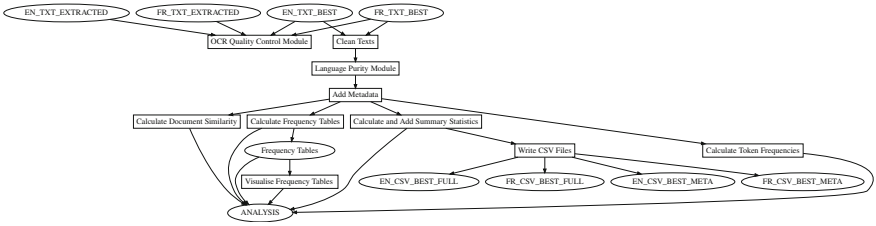

FIGURE 3 Workflow schematic, part 2 (CD-ICJ)

to a smaller and partially randomised set of case numbers in order to decrease compilation time and bandwidth usage during development.

From the full list of individual document links the program separates all links to unlabelled files and downloads these. Nearly all of the Court's documents are labelled according to a useful and machine-readable scheme, but 21 documents are only available with numeric filenames that do not appear to contain semantic information. These files were downloaded and, based on manual review, labelling information compliant with the labelling scheme of the Court was devised. This labelling information is published as a CSV file.

Labelling information is programmatically added at compilation time to the set of links, replacing numeric labels with hand-coded semantic labels. This

scheme is verified in a strict test with regular expressions, which halts compilation with an error upon failure. All files are then downloaded from the ICJ website, with the correct labels applied, resulting in the full collection of raw PDF files. An average of 1 s wait time between requests ensures respectful use of the Court's bandwidth. All downloaded files are checked against a master list, with the results documented in the Compilation Report. A very few bilingual PDF documents are split with custom instructions into their component languages to ensure monolinguality of the data.

The filenames of all monolingual documents are further enhanced with additional variables containing a shortened case name, stage of proceedings and country or entity codes for the applicant and respondent, in addition to being transformed into a more complex and enhanced machine-readable filename scheme described in the Codebook. Shortened case names were custom-coded to contain as much information of the original case name as possible and be easily searchable within a standard folder setting. This scheme is verified in a second strict test with regular expressions, which halts compilation with an error upon failure. The resulting monolingual PDF collections of English and French documents form the first variant of the data set. Missing language counterparts are documented, with three English and 12 French documents not being available on the website, even after manual review and search.

Missing documents in just one of the languages of the Court are usually due to a time lag between the publication of documents in the original language (often English) and the availability of official translations. This may affect analyses that rely on the most recent output of the Court, although the missing number of documents is usually small. Analysts should bear this in mind and either take into account document availability or wait for the next update of the CD-ICJ, which will most likely contain the relevant documents. Some very few documents were never published in the other language and it is unclear if or when they will be made available.

Conversion to machine-readable plaintext

Digital international court documents, specifically those of the ICJ and PCIJ, are usually only available as PDF documents. The conversion of PDF files into machine-readable plaintext representations of the content may occur in a number of ways, generally either through the use of optical character recognition (OCR) software or by extracting an existing text layer from a PDF file. This existing text layer, in turn, could have been created by a previous application of OCR or derive from the document being digitally created in the first place (born-digital).

Perfect plaintext is only available if the original PDF documents were born-digital. The quality of the processed plaintext from OCR can vary significantly, depending on the source material, quality of scanning, pre-processing steps, the quality of the OCR software and the options set (such as language profiles). Table 6 shows the difference (in an admittedly extreme example) between the

TABLE 6 Comparison of OCR quality: Original OCR and Tesseract OCR (CD-ICJ)

Original OCR	Tesseract 4
Officia1 citation Arrned Activities on the Territory of' the Congo (New Applic ltion:2002) (Dernocrutic Repuhlic qf' the Congo v. Rivundu), Provi.siona1 Meusures, Order qf I O J u k 2002, 1.C.J. Reports 2002, p. 21 9	Official citation: Armed Activities on the Territory of the Congo (New Application: 2002) (Democratic Republic of the Congo v. Rwanda), Provisional Measures, Order of 10 July 2002, 1 C.J. Reports 2002, p. 219

original OCR text layer provided by the Court and a modern text layer generated by *Tesseract* using LSTM neural networks.

From the downloaded PDF documents plaintext files are generated with two methods: (a) the simple extraction of the text layer and (b) the LSTM neural network engine of the OCR software *Tesseract*.

Simple text layer extraction is performed with the *pdftools* package for R. The result of this step is the "extracted" variant of the data set.

To define the set of documents to be processed with OCR I manually reviewed the original PDF files and extracted plaintext files to detect the approximate time when the ICJ switched to born-digital documents. I estimate this to be sometime during the year 2004, as the documents from the *Avena* and *Construction of a Wall* documents of March and July 2004 show visible signs of being scanned and OCR errors in the plaintext files, whereas the *Use of Force* documents of December 2004 appear to be clean, digital documents.

OCR with *Tesseract* is performed for all documents dated 2004 and earlier. PDF documents are converted to TIFF format at 300 dpi, including removal of the alpha channel and setting of a white background via the system library *Image Magick*. TIFF files are then processed with the LSTM neural network engine of *Tesseract* using English (primary) and French (secondary) training data for English documents and French (primary) and English (secondary) training data for French documents. Bilingual training data ensures that direct quotations in the non-dominant language of the document are correctly processed. The results are high-quality plaintext files and enhanced PDF files for all documents dated 2004 and earlier.

In both cases footnotes are treated as an integral part of the text flow. Their raw textual content is positioned following the continuous text of the page they are printed on in the original PDF document. This mainly concerns appended opinions. Footnotes do occur in majority opinions of the ICJ, but appear to be quite rare in comparison.

The TXT and PDF variants entitled "best" are composed of files generated by *Tesseract* for documents dated 2004 and earlier and a subset of the extracted variant of documents dated 2005 and later (which are assumed to be born-digital). From this ideal PDF variant a subset consisting of all majority

opinions forms a focused variant intended for practitioners. All of the steps up to this point are visualised in Figure 2.

OCR quality control

The process continues with Figure 3. The content of all TXT files and their filename metadata is ingested into R. The texts in the resulting R objects are cleaned by removing hyphenation between words and replacing certain special characters (that occur due to OCR errors) with their intended equivalents.

The extracted and *Tesseract* variants of the plaintext files for both languages dated 2004 and earlier are then subjected to OCR quality control. A standard quantitative pre-processing workflow is simulated, including tokenisation with removal of numbers, punctuation, symbols and separators, followed by converting all tokens to lowercase and removal of English and French stopwords based on standard lists supplied with the *quanteda* package for R. From these tokens objects I create document-feature matrices according to the bag-of-words model, where each unique token is a column, each document is a row and the values are simple unweighted counts of the number of times a feature occurs in a document. The terminology "features" in the *quanteda* ecosystem differs from the commonly used "terms" (as in "document-term matrix") to convey a greater level of generality, as features may also include stemmed terms, parts-of-speech terms, n-grams and syntactic dependencies (Benoit, Watanabe, Wang, Nulty, et al., 2021, Vignette).

