

Words of Power: Introducing a Comprehensive Corpus of UN Security Council Resolutions

Seán Fobbe, Lorenzo Gasbarri and Niccoló Ridi

10 April 2025

Contents

Abstract	2
1 Introduction	3
2 Overview of United Nations Data	4
3 A New Corpus of UN Security Council Resolutions	6
3.1 Design Principles	7
3.2 Content	7
3.3 Variants	8
3.4 Open Source Code	9
3.5 Public Domain Dedication	9
4 Dataset Design and Computational Methods	10
4.1 Data Engineering and the Scientific Method	10
4.2 Reproducible End-to-End Workflow	11
4.3 Testing and Reporting	13
5 Detailed Description of the Dataset	14
5.1 Text Corpus	14
5.2 Thematic Metadata	16
5.3 Geographic Metadata	17
5.4 Voting Metadata	19
5.5 Citation Network Data	20
5.6 Source Variables	23
6 Conclusion	24
7 References	24

Abstract

We introduce the *Corpus of Resolutions: UN Security Council* (CR-UNSC), a novel international legal-political dataset containing texts and metadata for all resolutions of the United Nations Security Council (UNSC) from resolution 1 (1946) through resolution 2722 (2024), as published by the UN Digital Library (UNDL).

The dataset provides full resolution texts in all six official UN languages, draft texts and full meeting records in English, as well as dozens of metadata variables, for a grand total of 82 variables in structured tabular CSV format, accompanied by extensive narrative documentation. In addition to the main tabular dataset the CR-UNSC is accompanied by specialized variants for network analysis and bibliography management.

The network analysis variant of the dataset offers citation network data in GraphML format containing all full-text citations to UN Security Council and to UN General Assembly resolutions associated with 54 node-level metadata variables and citation counts.

The bibliography management variant reproduces a significant part of the dataset metadata in BibTeX format ready for import into widely used reference managers such as Zotero, Jabref, Citavi and Endnote. The bibliographic database ensures inclusive access to the dataset for traditional researchers with no expertise in data analysis.

The computational workflow is engineered as a fully automated end-to-end extract-transform-load (ETL) data pipeline with citation analysis and NLP components, unit tests and extensive reporting. The declarative workflow is fault-tolerant, resumable and stores intermediate results in over one hundred individual checkpoints. The code is published open source under the GNU General Public License Version 3 (GPLv3).

We intend to update the corpus at least once per year. The most recent version of the CR-UNSC corpus will always be available open access via Zenodo through its Concept DOI located at <https://doi.org/10.5281/zenodo.7319783>.

1 Introduction

The United Nations Security Council (UNSC) is the most influential and powerful of the principal UN organs. Under Article 24 of the UN Charter (UNC) member states have conferred upon the UNSC the “primary responsibility for the maintenance of international peace and security”. Chapter VII of the UNC grants the Council unique powers to adopt binding resolutions ranging from economic sanctions to military force in order to manage threats to peace and react to armed conflict. In addition to its peace and security mandate, the UNSC plays a prominent role in the admission of new UN members, the appointment of the Secretary-General, the election of judges of the International Court of Justice (ICJ), the calling of special and emergency sessions of the General Assembly, as well as the amendment of the UNC and of the ICJ Statute.

Composed of five permanent and ten non-permanent members, the functioning of the UNSC is constrained by the political context in which it operates. During the Cold War, the complex political relationships between the permanent members significantly affected the capacity of the UNSC to address violations of international peace and security, with only 646 resolutions passed from 1946 to 1989. Since the 1990s, the activity of the UN Security Council has increased dramatically and produced 2721 resolutions by the end of 2023 despite high tensions among the permanent members. The length, complexity and thematic breadth of the resolutions has also increased, prompting calls to redefine it as a quasi-legislative body.

To facilitate future research on the UNSC we introduce the “Corpus of Resolutions: UN Security Council” (CR-UNSC), a novel international legal-political dataset containing texts and metadata for all resolutions of the United Nations Security Council (UNSC) from resolution 1 (1946) through resolution 2722 (2024) as published by the UN Digital Library (UNDL). The dataset provides full resolution texts in all six official UN languages, draft texts and full meeting records in English, as well as dozens of metadata variables, for a grand total of 82 variables in structured tabular CSV format, accompanied by extensive narrative documentation. In addition to the main tabular dataset, the CR-UNSC is accompanied by specialized variants for network analysis and bibliography management.

The network analysis variant of the dataset offers citation network data in GraphML format containing all full-text citations to UN Security Council and to UN General Assembly resolutions associated with 54 node-level metadata variables and citation counts.

The bibliography management variant reproduces a significant part of the dataset metadata in BibTeX format ready for import into widely used reference managers such as Zotero, Jabref, Citavi and Endnote. The bibliographic database ensures inclusive access to the dataset for traditional researchers with no expertise in data analysis.

2 Overview of United Nations Data

For many years the open availability of high-quality international legal data was sharply limited. Quality data was either kept secret in commercial databases or hidden away behind dysfunctional user interfaces to public databases. Even academic datasets routinely allow only limited web access and restrict raw data availability to hand-picked users. This situation forces many quantitative scholars to create their own bespoke datasets, an entry barrier to the field that is time-consuming and difficult to surmount, with unfortunate consequences for scientific reproducibility and the development of the field.

Recent years have seen more and more datasets relevant to international law being published, primarily by political scientists. This is a mixed blessing because the substantive and methodical focus of international relations scholars has often followed rather narrow trends, with much research on the UN being directed towards voting patterns, political speeches and institutional structure. For example, Bailey, Strezhnev, and Voeten (2017) recorded 75 articles in the period 1998–2012 that analyzed UN votes in terms of revealed State preferences. Fjelstul, Hug, and Kilby (2025) later counted an astonishing 86 journal articles just for the year 2023 which relied on UNGA voting data published by Bailey, Strezhnev, and Voeten (2017).

While international relations scholars have long pushed the boundaries of empirical research and engaged in significant data collection efforts to map the international system, international lawyers have rarely left the comfortable perch of doctrinal legal arguments. In a new century dominated by ever more complex problems it is critical for the field of international law to develop its own empirical research agenda, create its own datasets, formulate specifically legal research questions and join hands with political scientists and empirical legal scholars to avoid duplication of effort.

Modern advances in natural language processing (NLP) have made quantitative approaches to textual data feasible and astonishingly productive (Alschner 2021). Large language models (LLM) have extended the frontier of what was previously thought possible and placed many interesting legal applications within reach, albeit with significant challenges in terms of ethics, security and equitable sharing of economic benefits. Sustained data collection efforts have further made the voluminous textual output of the UN, its organs, specialized organizations and the wider UN system more accessible than ever before.

Datasets composed primarily of texts are called *corpora*. Corpora are of special interest for two reasons: the texts produced in the environment of an international organization not only reproduce the actual practice of the organization and its stakeholders, but also its *intended* practice. Where voting data shows a snapshot of revealed preferences in the present on specific issues, textual data can show the past, present and possible futures in multidimensional context. This complexity of textual data is both blessing and curse. While the richness of political-legal texts is helpful for the study of many challenging questions, lack of clarity, questionable truth value and hidden political compromises can easily mislead careless analysts. In addition to these general features, UNSC resolutions, especially if adopted under

Chapter VII, have special legal value with binding impact on specific situations and may even influence the general course of international law.

The UN General Assembly has seen a number of significant research advances and notable corpora published in recent years. Baturo, Dasandi, and Mikhaylov (2017) introduced and Jankin, Baturo, and Dasandi (2024) extended the *UN General Debate Corpus* to document political speeches given before the UNGA during annual General Debates from 1946 to 2023. Mesquita and Pires (2024b) recently published a collection of all UNGA resolution texts for the years 1946 to 2019, partially based on the list by Warntjen (2016). The *UN General Assembly Sponsorship Dataset* (Seabra and Mesquita 2022) was recently extended by Mesquita (2024) to include the texts of draft resolutions (2000-2020).

Widely used UNGA voting data has been published for many years by Voeten et al. and now covers the period 1946 to 2023 (Voeten, Strezhnev, and Bailey 2009; Bailey, Strezhnev, and Voeten 2017). Fjelstul, Hug, and Kilby (2025) created an updated database on UNGA decision-making (UNGA-DM) with 7,312 decisions (recorded votes), 1,050,340 country-level votes and 116,967 non-votes by UN members from 1946 to 2023. Citation analysis of UNGA resolutions has been conducted as early as Bleicher (1969). Another notable UN resource is the UN Parallel Corpus (Ziems, Junczys-Dowmunt, and Pouliquen 2016) of texts in all six official UN languages intended for machine translation research.

The texts of the International Court of Justice (ICJ) have seen rising interest as well, with Alschner and Charlotin (2018) constructing an unpublished corpus of ICJ judgments and analyzing its internal citation network. Fobbe (2022) democratized research on ICJ and PCIJ judgments and opinions with the publication of the *Corpus of Decisions: International Court of Justice (CD-ICJ)* and the *Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)* which comprehensively document the jurisprudence of the two World Courts from 1922 to 2023.

