kremlin.ru_ru_2024

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

Giorgio Comai (OBCT/CCI)

2024-09-16

Scope of this dataset

This textual dataset is based on kremlin.ru, i.e. the Russian-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_ru_2024

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

Start date: 1999-12-31

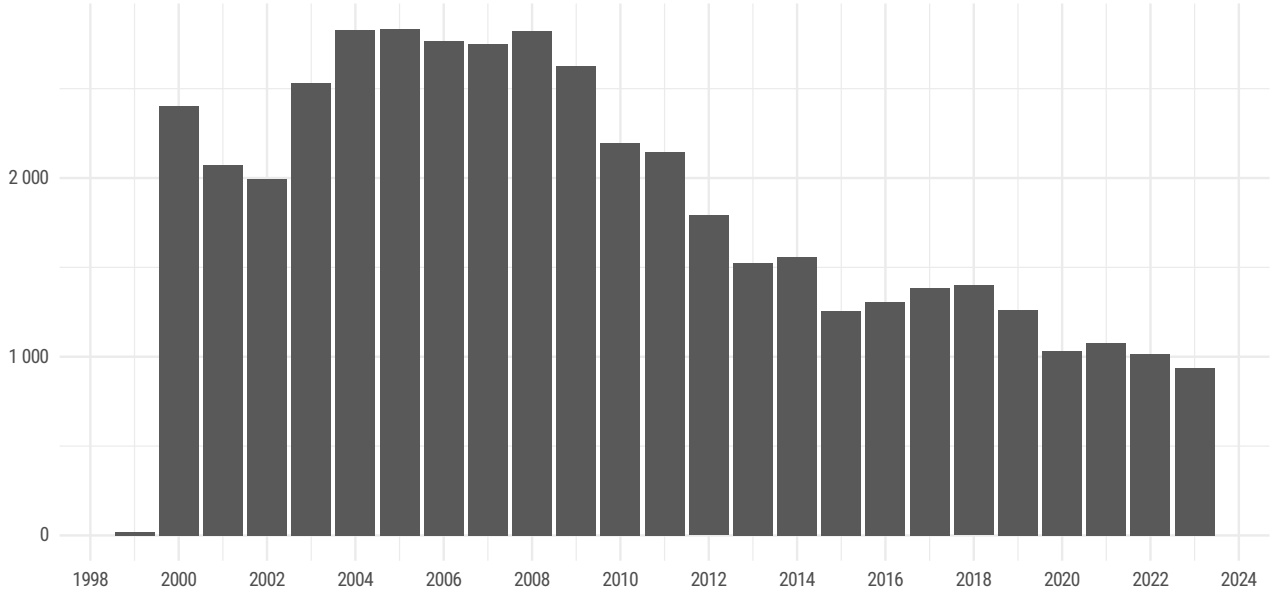
End date: 2023-12-31

Total items: 45 538

Available columns: doc_id; text; title; date; time; datetime; announcement; description; sections; themes; themes_id; persons; persons_id; countries; countries_id; location; location_w_qid; location_w_label_en; location_w_description_en; location_w_label_ru; location_w_description_ru; location_w_country_qid; location_w_country_name; location_w_country_code; location_w_latitude; location_w_longitude; url_id; url

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

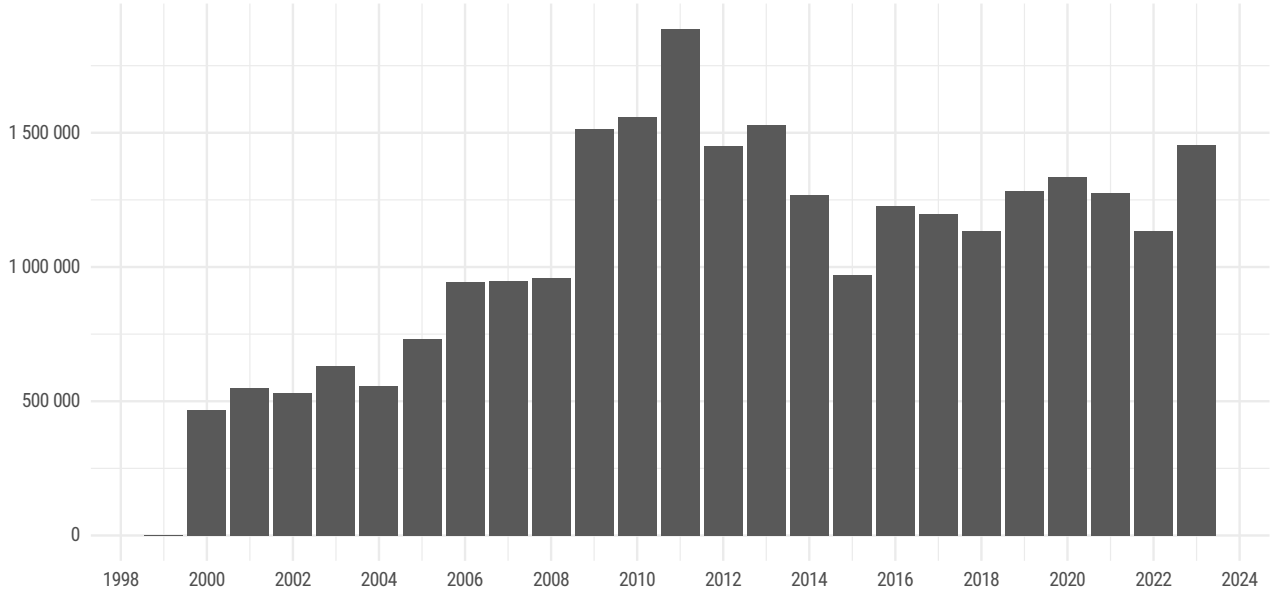
Based on 45 538 items published between 31 December 1999 and 31 December 2023