The theoretical justification for this test is that OCR generally works at the level of individual character recognition and OCR errors result in tokens which, in the best case, are almost, but not quite, correct (e.g., the word "International" might be recognised as "1nternati0nal") and, in the worst case, complete gibberish. For humans copy and pasting quotations or reading the plaintext files such OCR errors represent a nuisance, for computational analyses based on the bag-of-words model or exact pattern matching via regular expressions such OCR errors constitute a major problem, as even a single incorrect character may cause a token to become very rare or unique. Across thousands of documents and pages these errors compound and confound quantitative analyses with high numbers of unique tokens (features) of very low frequency. Low-frequency features are, in many standard workflows, simply discarded via threshold pruning of the matrix during analysis, although this is a crude way of dealing with the problem and discards valuable information.

A smaller, but higher quality document-feature matrix is important in many natural language processing applications, particularly frequency analysis, pattern matching and topic modelling. Topic modelling profits especially from smaller matrix size, resulting in lower runtime and higher quality models. Traditional research based on enhanced PDF files will benefit from enhanced

keyword searches within a document and require less revision when copy and pasting for the purposes of direct quotation.

The reduction in features between the extracted and *Tesseract* variants serves as a lower boundary of corrected OCR errors. The true number is likely higher, as tokens correctly recognised by *Tesseract* that were routinely misrecognised before will form new, but semantically correct features. In other cases tokens recognised with perfect accuracy will be counted as known features, increasing the quality of the matrix. Table 7 compares the number of features for the extracted and *Tesseract* variants and shows that the *Tesseract* variant is demonstrably superior, with up to 50.19% reduction in matrix size for the critically important English version and an equally persuasive reduction of 42.54% for the French version.

Further processing and automated testing

Following OCR quality control the language purity of each monolingual variant is assessed with automatic language detection based on an analysis of n-gram patterns with the *textcat* package for R. Detection is limited to English and French, as these are the only primary languages in which the Court's documents are published. Prior unrestricted analysis did not detect any anomalies, except for the occasional mislabelling of French documents as Catalan by *textcat*. Including certain corrections made at the level of the hyperlink list and the file splitting module the analysis shows that all documents are indeed monolingual, with the language labels in the metadata matching the language of the content.

To the R objects a number of metadata variables are added, including the full case name, a dummy variable indicating minority opinions, geographical classifications of applicants and respondents according to the UN M49 standard, the DOIs of version and concept, the version number in the YYYY-MM-DD format and the Creative Commons Zero waiver. Further automated testing generates frequency tables for all categorical variables and visualisations for a selection, as shown by way of example in Figures 4–6. Linguistic metrics (counts of characters, tokens, types and sentences) are calculated, added to the data set, numerically summarised and visualised.

TABLE 7 Quality comparison of extracted and Tesseract plaintext variants dated 2004 and earlier (CD-ICJ)

Language	Extracted features	Tesseract features	Difference (abs)	Difference (rel)
English	115,816	57,686	−58,130	−50.19%
French	135,811	78,031	−57,780	−42.54%

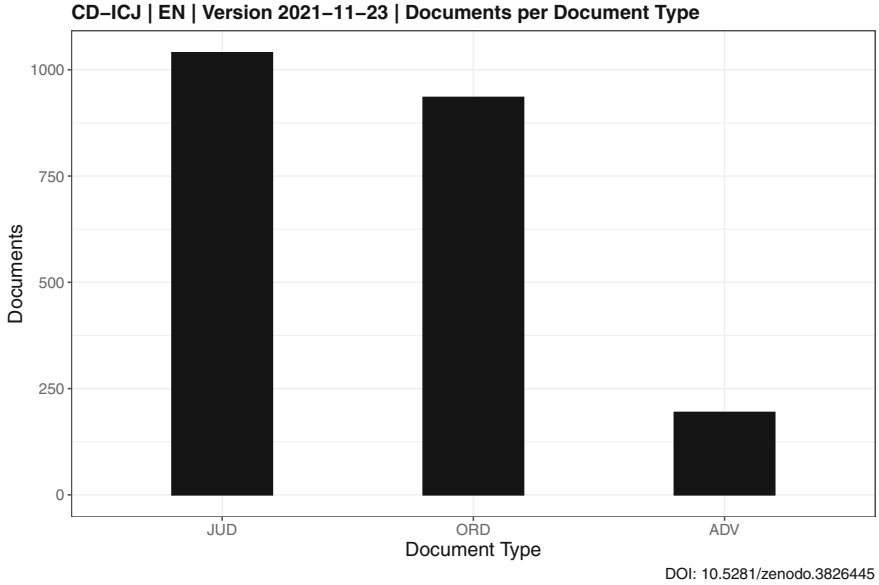

FIGURE 4 The number of documents per document type (CD-ICJ). The counts include both majority and minority opinions

Token frequencies and similarity statistics are calculated, reported and visualised. All results of automated tests are printed in the Compilation Report and most results are additionally stored as machine-readable CSV files and diagrams in the "analysis" archive included with the data set, published under the same public domain waiver. Identical tests are run for both the English and French versions of the data set.

The complete machine-readable data set and its metadata are then written as CSV files. The number and size of files for all PDF and TXT variants is counted and reported. File sizes are analysed, their distribution visualised and files larger than 10 MB reported. Due to the conversion at 300 dpi a very few of the resulting PDF files are quite large. This problem appears to be related to their initial file size and further research will be conducted to reduce file size.

Compilation of the data set concludes by creating ZIP archives for all variants, including source code and analysis data, and calculating cryptographically secure SHA2-256 and SHA3-512 hashes. Hashes are reported in the Compilation Report and written to a CSV file. The Codebook is compiled in the same manner and integrates a number of tables and diagrams generated during the corpus creation process to provide prospective analysts a fuller picture of the data set and its variables. The Codebook also verifies the cryptographic hashes to ensure that files have not corrupted.

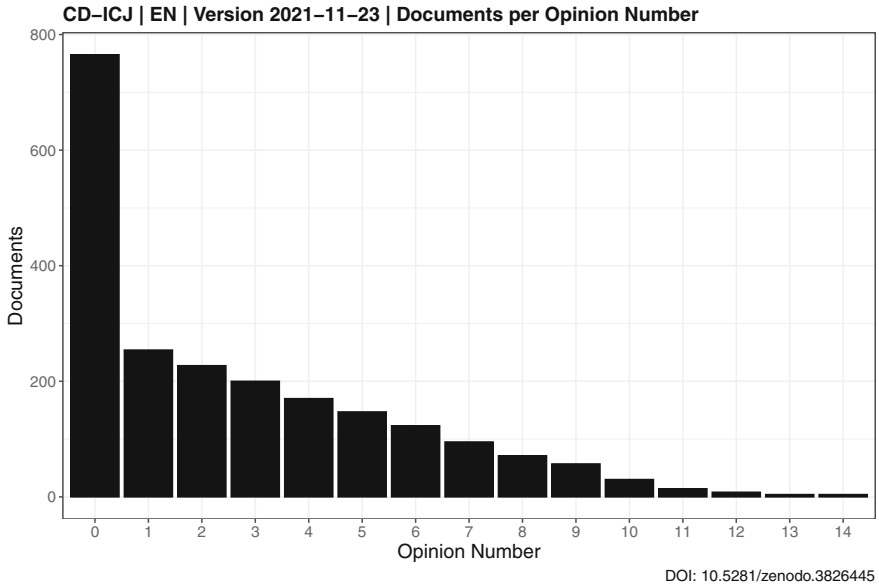

FIGURE 5 The number of documents per opinion number (CD-ICJ). Majority opinions are coded by the Court as "0." Minority opinions are numbered starting at "1" in ascending order to the maximum number of minority opinions appended to a given majority opinion

CORPUS OF DECISIONS: PERMANENT COURT OF INTERNATIONAL JUSTICE (CD-PCIJ)

Description

The *Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)* collects and presents for the first time in human- and machine-readable formats all documents of PCIJ Series A, B and A/B. It contains 259 documents in English, 261 documents in French and a multilingual collection of 265 original documents. The documents cover the years 1922 to 1940. Among these are judgments, advisory opinions, orders, a single "decision," appended minority opinions, annexes, applications instituting proceedings and requests for an advisory opinion. This data set is designed to be complementary to and fully compatible with the *Corpus of Decisions: International Court of Justice (CD-ICJ)*, which is also available open access.