Corpora of relevance to peace and conflict studies are slowly gaining ground as well. Pomeroy (2017) published a corpus of 888 speeches given in the UN Committee on the Peaceful Uses of Outer Space between 1961 and 1993. Bell and Badanjak (2019) constructed *PA-X*, a corpus of peace agreements. Clayton, Dorussen, and Böhmelt (2021) collected 469 UN peace initiatives in the *United Nations Peace Initiatives (UNPI)* dataset (1946–2015). The *Peacekeeping Mandates (PEMA)* dataset by Di Salvatore et al. (2022) codes 41 different tasks in 27 peacekeeping operations in Africa authorized and amended in 365 UNSC resolutions.

Research and data availability on the UN Security Council (UNSC) has been far more limited compared to the UNGA. A dataset by Beardsley (2013) coded metadata for UNSC resolutions from 1946 to 2008 in terms of adoption date, affected States and actions taken. (Benson and Tucker 2021, 2022a) coded dozens of metadata variables for UN peacekeeping resolutions and (Benson and Tucker 2022b) did the same for civil war related resolutions, covering the period 1946-2015.

The *UN Security Council Debates* corpus by Schoenfeld et al. (2019) containing 82,165 speeches for the period 1995–2020 is a welcome exception and one of the few

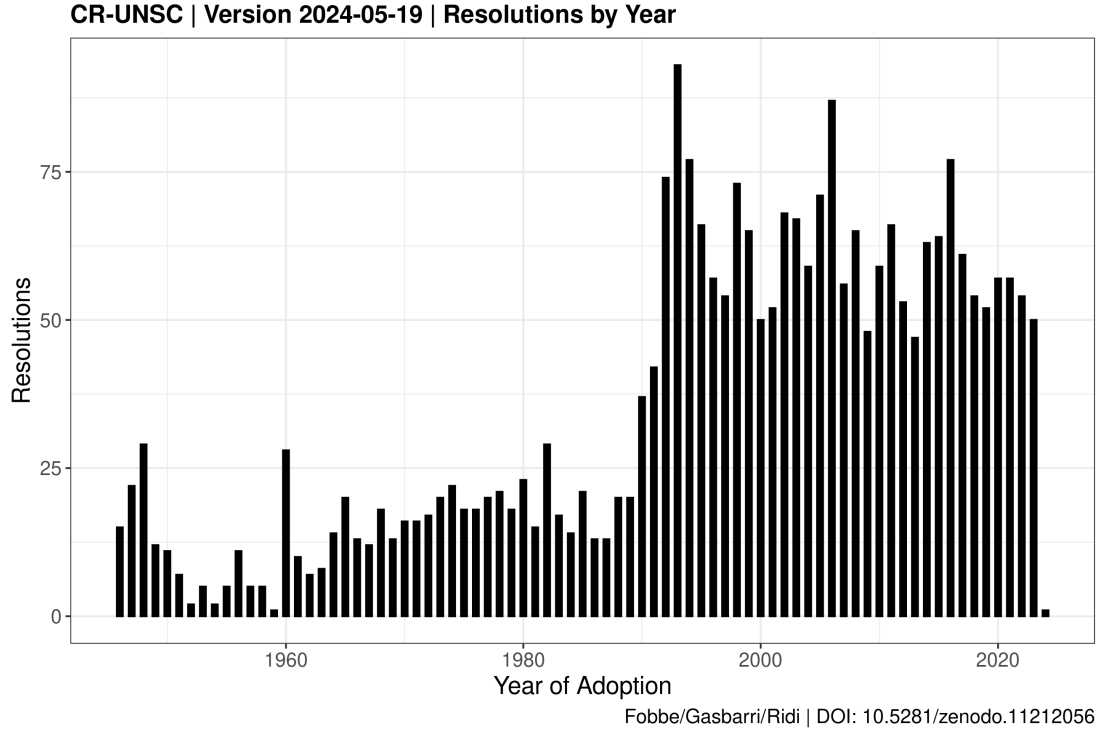


Figure 1: Count of UNSC resolutions adopted by year contained in the CR-UNSC dataset.

comprehensive UNSC corpora openly published. Badache, Hellmüller, and Salayme (2022) specifically collected and published open access UNSC speeches (1991-2020) related to peace given by the P-5, Brazil, South Africa and Turkey in the *UN Security Council Peace-Related Speeches* (UNSCPeaS). Mesquita and Pires (2024a) claim to have produced a corpus of 2,503 UNSC resolutions for the period 1946–2019, but the corresponding Harvard Dataverse entry Pires (2024) does not contain the texts of UNSC resolutions, only metadata.

3 A New Corpus of UN Security Council Resolutions

We present the *Corpus of Resolutions: UN Security Council (CR-UNSC)* to fill this void in UNSC research data availability. The CR-UNSC collects and structures in human- and machine-readable form 2722 (1946–2024) published resolutions of the United Nations Security Council (UNSC) in all six official UN languages (Arabic, Chinese, English, French, Spanish and Russian) with detailed metadata for a grand total of 82 variables. Drafts and meeting records are included in English, as this is the primary language of international drafting (Tomuschat 2017, 203). Figure 1 shows the count of resolutions by year contained in the corpus.

Specialized variants of the dataset provide citation data and a bibliographic database compatible with a wide range of end-user applications. We use the UN Digital

Library (UNDL) as the primary data source, augmented with extensive hand-coded data and automatic error detection and correction.

To the best of our knowledge, no comparable corpus of UNSC resolutions is available open access at the time of writing. The first version of the CR-UNSC was made available open access on Zenodo on 5 May 2024 (Version 2024-05-03), prior to any comparable UNSC corpus. This paper discusses the most recent Version 2024-05-19, which contains some minor bug fixes and adds new features. It builds on previous work by Ridi and Gasbarri (2023) to conduct citation network analyses on a corpus of UNSC resolutions. Table 1 provides an overview of how to access the corpus.

Table 1: Availability of the CR-UNSC Dataset.

Attribute	Detail
Full Name	Corpus of Resolutions: UN Security Council
Acronym	CR-UNSC
Initial Version	2024-05-03
Initial Release	https://doi.org/10.5281/zenodo.7319781
Newest Version	https://doi.org/10.5281/zenodo.7319780
License	CC Zero 1.0 Universal Public Domain Dedication

3.1 Design Principles

The CR-UNSC is designed for a) comprehensive coverage of the subject matter, b) generality of application and c) full conceptual and computational reproducibility.

Design, construction and compilation of this data follow a principled approach to open licensing, strict transparency and full scientific reproducibility. The *FAIR Guiding Principles for Scientific Data Management and Stewardship* (Findable, Accessible, Interoperable and Reusable) underpin both design and manner of publication (Wilkinson et al. 2016). We aim to observe FAIR principles in relation to the source code as well (Lamprecht et al. 2020).

3.2 Content

In version 2024-05-19 the CR-UNSC contains all UNSC resolutions from 1 (1946) through 2722 (2024) documented by 82 variables in seven substantive groups (ID, text, thematic, voting, geographic, sources, meta). The size of the text corpus in English is approximately 3.7 million tokens.

1. *ID* variables provide identifying symbols and temporal information (date and year) for the adopted resolution, the draft, its meeting record and various related documents.
2. *Text* variables offer the full text of the adopted resolution in all six UN languages, plus the text of the associated draft and meeting record in English. We further provide pre-computed statistical measures of the length of English resolution texts as characters, tokens, types, and sentences.

3. *Thematic* variables contain different thematic classifications of UNSC resolutions help identify relevant texts. They include the presence of Chapter VI/VII/VIII invocations, mentions of key terms (human rights, threat to peace/breach of peace/aggression), a title, a UNDL-curated summary, topics, subjects and notes.
4. *Vote* variables give counts of yes, no, abstaining and non-voting members, the total number of members at the time, the date of the vote and a breakdown of votes by State.
5. *Geographic* variables indicate mentions of countries in the full text of UNSC resolutions. They are provided as ISO-3 shortcode and ISO name, M49 code, M49 region, M49 intermediate region and M49 subregion.
6. *Source* variables collect URLs for all full text resolutions, drafts and meeting records in all six official UN languages, as well as the record pages for resolution, draft and meeting record.
7. *Meta* variables store information about the dataset itself, such as license, Version DOI, Concept DOI and dataset version.

3.3 Variants

The CR-UNSC is published in several variants to cover different quantitative, qualitative and mixed-method use cases.

The *CSV files* are the primary variant intended for quantitative analysis. The main CSV file contains the full spectrum of 82 variables. A much smaller metadata-only CSV file contains the same complement of variables, save only the texts.

The *TXT variants* are a possible alternative for quantitative researchers who are unaccustomed to using CSV files. The primary TXT variant collects all English language texts of UNSC resolutions in the best quality available to the dataset authors, equivalent to the “text” variable in the main CSV file. Resolutions 1 through 899 in the are expert-revised OCR texts, resolutions from 900 onwards are born-digital.

Another TXT variant collects all TXT files as they are produced during dataset compilation (OCR and extracted) for reproducibility purposes. Qualitative researchers with slow internet connections or who wish to save disk space should consider one of the TXT variants, as they still provide a reasonable visual approximation of the original documents, but offer the advantage of drastically reduced file size.

The three *PDF variants* (resolutions, drafts and meeting records) are intended to assist qualitative research and enable mixed-methods approaches. The first PDF variant republishes the original PDF files of all resolutions in all six languages as provided by the UNDL. It further contains resolutions 1 through 899 with an enhanced OCR text layer created with the Tesseract LSTM neural network machine learning engine. The second and third PDF variants republish all original PDF files of draft texts and meeting records in English only. Files are named after the associated resolution to assist discovery.

