kremlin.ru_en_2024

Corpus based on Russia's president website (in English, 1999-2023)

Giorgio Comai (OBCT/CCI)

2024-09-16

Scope of this corpus

This textual dataset is based on en.kremlin.ru, i.e. the English-language version of the official website of the president of the Russian Federation. It includes only its main sections with news and updates; it does not include other sections of the website such as legal documents, the Constitution, etc.

This dataset includes contents published between 31 December 1999 and 31 December 2023, under two Russian presidents: Vladimir Putin and Dmitri Medvedev.

Summary statistics

Dataset name: kremlin.ru_en_2024

Dataset description: all news items published on the English-language version of Kremlin.ru

Start date: 1999-12-31

End date: 2023-12-31

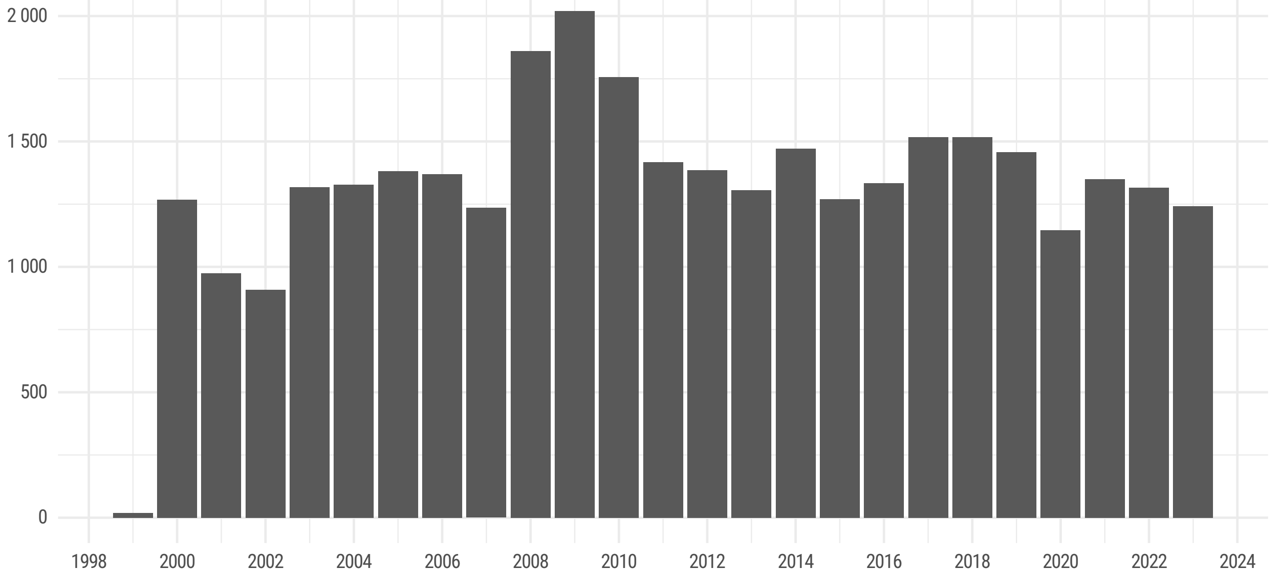
Total items: 33 165

Available columns: doc_id; text; title; date; time; datetime; location; description; keywords; tags; tags_links; persons_id; persons_name; url_id; url