Table 8 contains summary statistics of key linguistic metrics for the CD-PCIJ. Figure 7 shows distributions of document length, calculated in characters, tokens, types and sentences (without lowercasing or removal of any features). Metrics and diagrams are the same as described for the CD-ICJ above. Diagrams in this article only analyse the English version of the corpus. Equivalent diagrams for the French version are published as part of the data set.

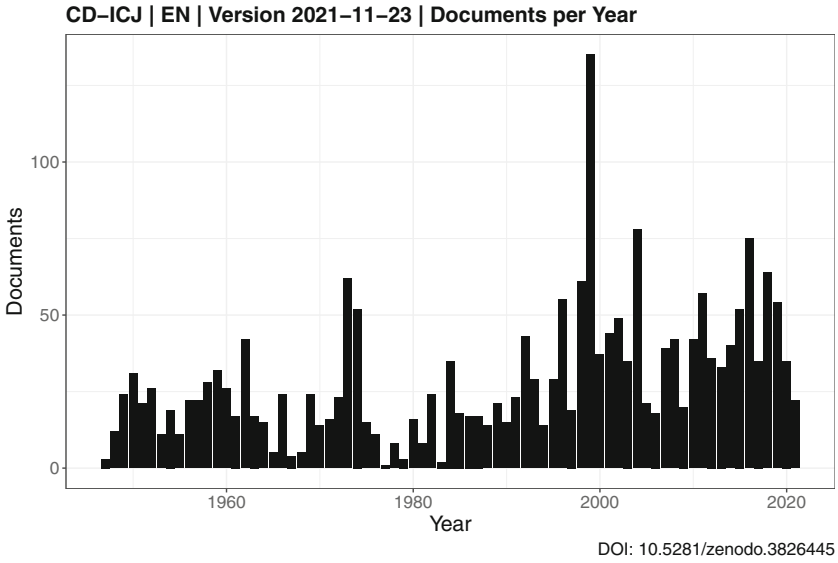

FIGURE 6 The number of documents per year (CD-ICJ). The counts include both majority and minority opinions. The abnormally high count for 1999 is due to a large number of documents emanating from the *Use of Force* cases, a known source of duplicates and quasi-duplicates. See section "Textual similarity and duplicate detection" on approaches to dealing with this challenge

TABLE 8 Linguistic metrics, English version (CD-PCIJ)

Metric	Sum	Min	Quart1	Median	Mean	Quart3	Max
Characters	6,830,889	310	6326.5	16,267	26,374.09	35,301.5	180,875
Tokens	1,298,030	62	1294.0	3107	5011.70	6726.0	33,652
Types	22,517	46	368.0	701	859.24	1174.0	3157
Sentences	38,266	6	38.0	97	147.75	203.5	821

A schematic representation of the compilation workflow for the CD-PCIJ is shown in Figure 8. Readers are advised to refer to the Compilation Report or directly to the source code for a full accounting of the process, as the schematic and a narrative summary cannot convey the full complexity of the calculations performed.

Data collection

All links to court documents provided on the web pages for PCIJ Series A, B and A/B on www.icj-cij.org are programmatically collected. After collection and selection of relevant document hyperlinks the original bilingual PDF documents are

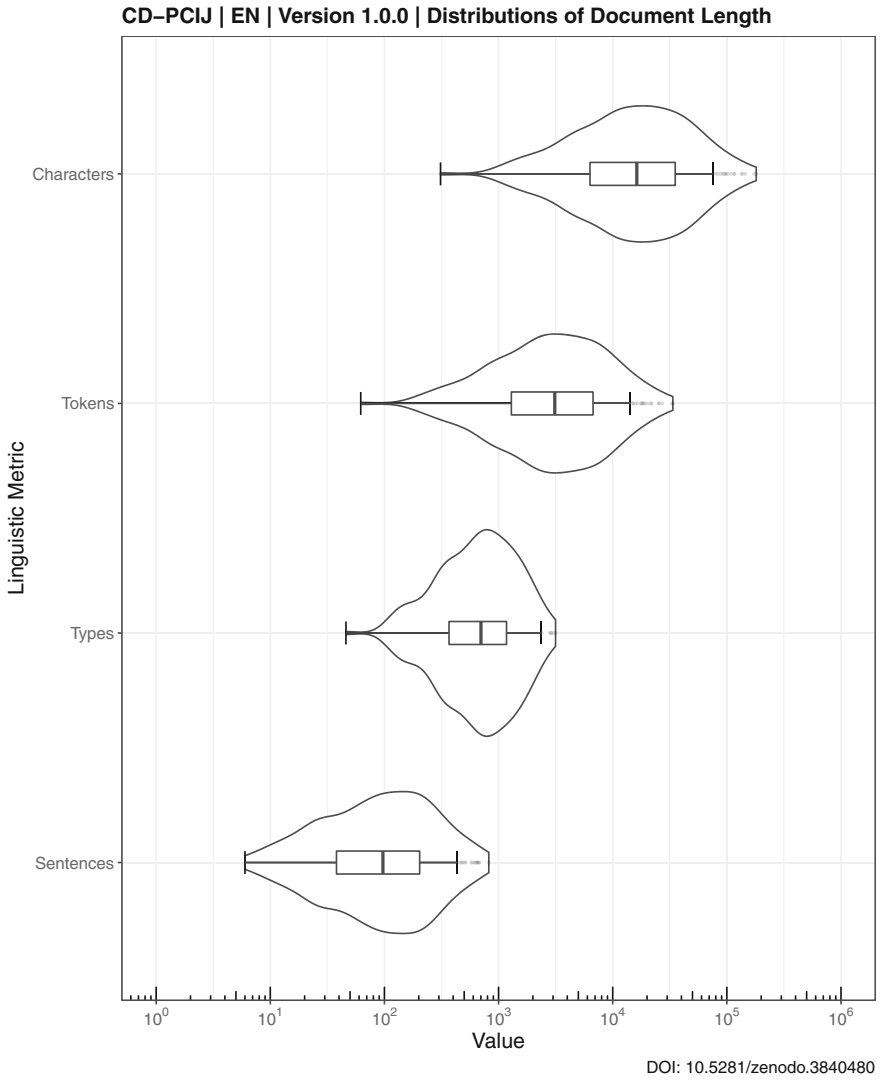

FIGURE 7 Distributions of document length in characters, tokens, types and sentences (CD-PCIJ)

downloaded. An average of 1 s wait time between requests ensures respectful use of the Court's bandwidth. All downloaded files are checked against a master list, with results documented in the Compilation Report.