Citations from UNSC resolutions to other UNSC and UNGA resolutions are stored

in a specialized variant in GraphML format. This network data format is widely supported and can easily be imported into graphical software such as Gephi. The UNSC resolution nodes are associated with 54 node-level metadata variables, but exclude full-text variables.

A *bibliography* file in BiBTeX format is available for researchers to allow quick import of the entire set of UNSC resolutions with much useful metadata into personal bibliographic databases. This includes thematic information, summaries and voting data. BiBTeX is an open and widely used format that can be processed with Zotero, Endnote, Citavi, Jabref and other popular reference managers.

Pre-computed *analysis diagrams* and frequency tables are available in the analysis variant. These are intended as an overview for researchers approaching the dataset for the first time and to aid teachers.

3.4 Open Source Code

We have published the full source code underlying the creation of the CR-UNSC under a the GNU General Public License Version 3 (GPLv3) and permanently archived it with Zenodo. The source code also exactly specifies the computational environment as a Dockerfile with associated versioning information. Table 2 shows the availability of the source code.

The integrity and veracity of the source code and publication ZIP archives are documented with cryptographically secure hash signatures (SHA2-256 and SHA3-512). Hashes are stored in a separate CSV file created during the data set compilation process. The hashes are then signed with a personal GPG key to document provenance.

Table 2: Availability of the CR-UNSC Source Code.

Attribute	Detail
Full Name	Source Code for the ‘Corpus of Resolutions: UN Security Council’
Acronym	CR-UNSC
Initial Version	2024-05-03
Initial Release	https://doi.org/10.5281/zenodo.7319784
Newest Version	https://doi.org/10.5281/zenodo.7319783
Development	https://codeberg.org/seanfobbe/cr-unsc
License	GNU General Public License Version 3 (GPLv3)

3.5 Public Domain Dedication

The copyright status of UN documents, resolutions, drafts and meeting records is governed by UN Administrative Instruction ST/AI/189/Add.9/Rev.2 of 17 September 1987.4 UN Administrative Instruction ST/AI/189/Add.9/Rev.2/Add.2 extended the temporal scope of the former until “further revision to that instruction”. As no revision has been published to this day, the rules remain in force. The most recent

Index to Administrative Issuances ST/IC/2019/1 lists Administrative Instruction ST/AI/189/Add.9/Rev.2 as active.

UN Administrative Instruction ST/AI/189/Add.9/Rev.2 of 17 September 1987 clearly states that, inter alia, Official Records and United Nations documents are to be “left in the public domain”. We wish to honor the letter and spirit of this UN policy. To ensure the widest possible distribution of official UN documents and to promote the international rule of law we waive any copyright that might have accrued by creating the dataset under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.

4 Dataset Design and Computational Methods

4.1 Data Engineering and the Scientific Method

In the presentation of this novel UNSC corpus we give priority to description and the *method* of its creation as an independent contribution to the scientific literature. We agree with Gerring that it is preferable to describe and discuss a dataset as the actual scientific advance instead of adding minor theoretical investigations or superficial causal analyses in an effort to approach a conventionally pleasing publication (Gerring 2012, 721). The focus on newsworthy results has had a detrimental effect on the scientific literature (Ioannidis 2005).

The creation and publication of scientific datasets is not as dazzling an activity as the production of disruptive theoretical claims. However, in the long term the slow and steady accumulation of evidence and sustained development of method may be as important or even more important than theoretical contributions, even if these are backed by convincing empirical data. This is because the pursuit of a specific theory in data collection leads to close-mindedness and limits collection efforts only to data which may potentially support or refute the hypothesis (Gerring 2012, 734).

This principle is evident at the highest levels of scientific quality. Greenwald reviewed 77 Nobel Prizes for scientific contributions in medicine, chemistry and physics in the 21-period from 1991 to 2011 and discovered that 82% (63 prizes) were awarded for services to method, compared to only 18% for theory (14 prizes) (Greenwald 2012, 103). He concludes:

The available documentation of Nobel awards reveals two forms of method–theory synergy: (a) existing theories were often essential in enabling development of awarded methods, and (b) award-receiving methods often generated previously inconceivable data, which in turn inspired previously inconceivable theories. (Greenwald 2012, 99)

The scientific method is the unifying principle of science. The heart of science consists of methodically robust and reproducible research (Munafò et al. 2017). Scientific results derive their legitimacy from reproducible method, not the other way around. Gauch proposed the PEL model to guide the full disclosure of scientific arguments: presuppositions (P), evidence (E) and logic (L) (Gauch Jr. 2002, 128).

Presuppositions are reasonable and necessary — but unprovable — assumptions that are combined with empirical evidence according to a clearly defined logic to produce a scientific conclusion.

The PEL model is primarily intended for theoretical arguments, but also provides a good framework for analyzing the production of evidence. The collection of low-level evidence and transformation into a dataset of high-level evidence such as the CR-UNSC requires a significant number of assumptions and complicated logic to succeed. It is critical for the scientific record that dataset authors not only focus on the content of the dataset, but also on the data engineering and methods that went into creating it by documenting their assumptions, data and the logic that links them. With our focus on method we hope to help others with their dataset creation efforts and contribute to the art of scientific data engineering.

4.2 Reproducible End-to-End Workflow

The computational workflow is conceptualized as an end-to-end data pipeline in extract-transform-load (ETL) style, fully automated from first contact with the UNDL database to production of the final publication-quality ZIP archives. The access-controlled upload to Zenodo is excluded from the pipeline to permit anyone to computationally reproduce the workflow. In this section we wish to acknowledge the critical software components that underpin the project and which should be cited like any other publication (A. M. Smith, Katz, and Niemeyer 2016).

We provide the complete computational environment in which the pipeline is to be run, specified as a custom Dockerfile based on the *r-ver* line of Docker images published by the Rocker Project (Boettiger and Eddelbuettel 2017; Nüst et al. 2020). The Rocker image uses Ubuntu 22.04.2 LTS, a derivative of the Debian distribution of the Linux operating system.

Additional configuration files referenced by the Dockerfile name the full set of system and R packages to be installed. The R packages are version-locked to the R version named in the *r-ver* Docker image. We further provide a script to compile Tesseract (R. Smith 2007; R. Smith, Antonova, and Lee 2009; R. Smith et al. 2024) from source to allow for version-controlled access to more recent versions of the library.

The project is implemented with the `{targets}` pipeline framework for R (Landau 2021). We are heavily indebted to William Michael Landau for publishing `{targets}`, without which this project would have been impossible to realize in its full complexity. Additional pipeline constructs are sourced from the `{tarchetypes}` package (Landau 2024). `{targets}` is the successor to the `{drake}` framework (Landau 2018).

Parallel high-performance procedures are implemented with the `{future}` (Bengtsson 2021, 2024a) and `{future.apply}` (Bengtsson 2024b) packages for R.

The full workflow is split into over one hundred individually named components, each of which is created by a custom R expression, executed in its own R session with a reproducible seed and its result held in a separate R object stored on disk in the `qs` format (Ching 2024). We implement each component with referential transparency

according to functional programming principles. Components and their dependency relationships form a directed acyclic graph (DAG) shown in Figure 2.

The entire pipeline is published together with its source code on Zenodo to provide future researchers the opportunity to inspect every single R object at every step of the pipeline for errors or inconsistencies. There is only one limitation: pipeline components outputting large numbers of files (e.g. PDF files) store only the paths to the files on disk, not the content of the files themselves. However, if their contents are read back into R (e.g. TXT files) the pipeline contains the file content in the following step. Since most final PDF files are also published in the final ZIP archives, only some intermediate PDF files remain transient.

The full workflow can be completed in 40–50 hours on a commodity Ryzen 3700X (8 cores/16 threads) CPU using only 65 W thermal design power (TDP) and 64 GB RAM. No graphics processing units (GPU) or tensor processing units (TPU) are necessary, making this an energy-efficient project with limited environmental impact. To further cut down on development time and computational cost the pipeline is designed to be fault-tolerant and resumable, including the download components. If the pipeline is interrupted for any reason (e.g. internal error, test failures, internet disruption, power outage, manual shutdown) it can be continued from the last successful step.

In comparison with previous work on data pipelines built in R, for example by Fobbe (2022), the CR-UNSC pipeline is much more streamlined, offers greater maintainability and greater efficiency through the extensive use of functional programming, static branching and modularity. While the source code of the older CD-ICJ (Version 2023-10-22) and the modern CR-UNSC (Version 2024-05-19) both comprise ca. 3,100 lines of code, the CR-UNSC offers many more features, processes three times as many languages, runs a larger suite of tests and extracts citation data to produce a citation network of UNSC resolutions.

4.3 Testing and Reporting

The dataset compilation process includes an intense testing and quality control regime. We implement hard and soft quality tests, as well as automated reporting of results in a Compilation Report, Quality Assurance Report and the automated creation of the Codebook.

Hard tests measure clear and unambiguous expectations. They cause the pipeline to halt on error and force a choice between fixing the upstream database or adding an explicit correction to the source code. An example of a hard test is the requirement for all votes to sum to either 11 or 15 votes (depending on period). In total we test 86 individual expectations, grouped into 32 thematic tests. Hard tests are positioned either at the beginning or end of a function, to ensure that the input and output of a component meet the highest quality standards.