License: Creative Commons Attribution 4.0 International

Number of items per year published on the English-language version of Kremlin.ru

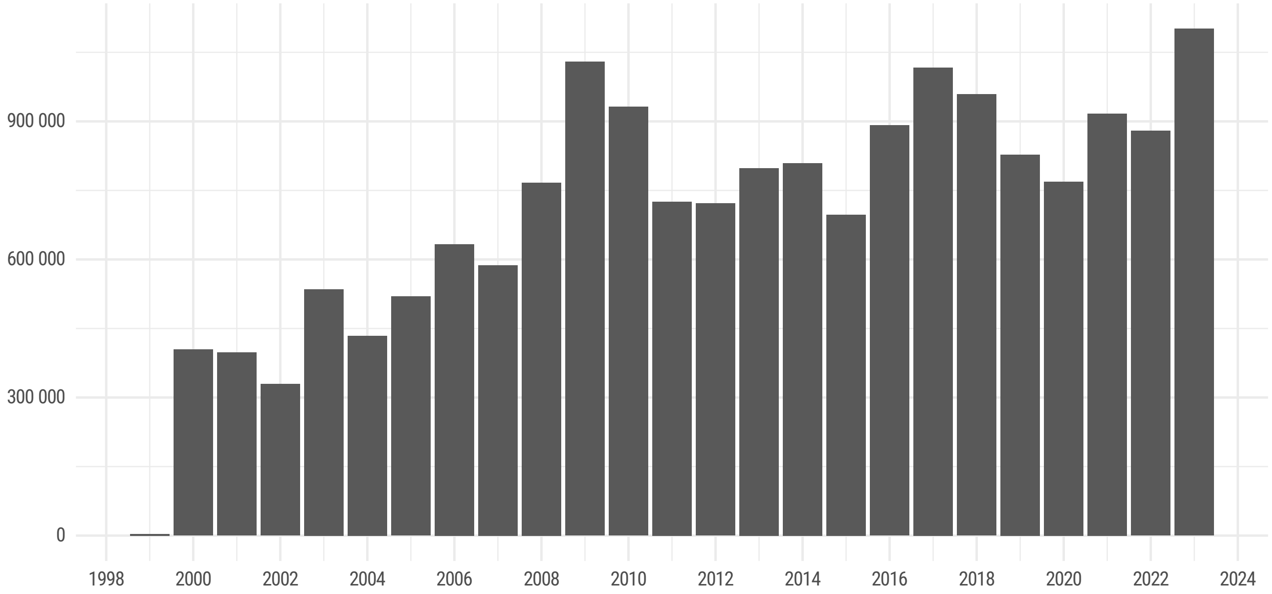
Based on 33 165 items published between 31 December 1999 and 31 December 2023



Source: Giorgio Comai / tadadit.xyz / kremlin.ru_en_2024

Number of words per year published on the English-language version of Kremlin.ru

Based on 33 165 items published between 31 December 1999 and 31 December 2023



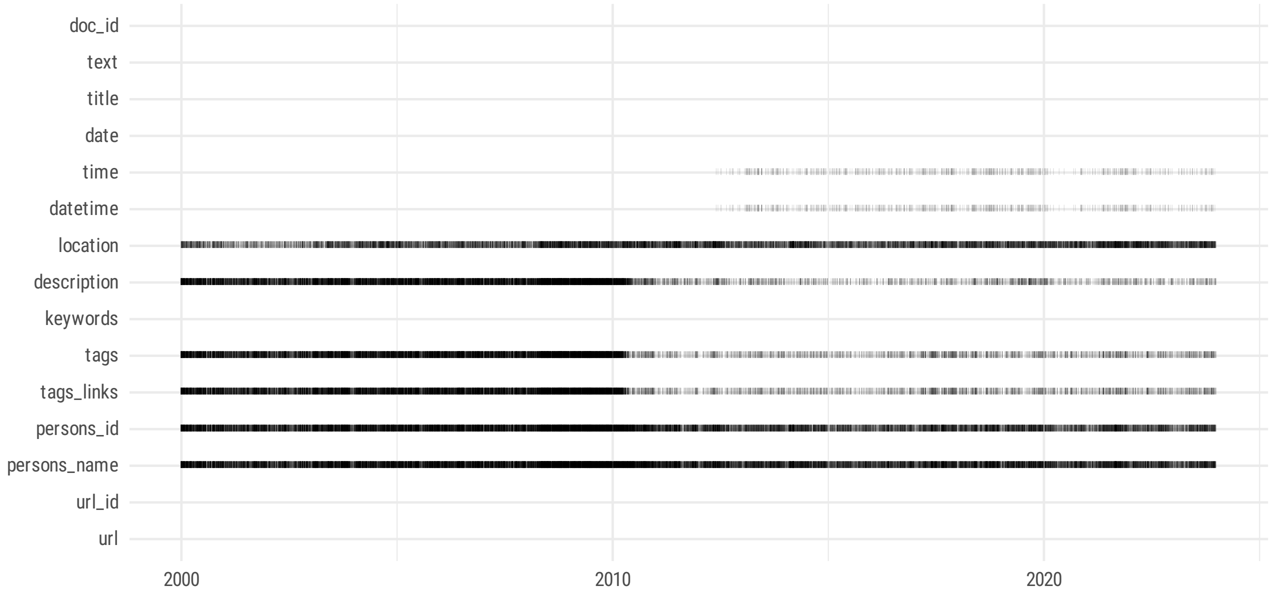
Source: Giorgio Comai / tadadit.xyz / kremlin.ru_en_2024