The filenames of the PCIJ documents contain some semantic information, but not in any useful or consistent machine-readable format. I hand-coded filenames according to a scheme that closely tracks that of the final CD-ICJ scheme, with

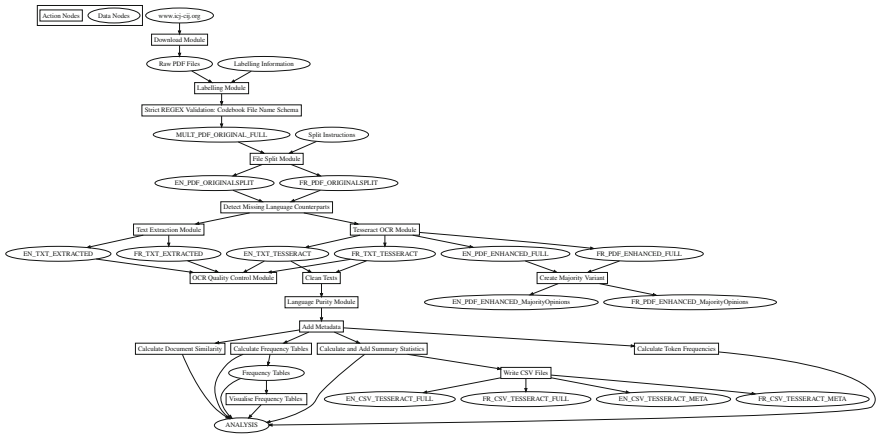

FIGURE 8 Workflow schematic (CD-PCIJ)

minor exceptions to allow for the separate coding of the Series (e.g., "A") and the Series number (e.g., "4"). These file names are stored in a CSV file and are then programmatically applied to the downloaded PDF files at compilation time. Downloaded and renamed files represent the first variant of multilingual original PDF files.

Splitting into component languages

Published documents of the PCIJ are nearly always bilingual (English and French), with monolingual documents a rarity. Some very few documents are available in German. To generate monolingual documents I manually reviewed every single page of every original document and created hand-coded split instructions for each original document, stored as a CSV file and executed programmatically at compilation time. Documents are processed according to one of four profiles:

1. English on even pages and French on odd pages
2. English on odd pages and French on even pages
3. Custom instructions due to non-standard alternation or inclusion of German
4. Do not split (monolingual documents)

The English and French versions produced in this step are stored separately and checked against each other to detect missing counterparts. The French version contains three surplus documents that are not available in English (although one is available in German); the English version contains one surplus document that is not available in French. Five originally monolingual

documents (two German, two French and one English) are stored in the multilingual variant. Strictly speaking one of the German documents (the *Danzig Courts* annex) is not a true original, as it was split from a bilingual file. However the quality of the document (scan and OCR) is original, so it is stored with the other originals to avoid creating another variant for a single document. All of these documents are annexes, that is, full texts or lists of documents submitted by the Parties to a case.

Conversion to machine-readable plaintext

From English and French PDF documents I generate plaintext files with two methods: (a) the simple extraction of the text layer and (b) the LSTM neural network engine of the OCR software *Tesseract*.

Simple text extraction is performed with the *pdftools* package for R. The result of this step is the "extracted" variant of the data set.

OCR with *Tesseract* is performed for all documents. PDF documents are converted to TIFF format at 300 dpi, including removal of the alpha channel and setting of a white background via the system library *Image Magick*. TIFF files are then processed with the LSTM neural network engine of *Tesseract* using English (primary) and French (secondary) training data for English documents and French (primary) and English (secondary) training data for French documents. Bilingual training data ensures that direct quotations in the non-dominant language of the document are correctly processed. The results are high-quality plaintext files and enhanced PDF files for all documents. From this enhanced PDF variant a subset consisting of all majority opinions forms a focused variant intended for practitioners.

In both cases footnotes are treated as an integral part of the text flow. Their raw textual content is positioned following the continuous text of the page they are printed on in the original PDF document. Footnotes appear to be rare in general.

OCR quality control

The content of all TXT files and their filename metadata is ingested into R. The texts in the resulting R objects are cleaned by removing hyphenation between words and replacing certain special characters (that occur due to OCR errors) with their intended equivalents.

The extracted and *Tesseract* variants of the plaintext files for both languages are then subjected to quality control. A standard quantitative pre-processing workflow is simulated, including tokenisation with removal of numbers, punctuation, symbols and separators, followed by converting all tokens to lowercase and removal of English and French stopwords from standard lists supplied with the *quanteda* package for R. From these tokens objects I create document-feature

matrices according to the bag-of-words model, where each unique token (i.e., a feature) is a column, each document is a row and the values are simple counts of the number of times a feature occurs in a document. The theoretical justification for this test is the same as for the CD-ICJ and described in more detail above.

Table 9 compares the number of features for the extracted and *Tesseract* variants and shows that the *Tesseract* variant is appreciably superior, with up to 30.73% reduction in matrix size. The French extracted variant is surprisingly close in matrix size to the French *Tesseract* variant. This is probably due to a French language OCR profile being used for all documents, with 261 documents being too few to accumulate many OCR errors, although 9.95% is still a notable difference that may affect computational results.

Further processing and automated testing

Following OCR quality control the language purity of each monolingual variant is assessed with automatic language detection based on an analysis of n-gram patterns with the *textcat* package for R. Detection is limited to English and French, as these are the only primary languages in which the Court's documents are published. Prior unrestricted analysis did not detect any anomalies, except the occasional mislabelling of French documents as Catalan. Including certain corrections made at the level of the file splitting module the analysis shows that all documents are indeed monolingual, with the language labels in the metadata matching the language of the content.

To the R objects a number of metadata variables are added, including the full case name, a dummy variable indicating minority opinions, geographical classifications of applicants and respondents according to the UN M49 standard, the DOI of version and concept, the version number in the YYYY-MM-DD format and the Creative Commons Zero waiver. Further automated testing generates frequency tables for all categorical variables and visualisations for a selection, as shown by way of example in Figures 9–11. Linguistic metrics (counts of characters, tokens, types and sentences) are calculated, added to the data set, numerically summarised and visualised. Token frequencies as TF and TF-IDF, as well as similarity statistics are calculated, reported and visualised. All results of automated tests are printed in the Compilation Report and most results are additionally stored as machine-readable CSV files and diagrams in

TABLE 9 Quality comparison of extracted and Tesseract plaintext variants (CD-PCIJ)

Language	Extracted features	Tesseract features	Difference (abs)	Difference (rel)
English	30,229	20,940	−9289	−30.73%
French	30,690	27,637	−3053	−9.95%