Soft tests examine the distributions of many variables by reporting summary statistics, frequency tables and diagrams of discrete and continuous distributions. They do not cause the pipeline to fail, but are reported in the Codebook (general interest) or

in the Quality Assurance Report (technical interest). The diagrams shown in this paper serve double duty as both analysis and continuous quality monitor of the data. The diagrams are recomputed whenever the underlying data changes.

5 Detailed Description of the Dataset

5.1 Text Corpus

The main feature of the CR-UNSC is its comprehensive text corpus of UNSC resolutions, drafts and meeting records. We source the PDF files for (almost) all UNSC resolutions (Arabic, Chinese, English, French, Spanish and Russian), drafts (English) and meeting records (English) from the UNDL database. One resolution each we acquired from UNDOCS and one from UNSMIL because they were not available from the UNDL. The workflow automatically tests the download manifest for completeness, duplicates and URL validity.

The text layers of all PDF documents are extracted with {pdftools} (Ooms 2024) and stored on disk as TXT files. Resolutions 1 through 899 and associated drafts and meeting records are additionally treated with optical character recognition (OCR) via the LSTM neural network engine of Tesseract 5 (R. Smith 2007; R. Smith, Antonova, and Lee 2009; R. Smith et al. 2024). For the English text of resolutions 1 through 899 one of us (Gasbarri) produced gold-standard texts based on revised OCR output.

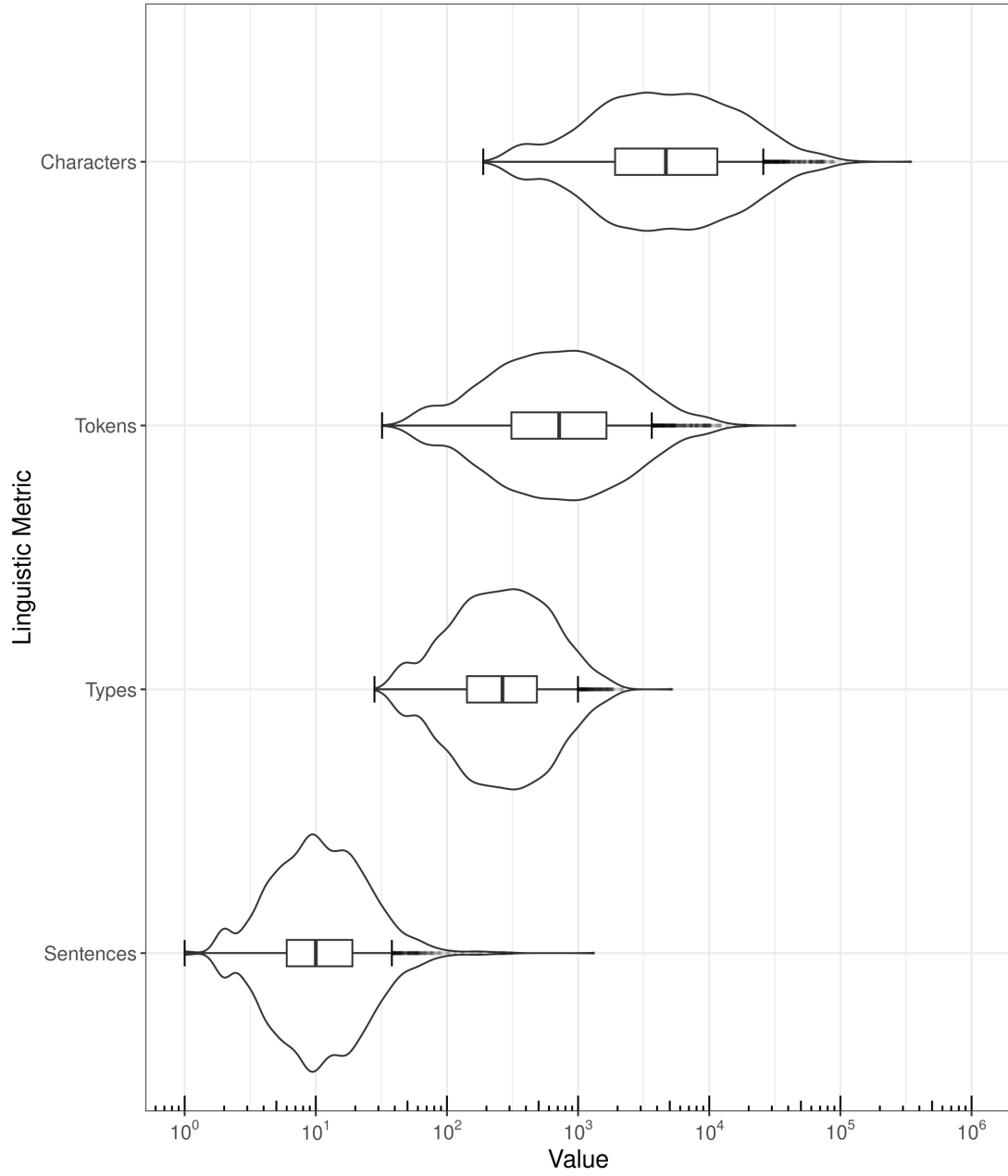
All results of the previous steps are merged to produce a best variant of all texts for the final dataset. The best variant of texts includes gold-standard (English) or Tesseract OCR versions (other languages, drafts and meeting records) of resolutions 1 through 899 and the digital text layer embedded in the published PDF files for resolutions 900 and later, which we assume to be born-digital. The merged text variables are then transformed into a data table.

We calculate measures of length (characters, tokens, types and sentences) for all documents in the English-language resolution variable and for the English resolution corpus as a whole with {quanteda} (Benoit et al. 2018, 2024). The summary statistics of these four linguistic metrics are recorded in Table 3. Figure 3 visualizes their distribution.

Table 3: Size of the corpus and individual documents.

Variable	Sum	Min	Quart1	Median	Mean	Quart3	Max
nchars	26071742	189	1914.75	4669.5	9578.16	11515.00	343887
ntokens	3704016	32	310.00	717.0	1360.77	1643.75	45030
ntypes	32316	28	142.00	265.0	367.16	485.00	5179
nsentences	43302	1	6.00	10.0	15.91	19.00	1313

To compare the text quality of extracted, OCR and gold (i.e. expert revised) variants in resolutions 1 through 899 we simulate a standard pre-processing workflow and calculate the number of features (unique tokens) with {quanteda} for each language



Fobbe/Gasbarri/Ridi | DOI: 10.5281/zenodo.11212056

Figure 3: Distributions of English language resolution length in the CR-UNSC. The box plots show the median (inner thick line), first and third quartile (outer box lines), 1.5 times the inter-quartile range (whiskers) and outliers (individual data points). The outer violin plots show the continuous distribution of values.

and type of process. OCR errors on individual characters cause a large number of unique tokens to appear, so the number of features is reasonable proxy for relative OCR quality. During pre-processing we remove numbers, punctuation, symbols and separators, lowercase all tokens and remove stopwords in English, French and Spanish. The resulting number of features determines the quality and speed with which document-feature matrix (or document-term matrix) oriented workflows, for example as the basis of topic models, can proceed. A lower feature count is better.

Table 4 shows the results for the text quality test. The quality of Tesseract OCR is clearly better than the quality of the original OCR that was included in the PDF documents, with a reduction of between 43% to 50% features, depending on language version. The gold-standard expert revision of the text is again clearly superior with a reduction of 74% of features.

Table 4: Results of the OCR quality test.

Language	Process	Features	Reduction (abs)	Reduction (rel)
English	Extracted	28761	0	0.00
English	OCR	16495	-12266	-0.43
English	Gold	7371	-21390	-0.74
French	Extracted	27466	0	0.00
French	OCR	13652	-13814	-0.50
Spanish	Extracted	28547	0	0.00
Spanish	OCR	15219	-13328	-0.47

We further test the texts for correctness of language label and purity of language content with {textcat} (Hornik et al. 2013, 2023). While each resolution text should be monolingual, the unprocessed nature of most PDFs (multiple resolutions per PDF, multiple languages per PDF) means that the non-English text variables are likely contaminated with superfluous text. The uses they are put to should be carefully considered. This test does not work well for non-Western languages and scripts, so the tests have been omitted for Arabic, Chinese and Russian. All English gold-standard resolutions pass this test, 15 French and 7 Spanish OCR resolutions fail.

Finally, we test all resolutions for suspiciously small character counts and document their full-text (except Chinese) in the quality control report. Documents with low character count indicate potential errors in processing. This could mean an absence of documents in the UNDL, an empty PDF file, an extraction/OCR failure or a refining failure. All English documents pass the test. One Arabic, two French, two Spanish, two Russian and 366 Chinese documents fail the test.

5.2 Thematic Metadata

Thematic metadata can reduce entry barriers for researchers unfamiliar with quantitative methods and ease the learning curve for new dataset users. Some thematic

metadata is provided by the UNDL and cleaned by ourselves, some of it is based on regular expressions applied to the English full text of UNSC resolutions.

