Annex to "What do I need to say to get your signature? Adding draft resolution text to the UN General Assembly Sponsorship Dataset"

Downloading the PDFs of the draft resolutions

The original dataset by was based on metadata retrieved from custom queries at the UN Digital Library (see the original Seabra and Mesquita article for a full description of the process). The metadata included information sufficient to recover the PDF files containing the text of each draft. The PDFs are stored under predictable URLs (e.g., L-Doc "A/72/L.10" is at https://digitallibrary.un.org/record/**1323260**/files/**A_72_L.10**-EN.pdf?ln=en). The original dataset contained the numeric record (in this case, 1323260) and the UN Document Symbol (A/72/L.10), so that it was possible to reconstruct candidate URLs for this record, with some trial and error for the remaining parts of the string. For instance, the example URL requires replacing the "/" from A/72/L.10 to "_", but for some entries the "." also needed to be replaced by "-". The replacements for these punctuation characters were the main source of variation in URLs

From the 7,111 records of interest, 27 could not be matched to an online file using URL reconstruction, but were located later through manual search. We also found that 76 PDFs had some form of file error that required manual correction. The most common was that a PDF filename did not correspond to the L-Document inside it (e.g. A/58/L.62 file named A/58/L.62 but content is A/C.1/57/L.13). Only two cases differed: A/55/L.75 had illegible encoding and had to be re-typed by hand, and A/C.3/65/L.14 was only available in Russian. No version in English was found so its entry is blank at the dataset.

Extracting and processing text

To import not only the text of the PDFs but also rich data on formatting, (i.e., boldface, italics, hyperlinks), we converted the PDFs to HTML files. The reason for converting into HTML is because this markup language preserves information on text formatting in delimiters that can be easily parsed with R packages. We converted the PDFs into HTMLs using the Linux program *pdftohtml,* imported the HTML files into R and, using *rvest* and *xml2,* extracted the integral text of the L-Document, information on the string logcation of bold and italic characters, and removed line breakers ("\n") so that the final text contained overall less debris. This cleaned text was stored at the column *text_ed* on the dataset, while the original text was preserved at the column *text_raw*.

Considering that this cleaning process sought to separate substantive data from metadata, we left the *text_ed* field as *NA* for all L-Documents that were addenda. This is because addenda do not alter the content of a draft resolution, but only add to its list of sponsors. This increment is captured by the metadata and hence there was no need to keep anything on the *text_ed*.

The verbs opening the paragraphs of the preambular and operative parts of an L-Document are of special significance. These are always presented in italics, so that we can use the same procedure as above to first locate the position of these words and then, instead of erasing them, copying them into a separate field. Other elements of an L-Document text can also appear in italics. Three have been the most common: the opening phrase "The General Assembly," at the top of resolutions; the text of footnotes; and some rare elements on occasion, such as page numbers or parts of the resolution identification. To remove those and only retain operative verbs, we erased from the

results: the string "The General Assembly", any sentence longer than 5 words (the longest operative verbs we found were sentences like "expressing its deep concern"), and any number character. The retrieved strings were stored in the column *par_verbs* and the verbs were separated from one another with double dashes ("--").

After these procedures, the dataset contained four columns for each L-Document: an ID column with the UN Document Symbol; one with the raw text, one with the text stripped of resolution title, list of sponsoring countries, and line breakers; and one with the verbs and verb phrases that opened preambular and operative paragraphs.