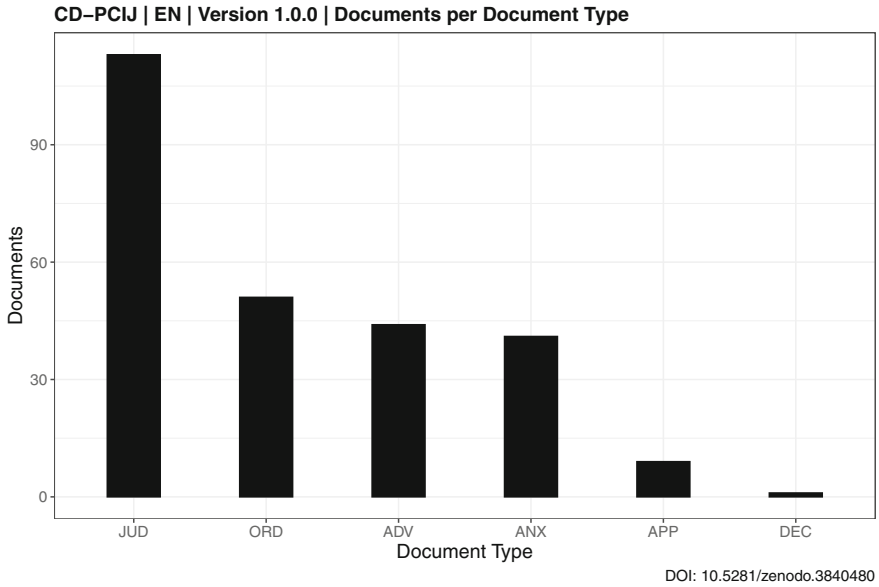

FIGURE 9 The number of documents per document type (CD-PCIJ). The counts for types JUD, ADV and ORD include both majority and minority opinions

the "analysis" archive included with the data set, published under the same public domain waiver. Identical tests are run for both the English and French versions of the data set.

The complete machine-readable data set and its metadata are then written as CSV files. The number and size of files for all PDF and TXT variants is counted and reported. File sizes are analysed, their distribution visualised and files larger than 10 MB reported, of which there are none. Compilation of the data set concludes by creating ZIP archives for all variants, including source code and analysis data, and calculating cryptographically secure SHA2-256 and SHA3-512 hashes. Hashes are reported in the Compilation Report and written to a CSV file. The Codebook is compiled in the same manner and integrates a number of tables and diagrams generated during the corpus creation process to provide prospective analysts a fuller picture of the data set and its variables. The Codebook also verifies the cryptographic hashes to ensure that files have not corrupted.

TEXTUAL SIMILARITY AND DUPLICATE DETECTION

The adoption of formally different decisions for each Applicant-Respondent pair in the course of the same substantive proceedings is a well-known and

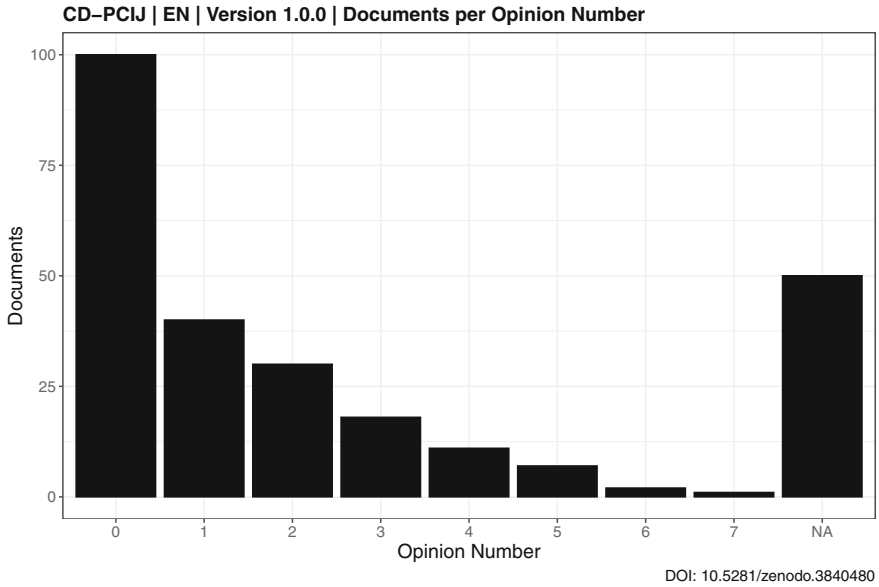

FIGURE 10 The number of documents per opinion number (CD-PCIJ). Majority opinions are coded as "0." Minority opinions are numbered starting at "1" in ascending order to the maximum number of minority opinions appended to a given majority opinion. Documents labelled "NA" are annexes and applications to institute proceedings

long-standing aspect of the ICJ's judicial practice. The *Use of Force* cases are particularly notable in this respect, having been assigned 10 formally different case numbers in the General List with eight of them resulting in nearly identical judgments. The other two were dismissed for lack of jurisdiction at an earlier stage of proceedings.

To provide a reproducible assessment of the number of duplicate documents in the CD-ICJ every compilation run includes the computation of pairwise similarity scores with the *correlation* method implemented in the *quanteda.textstats* package for R (Benoit, Watanabe, Wang, Lua, & Kuha, 2021). The documents were pre-processed in the same manner as for the OCR quality control tests: unigram tokenisation, removal of numbers, removal of special characters, removal of standard stopwords in English and French, as well as lowercasing. I investigated other pre-processing workflows without the removal of features or lowercasing, as well as bigrams and trigrams, but, based on my own qualitative assessment of the results, these performed no better and sometimes even worse than the standard workflow.

Figure 12 shows the number of duplicate documents as a function of document correlation similarity. For thresholds between 0.80 and 0.99, in intervals of 0.1, I further store both the individual IDs of duplicate

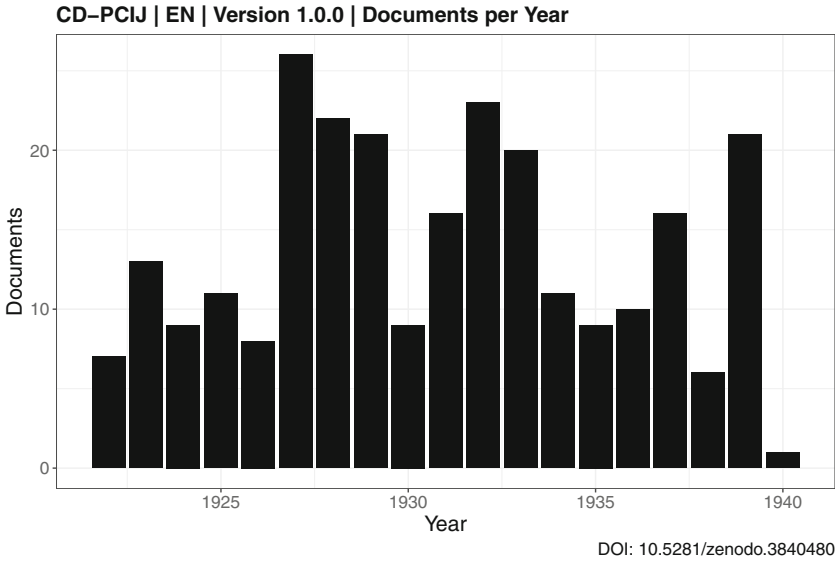

FIGURE 11 The number of documents per year (CD-PCIJ). The counts include all documents, regardless of type