Given the UNSC’s responsibility for international security and the importance of different legal bases in the UN Charter we chose to highlight a number of related keywords and key phrases. All of these are extracted from the full text. We document the mentions of “Chapter VI”, “Chapter VII” and “Chapter VIII”, which indicate the powers invoked by the UNSC in the operative part of the resolution. In the context of Chapter VII resolutions it matters whether phrases such as “threat to peace”, “breach of peace” and “aggression” occur, where threats to the peace are the most common invocation and the other two are exceedingly rare. We further document mentions of “self-defense” (with alternate spellings) to round out this preliminary analysis of legal bases in the UN Charter.

In addition to the extracted metadata we package a wide variety of thematic information curated by the UNDL. Of particular note are the title and alternative title of the resolution, three separate topical classifications (topic, subject, series), additional notes and a narrative summary of the resolution. Figure 4 shows a count of UNSC resolutions by subject matter. Another variable indicates if the resolution contains a legal instrument, such as the statute of an international court. Some agenda information is included as well.

5.3 Geographic Metadata

The dataset includes a limited amount of geographic metadata to showcase potential applications of the corpus, to assist with teaching and to facilitate preliminary analyses. Geographic variables indicate if a country or geographical region has been mentioned in the full text of UNSC resolutions. We provide these as ISO-3166-1 Alpha-3 country codes and names, M49 codes, M49 regions, M49 intermediate regions and M49 subregions.

We extract these indicators by parsing the English resolution texts for the short names contained in the current UN M49 standard as provided by {ISOcodes} (Buchta and Hornik 2024), with slight custom modifications to increase accuracy and align them more closely with the short names used in UN documents. Examples of such modifications are the addition of “DPRK” for North Korea (a common acronym in UN documents), breaking up the full UN name of the UK (which, inconveniently, is also its UN short name) into components (“Northern Ireland”, “Great Britain”, and “United Kingdom”), reducing the short name to the closest unique string (e.g. Hong Kong) or adding updated names (Türkiye for Turkey). All custom additions are documented in the source code. The results are then mapped to and returned as ISO-3166 Alpha-3 codes. We operate with ISO-3 codes instead of M49 numeric codes to make the source code more readable and to reduce coding errors.

We map the obtained ISO-3166 Alpha-3 codes to ISO-3166 names and M49 codes with {ISOcodes} (Buchta and Hornik 2024). With {countrycode} (Arel-Bundock 2024) we map ISO codes to M49 regions (e.g. Africa), M49 intermediate regions (e.g. Sub-Saharan Africa) and M49 subregions (e.g. Eastern Africa). Note that

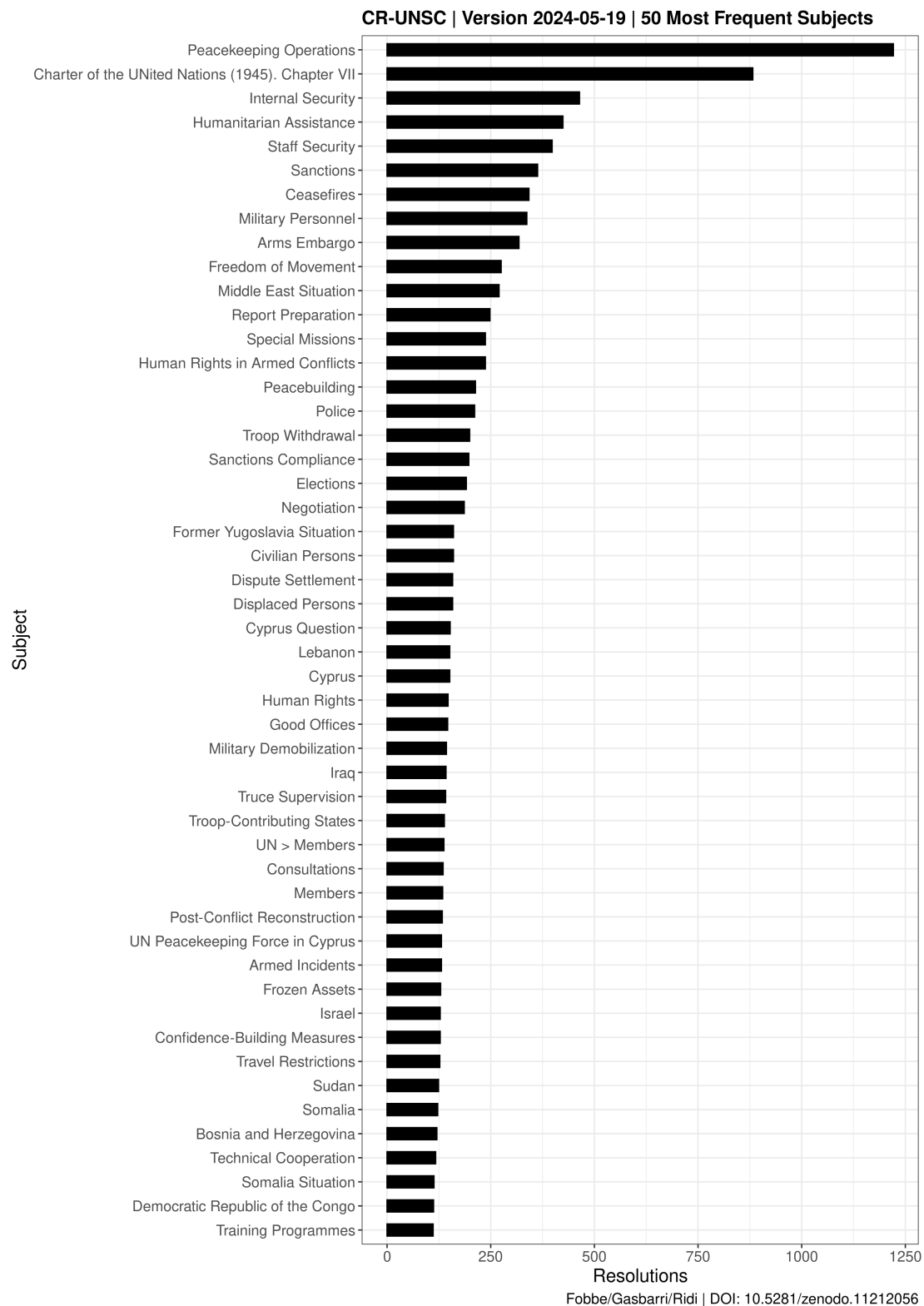


Figure 4: Number of UNSC resolutions by subject matter as coded by the UN Digital Library.

while regions and subregions are available for all countries, not all countries have intermediate regions. Figure 5 shows the count of UNSC resolutions in the CR-UNSC by UN M49 sub-regions.

This dictionary approach has some notable limitations. The first limitation is that some countries whose short names are strict subsets of other countries (Congo and Democratic Republic of the Congo; Sudan and South Sudan) cannot be distinguished. This means that mentions of the country with the superset name will always be an overestimate. This problem affects very few countries, although unfortunately those that have been of critical interest to the UNSC in the past.

The second limitation is that we chose to adopt the current revision of the M49 standard. Historical countries (e.g. Yugoslavia, Soviet Union) are not included. Historical names of countries (e.g. Birma/Burma for Myanmar) are also not included at this point, although we intend to add them in the future.

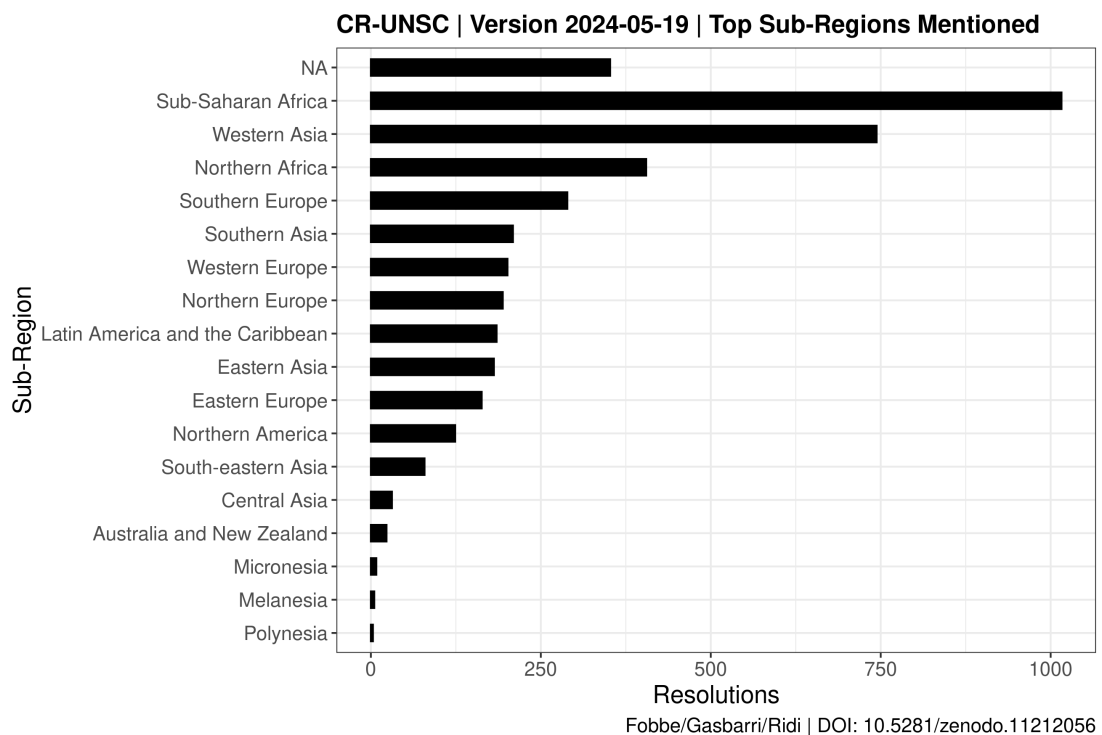


Figure 5: Number of UNSC resolutions by sub-region mentioned.