Missing data

field	present	missing	missing_share
doc_id	33 165	0	0.0%
text	33 165	0	0.0%
title	33 165	0	0.0%
date	33 165	0	0.0%
time	32 552	613	1.8%
datetime	32 552	613	1.8%
location	17 635	15 530	46.8%
description	16 734	16 431	49.5%
keywords	33 165	0	0.0%
tags	16 708	16 457	49.6%
tags_links	16 708	16 457	49.6%
persons_id	9 253	23 912	72.1%
persons_name	6 354	26 811	80.8%
url_id	33 165	0	0.0%
url	33 165	0	0.0%

Missing metadata on the English-language version of Kremlin.ru

A thin line represents an empty field



Source: Giorgio Comai / tadadit.xyz / kremlin.ru_en_2024

Narrative explanation of how this textual dataset was built

Kremlin.ru publishes all of its news items in one ore more of the following sections:

- transcripts
- Presidential Executive Office
- State Council
- Security Council
- Commissions and Councils
- news

This dataset has been generated by parsing each of these sections, similarly to what would be accomplished by insistently clicking on the "show more" link at the bottom of the relevant index pages until the oldest post has been reached.

Some items are posted in more than one section with different urls; they however keep the same internal id: a series of up to 5 digits included at the end of each url. For example, the article "Meeting with permanent members of the Security Council" has been posted on 4 February 2011 at both of the following urls:

- http://en.kremlin.ru/events/president/news/10235
- http://en.kremlin.ru/events/security-council/10235

In order to prevent duplication of contents, only one of these articles is preserved in the final dataset; for consistency, only the first match, according to the order in which sections are listed above, is kept. This allows to see easily which posts are defined as "transcripts" and gives precedence to more specific sections (the generic "news" is used only if the given item was not posted in previous sections). This choice should be substantively irrelevant for most use cases, as all sections are anyway included in a separate field.

Dataset cleaning and reordering

The following steps are conducted on the original dataset before exporting:

- ensure all items have a date
- ensure no post following the cut-off date (2023-12-31) is included
- introduce a `doc_id` column (composed of the website base url, the language of the dataset, and the `url_id`) and set this as the first column of the dataset

Useful links

- the Russian-language version of this corpus: kremlin.ru_ru_2024
- a detailed walkthrough of the technicalities involved in creating this corpus: Extracting textual contents from the Kremlin's website with castarter
- a blog post using a previous version of this dataset: Russophobia in Russian official statements and media

License information

The footer of kremlin.ru as well as the dedicated copyright page make clear that:

> "all materials published on this website are available with the following license"Creative Commons Attribution 4.0 International"

This license gives the right to "copy and redistribute the material in any medium or format", and to "remix, transform, and build upon the material for any purpose, even commercially", as long as appropriate credit is given to the source and the license is included.

The contents of this dataset - "kremlin.ru_en" - are distributed within the remits of this license. To the extent that it is possible, the dataset itself is also distributed by its creator, Giorgio Comai, with the same CC-BY license, as well as under the Open Data Commons Attribution license (ODC-BY).

Funding and disclaimers

This dataset and accompanying materials have been produced within the scope of a project conducted with the support of the Unit for Analysis, Policy Planning, Statistics and Historical Documentation - Directorate General for Public and Cultural Diplomacy of the Italian Ministry of Foreign Affairs and International Cooperation, in accordance with Article 23 – bis of the Decree of the President of the Italian Republic 18/1967.

The views expressed are solely those of the authors and do not necessarily reflect the views of the Ministry of Foreign Affairs and International Cooperation.