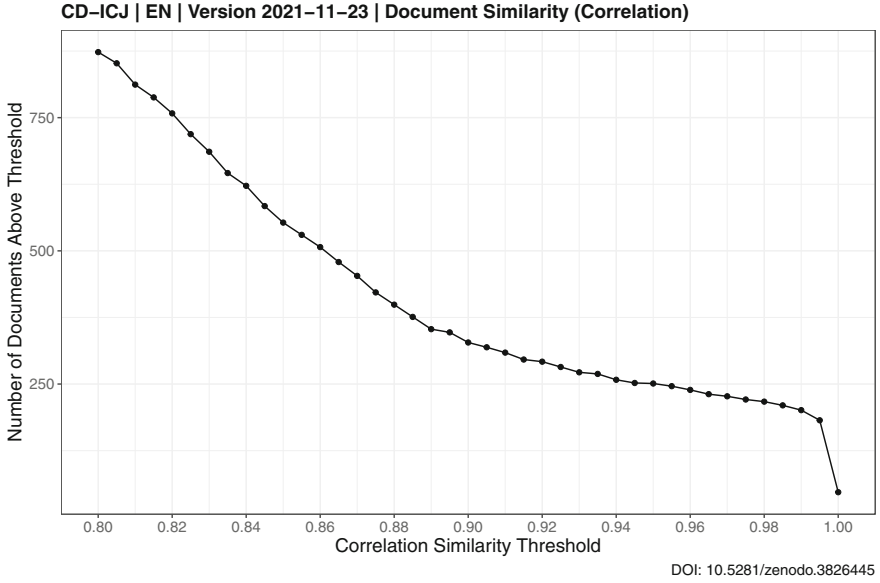

FIGURE 12 Duplicate documents as a function of correlation similarity (CD-ICJ)

documents to be excluded and the duplicate document pairs as CSV files (item 17 in the "analysis" archive). These serve two functions: one, to permit qualitative review of this method and two, to serve as convenient ready-to-use and reproducible sets of duplicates that analysts may wish to exclude from their own analyses.

Based on this similarity test the number of duplicate documents within the CD-ICJ is non-negligible for the purposes of computational analysis. For a correlation similarity threshold of 0.95 this equates to 251 duplicate documents in the English version, about 11.6% of the full corpus.

It is not my intention to present a definitive and recommended method of de-duplicating the CD-ICJ in this article. The manner of de-duplication will, however, substantially affect analytical results and should be made after careful consideration of a project's research goals, methodology and the subset of data to be analysed. The question of whether duplicate documents should be removed at all, the choice of similarity algorithm and the threshold for marking a document as duplicate remain matters of individual professional discretion until such time as further research can establish an authoritative methodology.

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 during the pre-processing step of an analysis. Nevertheless, my own qualitative review of the IDs of documents to be excluded showed that a correlation similarity threshold of approximately 0.95 appears to provide results that are generally in line with expectations based on my own experience in international law and what is known of the Court's work.

An alternative and reasonably efficient qualitative method of de-duplication would be to exclude from an analysis all documents assigned to certain case numbers which are assumed to be duplicate cases, based on a traditional assessment of the totality of the circumstances of each case, such as the case name, what has been reported of the proceedings and whether contentious parties overlap. The danger of this approach is that, even if proceedings are duplicate in principle, there may still be individual court documents that are unique or contain unique sections specific to the situation of a certain party, such as I describe for the CD-PCIJ below. De-duplicating the CD-ICJ based on a qualitative document-by-document assessment would in all likelihood require a prohibitive amount of human resources and, due to compound human error, may be a doubtful improvement over computational similarity analysis.

The same precautions need not be taken for analyses of the CD-PCIJ, as the only document pair with concerning similarity scores ($\geq$0.920) are the minority opinions of Judge Pessôa in the *Serbian Loans* and *Brazilian Loans* cases. These nonetheless differ slightly in a manner that is relevant to legal analysis. Figure 13 shows the same similarity analysis performed for the PCIJ.

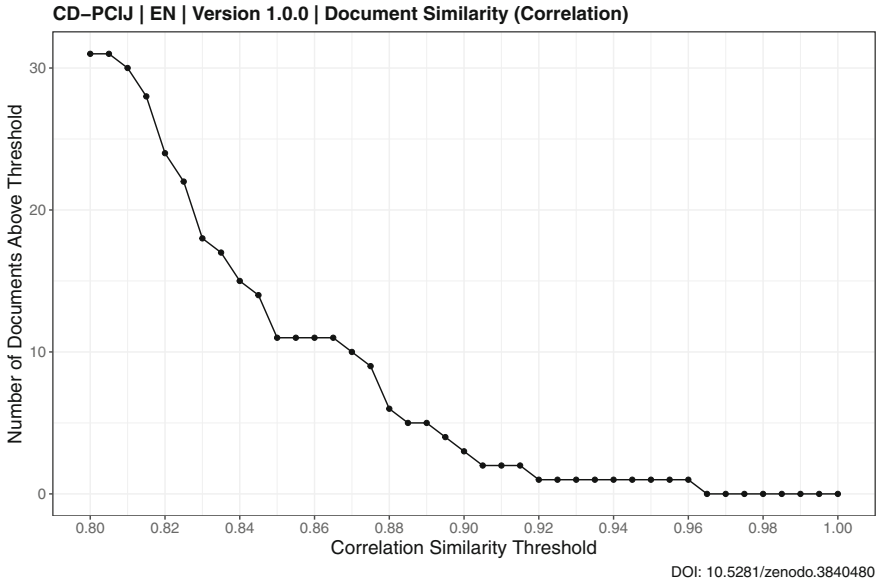

FIGURE 13 Duplicate documents as a function of correlation similarity (CD-PCIJ)

LIMITATIONS

Publication bias

Publication bias must be taken into account, but appears to have limited impact on the data. Given that the primary data source is the website of the ICJ, only those documents of the ICJ and PCIJ which have been digitised and published by the ICJ are included in the twin corpora. For a small number of older ICJ documents no English translations are available and a request for these files addressed to the Registry of the Court went without reply. Newly published documents are sometimes made available only in the original language, with translations published weeks or months later. Full bilinguality may therefore have a time lag of one or two versions of the CD-ICJ, but probably not more. For the CD-PCIJ a very few documents are available in only one language and it is unlikely that official translations will be forthcoming and retroactively added to the respective PCIJ Series. However, this affects only annexes, which contain full texts of documents or lists of documents provided by the Parties to a case, such as the text of the Treaty of Neuilly. The judicial output of the PCIJ appears to be published comprehensively in both languages.

OCR bias

Current *OCR* technology is not perfect. A significant fraction of the CD-ICJ and the entirety of the CD-PCIJ rely on OCR to convert scanned images to plaintext. While *Tesseract* in its fourth iteration with an LSTM neural network engine yields an astonishingly high OCR quality, errors still occur. The OCR output has not been corrected due to the scale of the undertaking and OCR bias must therefore be taken into account during analyses. However, the source code is built with the principle of modularity in mind. If better OCR technology is developed in the future, the *Tesseract* engine can easily be upgraded to a later version or be exchanged for different open source software.

Language mismatch

Language mismatches may have gone undetected. I define a language mismatch as the divergence of the dominant language in a document from the language declared in the metadata. The language metadata for the CD-ICJ was initially provided by the ICJ and was re-checked with automatic language analysis, while every single page of the CD-PCIJ was manually vetted and re-checked with automatic language analysis. Nonetheless, errors in automatic and manual review remain possible.