5.4 Voting Metadata

The study of voting behavior in the United Nations has drawn intense academic attention over its entire lifetime (Bailey, Strezhnev, and Voeten 2017), although the UNGA has enjoyed pride of place due to its universal representation of UN membership. Nevertheless, voting at the UNSC has seen its own brand of fierce attention (Todd 1969; Dreher, Sturm, and Vreeland 2009), with much of it focused on the veto powers of the permanent five (P-5) members (Winter 1996).

Our dataset takes a different approach to UNSC voting and focuses not on resolutions that failed (e.g. because of a veto), but resolutions that were successfully adopted. We accompany all UN resolutions with voting data that document the voting results and voting patterns of UNSC resolutions. We provide simple counts of yes, no, abstaining and non-voting members, the total number of members at the time, the date of the vote and a breakdown of votes by State.

The voting data is automatically extracted from the metadata contained in the UNDL records. We programmatically test all voting variables for plausibility, such as the presence of an integer with a minimum of 0, a maximum of 15 and all votes correctly summing to either 11 or 15. We implement hand-coded corrections for about a dozen resolutions that fail these tests when extracted from the UNDL database.

Note that the UNSC was enlarged from 11 to 15 members (growing from 6 to 10 non-permanent members) in 1965, so comparing absolute vote counts is only appropriate before and after this watershed moment. All other analyses should use proportions. Figure 6 displays an analysis of exact voting proportions, disaggregated by yes, no, abstention and nonvote. Note the large number of high yes percentages, which could be an indication of the intense diplomatic negotiations that precede UNSC votes. The few resolutions with zero percent yes votes were adopted without a vote.

A notable limitation of this voting data is that it is only available for successfully adopted resolutions, not for failed drafts. This impacts prediction models in particular, since no examples of the negative class are included. It may nevertheless be useful in studying revealed state preferences among the membership of the UNSC, especially when linked to the text of the resolution. Voting data can also indicate which adopted resolutions were especially contentious, making an analysis of the draft text versus the final text attractive.

5.5 Citation Network Data

The study of legal citation data has long been a mainstay of quantitative approaches to the law. Bleicher (1969) was one of the earliest to study UN data. Recent years have seen more and more interest in citation analysis of international law (Alschner and Charlotin 2018), particularly the citation network of the UNSC (Ridi and Gasbarri 2023; Mesquita and Pires 2024a). While many analysis results have been published, there remains a lacuna in terms of data availability.

We provide a specialized variant of the CR-UNSC dataset that contains all citations from UNSC resolutions to other UNSC resolutions, as well as UNSC resolutions to UNGA resolutions. Formally it is represented as a weighted, directed graph of citations (edges) between resolutions (nodes). The citation data is available in GraphML, a format that is widely importable into network analysis software and, in particular, can be used with popular graphical software such as Gephi.

Citations to UNSC resolutions contained in UNSC resolution text follow a predictable and easily extracted pattern of the form “resolution NUMBER (YEAR)”, e.g. “resolution 1 (1946)”. We extract all such citations with `{stringi}` (Gagolewski 2022; Gagolewski et al. 2023), map them to their source resolutions, clean them and create

CR-UNSC | Version 2024-05-19 | Voting Data

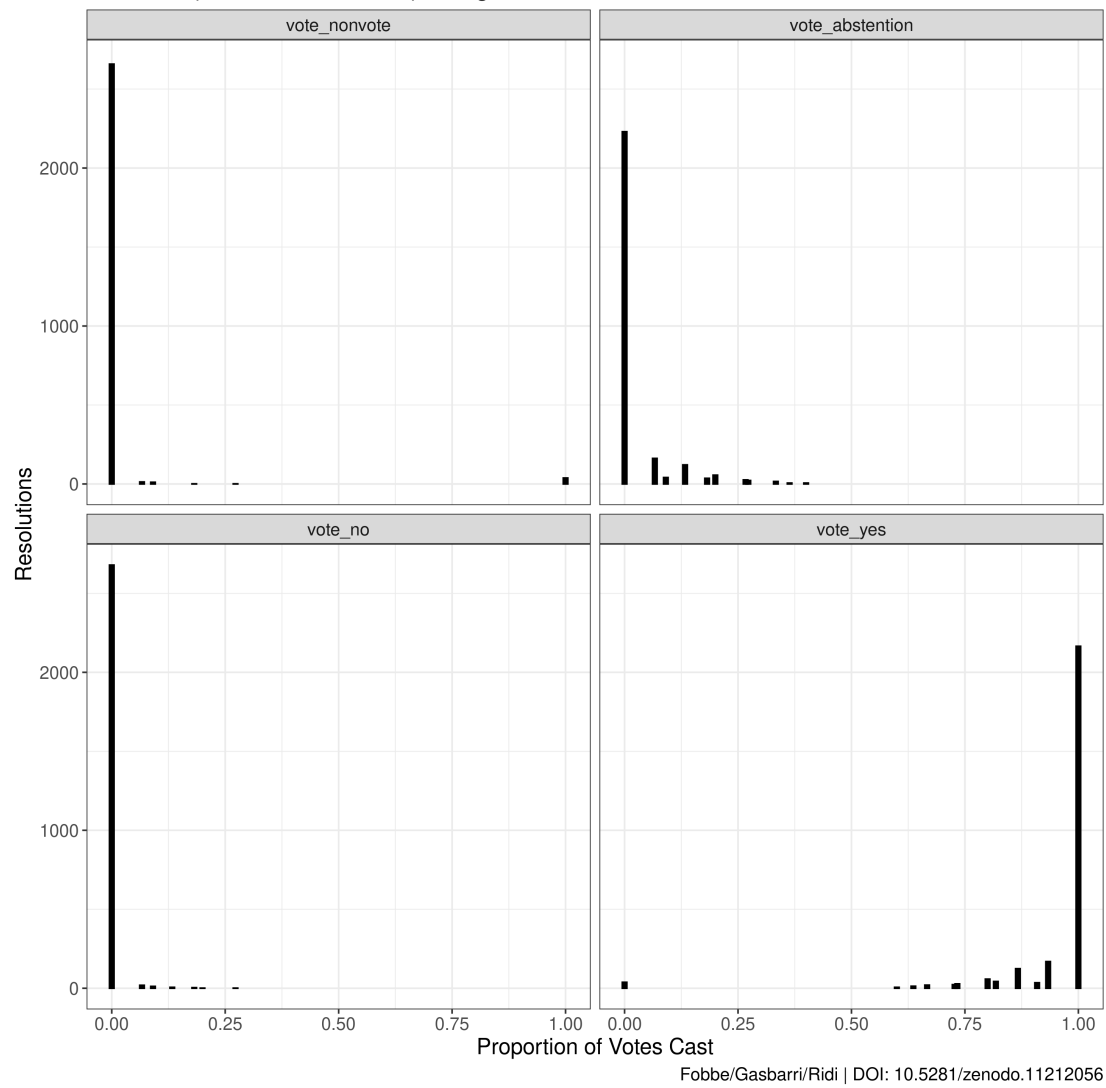


Figure 6: Counts of yes and no votes, abstentions and non-voting members by proportion in the CR-UNSC.

an edge list. From this edge list we create a graph representation with {igraph} (Csardi and Nepusz 2006; Csárdi et al. 2024) and add the metadata sourced from the main dataset. We visualize the network with ggraph (Pedersen 2024) in Figure 7.

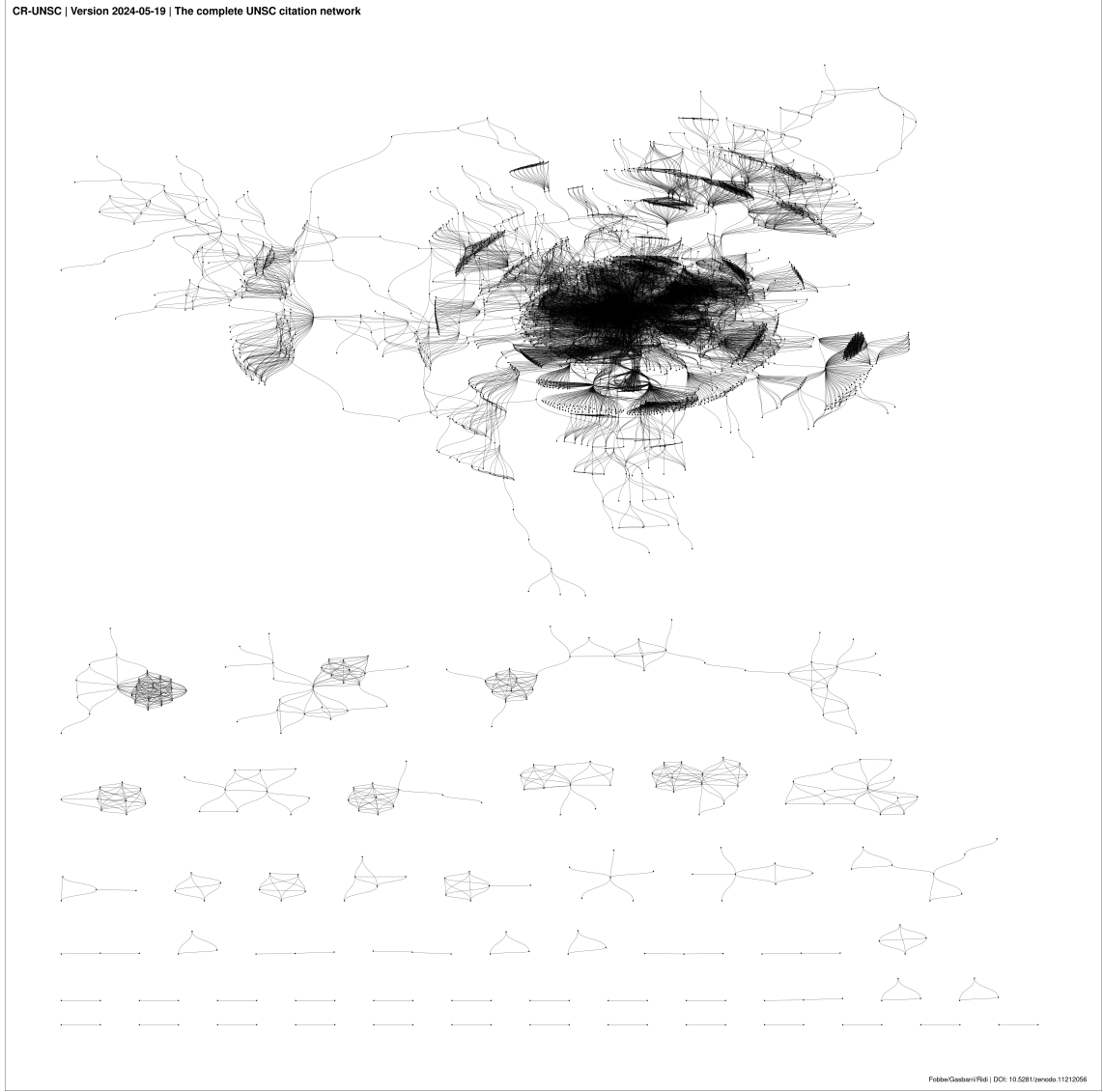


Figure 7: The citation network of the UNSC between UNSC-UNSC resolutions and UNSC-UNGA resolutions.

Cleaning steps for UNSC resolution nodes include standardizing resolution names, removing whitespace, removing resolutions that do not cite other resolutions, removing self-citations, and removing implausible citations (resolutions cannot cite later resolutions).

UNGA citations in UNSC resolutions are rare, but they do exist. Their format is more complicated than UNSC resolutions. We construct regular expressions for UNGA resolutions in regular sessions with Roman numerals in short form (“A/RES/1514(XV)”) and long form (“General Assembly resolution 1514(XV)”), Arabic numerals in short form (“A/RES/75/325”) and long form (“General Assembly

resolution 75/325”), as well as emergency special sessions.

Cleaning steps for UNGA resolution nodes include standardizing resolution names, removing whitespace and removing resolutions that do not cite other resolutions.

Table 5 contains basic metrics for the citation network published as part of the CR-UNSC. The number of nodes gives a count of resolutions with in- or outgoing citations. The number of edges gives a count of resolution-pairs that are connected by at least one citation. The individual weights for directed edges give the number of citations for the bilateral links. The out-strength gives the total count of citations between all resolutions (excluding self-citations).

Table 5: Basic Metrics for CR-UNSC citation network between UNSC-UNSC and UNSC-UNGA citations.

Metric	Value	Metric	Value
Number of Nodes	2,516.00	Min In-Degree	0.00
Number of Edges	15,507.00	Mean Out-Degree	6.16
Strength (Out)	26,866.00	Max Out-Degree	49.00
Mean Degree	12.33	Min Out-Degree	0.00
Max Degree	336.00	Diameter	40.00
Min Degree	1.00	Radius	1.00
Mean In-Degree	6.16	Assortativity (Degree)	0.10
Max In-Degree	331.00	Mean Distance	6.68

The UNSC resolution nodes are associated with as much metadata as makes sense from a network perspective. We include a total of 54 metadata variables from the main dataset. These include many ID variables, most thematic variables, most voting variables, most geographic variables and the meta variables related to the dataset. An important limitation of this data is that metadata variables are only available for UNSC resolution nodes, not UNGA resolution nodes.

The graph most importantly does not include the full texts of resolutions. To select nodes based on the contents of the full text it is necessary to run the analyses on the CSV file, export the resolution numbers and use these to select the appropriate nodes in the chosen network analysis software.

5.6 Source Variables

Source variables store URLs for all full texts of resolutions, drafts and meeting records in all six official UN languages, as well as the record pages for resolution, draft and the meeting record. The URLs are part of the dataset for several reasons. First, the UNDL explicitly requested we include the source identifiers of all resources. Second, a desire for transparency. Third, a possibility for end-users to extend the dataset by accessing PDFs of drafts and meeting records in languages other than English for research projects beyond that which the CR-UNSC can offer. Fourth, economy of process. We limited ourselves to including full texts and PDFs of drafts

and meeting records in English in order not to consume excessive bandwidth from the UNDL (saving $5 \times 2 \times 2722$ requests). The URLs offer a more economical way of accessing the data, if necessary.

6 Conclusion

We introduced the *Corpus of Resolutions: UN Security Council* (CR-UNSC), a novel international legal-political dataset containing texts and metadata for all resolutions of the United Nations Security Council (UNSC) from resolution 1 (1946) through resolution 2722 (2024) as published by the UN Digital Library (UNDL). The dataset provides full resolution texts in all six official UN languages, draft texts and full meeting records in English, as well as dozens of metadata variables, for a grand total of 82 variables in structured tabular CSV format with extensive narrative documentation. In addition to the main tabular dataset the CR-UNSC is accompanied by specialized variants for network analysis and bibliography management.

With the CR-UNSC we hope to contribute to a more empirical view of the international legal system. In an international community founded on the rule of law the activities of the United Nations must be public, transparent and defensible. In the 21st century this requires quantitative scientific review of decisions and actions. In these troubled times when the rule-based international order is under attack it becomes more important than ever to document, understand and promote systematic approaches to international law.

7 References

- Alschner, Wolfgang. 2021. “The Computational Analysis of International Law.” In *Research Methods in International Law: A Handbook*. Edward Elgar. <https://doi.org/10.4337/9781788972369.00022>.
- Alschner, Wolfgang, and Damien Charlotin. 2018. “The Growing Complexity of the International Court of Justice’s Self-Citation Network.” *European Journal of International Law* 29 (1): 83–112.
- Arel-Bundock, Vincent. 2024. *countrycode: Convert Country Names and Country Codes*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.countrycode>.
- Badache, Fanny, Sara Hellmüller, and Bilal Salayme. 2022. “United Nations Security Council Peace-Related Speeches (UNSCPeaS).” Harvard Dataverse. <https://doi.org/10.7910/DVN/00ROVZ>.
- Bailey, Michael A, Anton Strezhnev, and Erik Voeten. 2017. “Estimating Dynamic State Preferences from United Nations Voting Data.” *Journal of Conflict Resolution* 61 (2): 430–56.
- Baturo, Alexander, Niheer Dasandi, and Slava J Mikhaylov. 2017. “Understanding State Preferences with Text as Data: Introducing the UN General Debate Corpus.” *Research & Politics* 4 (2): 1. <https://doi.org/10.1177/2053168017712821>.
- Beardsley, Kyle. 2013. “The UN at the Peacemaking–Peacebuilding Nexus.” *Conflict Management and Peace Science* 30 (4): 369–86. <https://doi.org/10.1177/073889>