Language blurring

Language blurring is a fairly common occurrence in the documents presented here. I define language blurring as the inclusion of non-negligible amounts of text in a language other than the dominant language of the document. Historically, international law was practiced primarily in French. In the modern era it is practiced primarily in English. Quotations of French documents in English judgments or vice versa must be expected and may affect analyses to a greater or lesser degree. Quotations in other languages have not been taken into account at the OCR level to preserve performance, but may be included in future versions, particularly Spanish.

Limited segmentation

Texts have generally *not been segmented* beyond the differentiation between majority and minority opinions provided by the Court itself and my own manually added annotations of the stage of proceedings. However, documents converted with *Tesseract* preserve page breaks as the "form feed" (FF) special

character in the plaintext version, a fact that may be exploited by researchers willing to segment by page.

It may also be possible to craft a reasonable approximation of paragraph segmentation by splitting the texts according to the occurrence of at least two consecutive new lines, although this may introduce other issues (footnotes will be considered paragraphs, vertically spaced text, such as in headings, may also be misidentified as multiple paragraphs). Note that the ICJ only began providing official paragraph numberings in the late 1960s/early 1970s and paragraphs in PCIJ decisions were never officially assigned numbers.

Segmentation within a document (e.g., arguments of parties, reasoning of the Court, holding, disposition) is a desirable characteristic, but, in my professional opinion, will require extensive manual intervention to provide a high-quality result. I will seek to acquire the necessary funding and institutional support, and, if successful, provide segmentation in a future update.

Metadata

The metadata provided with the twin data sets is limited to high-quality and robust variables. This is intentional. Data sets—like programming languages and their associated packages—often represent a strict layer of abstraction in the workflow of downstream researchers. In other words, users tend to treat data sets and existing software as unquestioned ground truth. This entails a special responsibility on the part of data set and software authors to ensure their work is as robust as possible and to limit the potential for misunderstanding and misinterpretation. Of course, very few abstractions are perfect, as Spolsky (2002) noted in his *Law of Leaky Abstractions*: "All non-trivial abstractions, to some degree, are leaky."

Nonetheless, for this reason I have only included metadata I consider to be robust, that is, which could be automatically extracted with a high degree of accuracy based on persuasive methodologies or coded manually without compromising quality for speed. Metadata, in both cases, must also be open to automated testing to provide for a second—and most importantly independent—layer of verification.

In the future I hope to add additional metadata, such as the identities of presiding and participating judges, agents for the parties, linguistic annotations, named entity recognition, citation data and possibly summaries of the facts and law.

CONCLUSION

This article presented novel twin corpora of all published decisions and opinions of the ICJ and the PCIJ. Both corpora are permanently archived open access with Zenodo (CERN) and uniquely identified with DOIs.

Possible applications of the corpora are numerous, whether in traditional legal research or in the quantitative analysis of textual data (natural language processing).

Traditional researchers will profit from the enhanced text layer for the PDF files, which will make full text searches within documents more effective and allow for almost revisionless copy and pasting when directly quoting the Court's opinions. Researchers from developing nations, rural areas and other places with slow and/or unreliable internet connections may profit especially from the TXT variant of the corpora. The TXT files are rendered in smart ASCII plaintext which retains a semblance of the original PDF formatting and the associated page numbers. In this manner researchers and practitioners may acquire the full jurisprudence of the ICJ and PCIJ for offline work at a fraction of the original file size and in nearly the same quality as persons from developed nations. The corpora might also be useful to competitors in the annual Philip C. Jessup International Law Moot Court Competition, where this disparity in research opportunities is particularly pronounced. Practitioners of international law will find the collection of majority opinions useful in their daily work.

Quantitative researchers in academia, government, civil society and the private sector will be able to employ the full complement of advanced natural language processing techniques to analyse the jurisprudence of the two Courts, including frequency analysis, keyword-in-context analysis, co-occurrence analysis, similarity analysis, network analysis, topic models and supervised machine learning, to name but a few (for more ideas, see Alschner, 2019). The easy availability, public domain status and transparency-by-design open source nature of the corpora may even help to equalise differentials in international power relations by providing States and other international actors with limited resources access to cutting-edge technology (Deeks, 2020, pp. 596–597, 643–647). The bilingual nature and availability of equivalent texts in both English and French could aid the development of machine translation tools for international law. Quantitative analysts will further benefit from the *qualitative* uses the corpora can be put to in order to confirm or deepen their statistical findings. The persuasiveness of results may reach the greatest heights where "distant reading" (Moretti, 2000, 2013) based on statistical techniques and "close reading" of carefully chosen samples by skilled international lawyers unite.

ACKNOWLEDGMENTS

I would like to thank the Registrar of the International Court of Justice, Professor Philippe Gautier, for his kind permission to compile and publish the data sets. I am further indebted to Dr. Teoman Hagemeyer-Witzleb and two anonymous reviewers for helpful comments that much improved the paper and data sets. Funding from the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation) in the form of a PhD scholarship is gratefully acknowledged. Open access publication was funded through Projekt Deal.

REFERENCES

Alschner, W., Seiermann, J., & Skougarevskiy, D. (2017). *Text-as-data analysis of preferential trade agreements: Mapping the PTA landscape*. UNCTAD Research Paper, No. 5. UNCTAD/SER. RP/2017/5/Rev.1. https://unctad.org/en/PublicationsLibrary/ser_rp2017d5_en.pdf

Alschner, W. (2019). *The computational analysis of international law*. Ottawa Faculty of Law Working Paper, No. 2019–33. https://ssrn.com/abstract=3428762

Alschner, W., & Charlotin, D. (2018). The growing complexity of the International Court of Justice's self-citation network. *European Journal of International Law*, *29*(1), 83–112. https://doi.org/10.1093/ejil/chy002

Alschner, W., Elsig, M., & Polanco, R. (2021). Introducing the electronic database of investment treaties (EDIT): The genesis of a new database and its use. *World Trade Review*, *20*(1), 73–94. https://doi.org/10.1017/S147474562000035X

Alschner, W., Seiermann, J., & Skougarevskiy, D. (2018). Text of trade agreements (ToTA): A structured corpus for the text-as-data analysis of preferential trade agreements. *Journal of Empirical Legal Studies*, *15*(3), 648–666. https://doi.org/10.1111/jels.12189

Baturo, A., Dasandi, N., & Mikhaylov, S. (2017). Understanding state preferences with text as data: Introducing the UN general debate corpus. *Research & Politics*, *4*(2), 1–9. https://doi.org/10.1177/2053168017712821

Bell, C., & Badanjak, S. (2019). Introducing PA-X: A new peace agreement database and dataset. *Journal of Peace Research*, *56*(3), 452–466. https://doi.org/10.1177/0022343318819123

Bell, C., Badanjak, S., Beujouan, J., Forster, R., Epple, T., Jamar, A., McNicholl, K., Molloy, S., Nash, K., Pospisil, J., Wilson, R., & Wise, L. (2021). *PA-X Codebook. Version 5*. Political Settlements Research Programme, University of Edinburgh, Edinburgh. https://www.peaceagreements.org

Benoit, K., Obeng, A., Watanabe, K., Matsuo, A., Nulty, P., & Müller, S. (2020). *readtext: Import and handling for plain and formatted text files*. https://cran.r-project.org/package=readtext

Benoit, K., Watanabe, K., Wang, H., Lua, J.W. Kuha, J. (2021). *quanteda.textstats: Textual statistics for the quantitative analysis of textual data*. https://cran.r-project.org/package=quanteda.textstats