- 4213491354.
- Bell, Christine, and Sanja Badanjak. 2019. “Introducing PA-x: A New Peace Agreement Database and Dataset.” *Journal of Peace Research* 56 (3): 452–66. <https://doi.org/10.1177/0022343318819123>.
- Bengtsson, Henrik. 2021. “A Unifying Framework for Parallel and Distributed Processing in r Using Futures.” *The R Journal* 13 (2): 208–27. <https://doi.org/10.32614/RJ-2021-048>.
- . 2024a. *future: Unified Parallel and Distributed Processing in r for Everyone*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.future>.
- . 2024b. *future.apply: Apply Function to Elements in Parallel Using Futures*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.future.apply>.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. “quanteda: An r Package for the Quantitative Analysis of Textual Data.” *Journal of Open Source Software* 3 (30): 774. <https://doi.org/10.21105/joss.00774>.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, and William Lowe. 2024. *quanteda: Quantitative Analysis of Textual Data*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.quanteda>.
- Benson, Michelle, and Colin Tucker. 2021. “United Nations Peacekeeping Resolutions (UNPR).” Harvard Dataverse. <https://doi.org/10.7910/DVN/G3IGFH>.
- . 2022a. “The Importance of UN Security Council Resolutions in Peacekeeping Operations.” *Journal of Conflict Resolution* 66 (3): 473–503. <https://doi.org/10.1177/00220027211044205>.
- . 2022b. “United Nations Conflict Resolutions (UNCR).” Harvard Dataverse. <https://doi.org/10.7910/DVN/EUYIW0>.
- Bleicher, Samuel A. 1969. “The Legal Significance of Re-Citation of General Assembly Resolutions.” *American Journal of International Law* 63 (3): 444–78. <https://doi.org/10.2307/2198866>.
- Boettiger, Carl, and Dirk Eddelbuettel. 2017. “An Introduction to Rocker: Docker Containers for R.” *The R Journal* 9 (2): 527–36. <https://doi.org/10.32614/RJ-2017-065>.
- Buchta, Christian, and Kurt Hornik. 2024. *ISOcodes: Selected ISO Codes*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.ISOcodes>.
- Ching, Travers. 2024. *Qs: Quick Serialization of r Objects*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.qs>.
- Clayton, Govinda, Han Dorussen, and Tobias Böhmelt. 2021. “United Nations Peace Initiatives 1946-2015: Introducing a New Dataset.” *International Interactions* 47 (1): 161–80. <https://doi.org/10.1080/03050629.2020.1772254>.
- Csardi, Gabor, and Tamas Nepusz. 2006. “The igraph Software Package for Complex Network Research.” *InterJournal Complex Systems*: 1695. <https://igraph.org>.
- Csárdi, Gábor, Tamás Nepusz, Vincent Traag, Szabolcs Horvát, Fabio Zanini, Daniel Noom, and Kirill Müller. 2024. *igraph: Network Analysis and Visualization in r*.

- <https://doi.org/10.32614/cran.package.igraph>.
- Di Salvatore, Jessica, Magnus Lundgren, Kseniya Oksamytna, and Hannah M Smidt. 2022. “Introducing the Peacekeeping Mandates (PEMA) Dataset.” *Journal of Conflict Resolution* 66 (4-5): 924–51. <https://doi.org/10.1177/00220027211068897>.
- Dreher, Axel, Jan-Egbert Sturm, and James Raymond Vreeland. 2009. “Global Horse Trading: IMF Loans for Votes in the United Nations Security Council.” *European Economic Review* 53 (7): 742–57.
- Fjelstul, Joshua, Simon Hug, and Christopher Kilby. 2025. “Decision-Making in the United Nations General Assembly: A Comprehensive Database of Resolution-Related Decisions.” *The Review of International Organizations*, 1–18. <https://doi.org/10.1007/s11558-024-09580-1>.
- Fobbe, Sean. 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.
- Gagolewski, Marek. 2022. “stringi: Fast and Portable Character String Processing in R.” *Journal of Statistical Software* 103 (2): 1–59. <https://doi.org/10.18637/jss.v103.i02>.
- Gagolewski, Marek, Bartek Tartanus, Unicode Inc., et al. 2023. *stringi: Fast and Portable Character String Processing Facilities*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.stringi>.
- Gauch Jr., Hugh G. 2002. *Scientific Method in Practice*. Cambridge University Press.
- Gerring, John. 2012. “Mere Description.” *British Journal of Political Science* 42 (4): 721–46. <https://doi.org/10.1017/s0007123412000130>.
- Greenwald, Anthony G. 2012. “There Is Nothing so Theoretical as a Good Method.” *Perspectives on Psychological Science* 7 (2): 99–108. <https://doi.org/10.1177/1745691611434210>.
- 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*. <https://doi.org/10.32614/cran.package.textcat>.
- Ioannidis, John PA. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2 (8): e124. <https://doi.org/10.1371/journal.pmed.0020124>.
- Jankin, Slava, Alexander Baturo, and Niheer Dasandi. 2024. “Words to Unite Nations: The Complete United Nations General Debate Corpus, 1946–Present.” *Journal of Peace Research*. <https://doi.org/10.1177/00223433241275335>.
- Lamprecht, Anna-Lena, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, et al. 2020. “Towards FAIR Principles for Research Software.” *Data Science* 3 (1): 37–59. <https://doi.org/10.3233/ds-190026>.
- Landau, William Michael. 2018. “The drake r Package: A Pipeline Toolkit for Reproducibility and High-Performance Computing.” *Journal of Open Source*

- Software* 3 (21): 550. <https://doi.org/10.21105/joss.00550>.
- . 2021. “The targets r Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing.” *Journal of Open Source Software* 6 (57): 2959. <https://doi.org/10.21105/joss.02959>.
- . 2024. *tarchetypes: Archetypes for Targets*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.tarchetypes>.
- Mesquita, Rafael. 2024. “What Do i Need to Say to Get Your Signature? Adding Draft Resolution Text to the UN General Assembly Sponsorship Dataset.” *Research & Politics* 11 (4): 1. <https://doi.org/10.1177/20531680241310396>.
- Mesquita, Rafael, and Antonio Pires. 2024a. “Jurisprudence in Hard and Soft Law Output of International Organizations: A Network Analysis of the Use of Precedent in Un Security Council and General Assembly Resolutions.” *Artificial Intelligence and Law*, 1–30.
- . 2024b. “The References of the Nations: Introducing a Corpus of United Nations General Assembly Resolutions Since 1946 and Their Citation Network.” *Journal of Peace Research*.
- Munafò, Marcus R, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. 2017. “A Manifesto for Reproducible Science.” *Nature Human Behaviour* 1 (1): 1–9. <https://doi.org/10.1038/s41562-016-0021>.
- Nüst, Daniel, Dirk Eddelbuettel, Dom Bennett, Robrecht Cannoodt, Dav Clark, Gergely Daróczi, Mark Edmondson, et al. 2020. “The Rockerverse: Packages and Applications for Containerisation with R.” *The R Journal* 12 (1): 437–61. <https://doi.org/10.32614/RJ-2020-007>.
- Ooms, Jeroen. 2024. *pdftools: Text Extraction, Rendering and Converting of PDF Documents*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.pdfutils>.
- Pedersen, Thomas Lin. 2024. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. CRAN: Contributed Packages. The R Foundation. <https://doi.org/10.32614/cran.package.ggraph>.
- Pires, Antonio. 2024. “UNSC Resolutions and Agenda Items.” Harvard Dataverse. <https://doi.org/10.7910/DVN/5EPLTO>.
- Pomeroy, Caleb. 2017. “spaceTexts: A New Corpus of Speeches in the UN Committee on the Peaceful Uses of Outer Space.” In *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, 41–46. IEEE. <https://doi.org/10.1109/fads.2017.8253191>.
- Ridi, Niccolo, and Lorenzo Gasbarri. 2023. “The Role of Previous Resolutions in the Practice of the Security Council.” *Columbia Journal of Transnational Law* 61 (3): 571–640.
- Schoenfeld, Mirco, Steffen Eckhard, Ronny Patz, Hilde van Meegdenburg, and Antonio Pires. 2019. “The UN Security Council Debates.” Harvard Dataverse. <https://doi.org/10.7910/DVN/KGVSYH>.
- Seabra, Pedro, and Rafael Mesquita. 2022. “Beyond Roll-Call Voting: Sponsorship Dynamics at the UN General Assembly.” *International Studies Quarterly* 66 (2): sqac008. <https://doi.org/10.1093/isq/sqac008>.

- Smith, Arfon M, Daniel S Katz, and Kyle E Niemeyer. 2016. “Software Citation Principles.” *PeerJ Computer Science* 2: e86. <https://doi.org/10.7717/peerj-cs.86>.
- Smith, Ray. 2007. “An Overview of the Tesseract OCR Engine.” In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, 629–33. Washington, DC, USA: IEEE Computer Society. <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf>.
- Smith, Ray, Ahmad Abdulkader, Rika Antonova, Nicholas Beato, Jeff Breidenbach, Samuel Charron, Phil Cheadle, et al. 2024. *Tesseract OCR*. <https://github.com/tesseract-ocr/tesseract>.
- Smith, Ray, Daria Antonova, and Dar-Shyang Lee. 2009. “Adapting the Tesseract Open Source OCR Engine for Multilingual OCR.” In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR*, edited by Venu Govindaraju, Premkumar Natarajan, Santanu Chaudhury, and Daniel P. Lopresti, 1–8. ACM International Conference Proceeding Series. Barcelona, Spain: ACM. <https://doi.org/http://doi.acm.org/10.1145/1577802.1577804>.
- Todd, James E. 1969. “An Analysis of Security Council Voting Behavior.” *Western Political Quarterly* 22 (1): 61–78.
- Tomuschat, Christian. 2017. “The (Hegemonic?) Role of the English Language.” *Nordic Journal of International Law* 86 (2): 196–227.
- Voeten, Erik, Anton Strezhnev, and Michael Bailey. 2009. “United Nations General Assembly Voting Data (V34).” Harvard Dataverse. 2009. <https://doi.org/10.7910/DVN/LEJUQZ>.
- Warntjen, Andreas. 2016. “United Nations General Assembly Resolutions, 1946-2014 (V1).” Harvard Dataverse. 2016. <https://doi.org/10.7910/DVN/T8EIWO>.
- Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 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>.
- Winter, Eyal. 1996. “Voting and Vetoing.” *American Political Science Review* 90 (4): 813–23.
- Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. “The United Nations Parallel Corpus V1.0.” In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3530–34.