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (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, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A., Lowe, W., & Müller, C. (2021). *quanteda: Quantitative analysis of textual data*. https://cran.r-project.org/package=quanteda

Benoit, K., Watanabe, K., Wang, H., Obeng, A., Müller, S., Matsuo, A., & Fellows, I. (2021). *quanteda.textplots: Plots for the quantitative analysis of textual data*. https://cran.r-project.org/package=quanteda.textplots

Borrett, C., & Laurer, M. (2020). *The CEPS EurLex dataset: 142.036 EU laws from* 1952–2019 *with full text and 22 Variables (version V2)*. Harvard Dataverse. https://doi.org/10.7910/DVN/0EGYWY

Brewer, C. (2021). *Colorbrewer palette*. https://colorbrewer2.org/

Buckheit, J., & Donoho, D. (1995). WaveLab and reproducible research. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and statistics* (pp. 55–81). Springer. https://doi.org/10.1007/978-1-4612-2544-7_5

Cassese, A. (2012). It is high time to restyle the respected old lady. In A. Cassese (Ed.), *Realizing Utopia: The future of international law* (pp. 239–249). Oxford University Press.

Deeks, A. (2020). High-tech international law. *George Washington Law Review*, *88*, 574–653.

Donoho, D. (2010). An invitation to reproducible computational research. *Biostatistics*, *11*(3), 385–388. https://doi.org/10.1093/biostatistics/kxq028

Dörr, O. (2019). Use of force, prohibition of. In A. Peters (Ed.), *Max Planck encyclopaedia of public international law*. Oxford University Press.

Dowle, M., & Srinivasan, A. 2021. *data.table: Extension of data.frame*. https://cran.r-project.org/package=data.table
EU Publications Office. (2018). *Cellar: The semantic repository of the publications office*. European Union. https://doi.org/10.2830/064688
Ewing, M. (2020). *mgsub: Safe, multiple, simultaneous string substitution*. https://cran.r-project.org/package=mgsub
Fedora Project. (2021). *Fedora operating system*. https://getfedora.org/
Fobbe, S. (2022). *Open access code and analysis data by Seán Fobbe*. Zenodo https://zenodo.org/communities/sean-fobbe-code/
Garnier, S., Ross, N., Rudis, B., Sciaini, M., Camargo, A., & Scherer, C. (2021). *viridis: Colorblind-friendly color maps for R*. https://cran.r-project.org/package=viridis
Hester, J., & Wickham, H. (2020). *fs: Cross-platform file system operations based on libuv*. https://cran.r-project.org/package=fs
Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., & Feinerer, I. (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, K., Rauch, J., Buchta, C., & Feinerer, I. (2020). *textcat: N-gram based text categorization*. https://cran.r-project.org/package=textcat
Iannone, R. (2016). *DiagrammeRsvg: Export DiagrammeR graphviz graphs as SVG*. https://cran.r-project.org/package=DiagrammeRsvg
Iannone, R. (2020). *DiagrammeR: Graph/network visualization*. https://cran.r-project.org/package=DiagrammeR
ImageMagick Development Team. (2021). *ImageMagick*. https://imagemagick.org
King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452. https://doi.org/10.2307/420301
Moretti, F. (2000). Conjectures on world literature. *New Left Review*, *1*, 54–68.
Moretti, F. (2013). *Distant reading*. Verso Books.
Morin, A., Urban, J., Adams, P., Foster, I., Sali, A., Baker, D., & Sliz, P. (2012). Shining light into black boxes. *Science*, *336*(6078), 159–160. https://doi.org/10.1126/science.1218263
Myers, D. (1948). Liquidation of League of Nations functions. *American Journal of International Law*, *42*(2), 320–354. https://doi.org/10.2307/2193676
Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. https://cran.r-project.org/package=RColorBrewer
Ooms, J. (2021a). *rsvg: Render SVG images into PDF, PNG, postscript, or bitmap arrays*. https://cran.r-project.org/package=rsvg
Ooms, J. (2021b). *magick: Advanced graphics and image-processing in R*. https://cran.r-project.org/package=magick
Ooms, J. (2021c). *pdftools: Text extraction, rendering and converting of PDF documents*. https://cran.r-project.org/package=pdftools
Open Source Initiative. (2021). *Frequently answered questions*. https://opensource.org/faq#cc-zero
Open SSL Software Foundation. (2021). *OpenSSL: Cryptography and SSL/TLS toolkit*. https://www.openssl.org/
Ovádek, M. (2021). Facilitating access to data on European Union laws. *Political Research Exchange*, *3*(1), e1870150. https://doi.org/10.1080/2474736X.2020.1870150
R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Shaw, M. (2017). *International law* (8th ed.). Cambridge University Press.
Smith, A., Katz, D., & Niemeyer, K. (2016). Software citation principles. *PeerJ Computer Science*, *2*, e86. https://doi.org/10.7717/peerj-cs.86
Spolsky, J. (2002, November 11). *The law of leaky abstractions*. Joel on Software. https://www.joelonsoftware.com/2002/11/11/the-law-of-leaky-abstractions/

Tesseract Project. (2021). *Tesseract open source OCR engine*. https://github.com/tesseract-ocr/tesseract

Wallig, M., Microsoft Corporation, & Weston, S. (2020). *foreach: Provides foreach looping construct*. https://cran.r-project.org/package=foreach

Wallig, M., Microsoft Corporation, Weston, S. & Tenenbaum, D. (2020). *doParallel: Foreach parallel adaptor for the parallel package*. https://cran.r-project.org/package=doParallel

Wallig, M., Revolution Analytics, & Weston, S. (2020). *iterators: Provides iterator construct*. https://cran.r-project.org/package=iterators

Weeramantry, C. (1995). *Request for an examination of the situation in accordance with paragraph 63 of the court's judgment of 20 December 1974 in the nuclear tests (New Zealand v France) case (Dissenting Opinion of Judge Weeramantry)*. ICJ Reports 1995, pp. 317–362.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://ggplot2.tidyverse.org

Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations*. https://cran.r-project.org/package=stringr

Wickham, H. (2020). *httr: Tools for working with URLs and HTTP*. https://cran.r-project.org/package=httr

Wickham, H. (2021). *rvest: Easily harvest (scrape) web pages*. https://cran.r-project.org/package=rvest

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2020). *ggplot2: Create elegant data visualisations using the grammar of graphics*. https://cran.r-project.org/package=ggplot2

Wickham, H., & Seidel, D. (2020). *scales: Scale functions for visualization*. https://cran.r-project.org/package=scales

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman and Hall/CRC.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman and Hall/CRC. https://yihui.org/knitr/

Xie, Y. (2021). *knitr: A general-purpose package for dynamic report generation in R*. https://cran.r-project.org/package=knitr

Zenodo. (2021). *Zenodo—Research. Shared*. https://zenodo.org/

Zhu, H. (2021). *kableExtra: Construct complex table with kable and pipe syntax*. https://cran.r-project.org/package=kableExtra

How to cite this article: Fobbe, S. (2022). Introducing twin corpora of decisions for the International Court of Justice (ICJ) and the Permanent Court of International Justice (PCIJ). *Journal of Empirical Legal Studies*, *19*(2), 491–524. https://doi.org/10.1111/jels.12313