Source: Giorgio Comai / tadadit.xyz / kremlin.ru_ru_2024

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

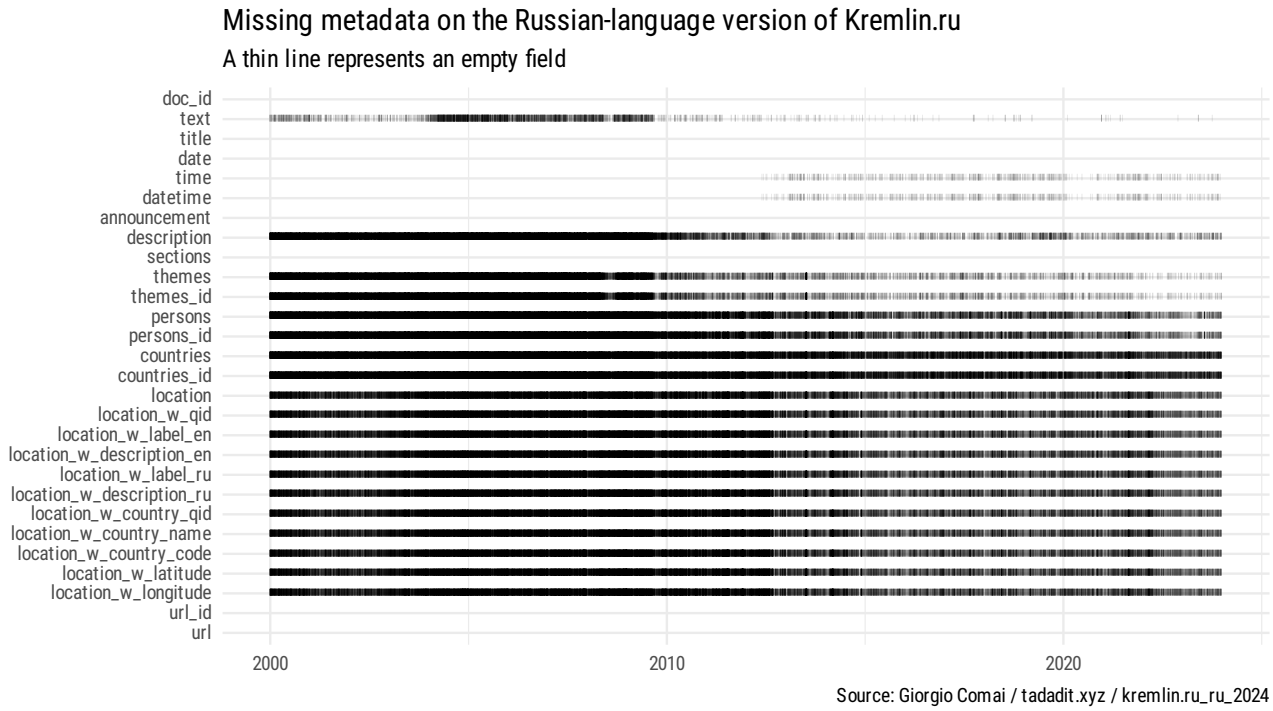
Based on 45 538 items published between 31 December 1999 and 31 December 2023



Source: Giorgio Comai / tadadit.xyz / kremlin.ru_ru_2024

Missing data

field	present	missing	missing_share
doc_id	45 538	0	0.0%
text	41 572	3 966	8.7%
title	45 538	0	0.0%
date	45 538	0	0.0%
time	44 925	613	1.3%
datetime	44 925	613	1.3%
announcement	45 538	0	0.0%
description	17 462	28 076	61.7%
sections	45 538	0	0.0%
themes	19 652	25 886	56.8%
themes_id	19 652	25 886	56.8%
persons	11 040	34 498	75.8%
persons_id	11 050	34 488	75.7%
countries	6 409	39 129	85.9%
countries_id	6 410	39 128	85.9%
location	22 321	23 217	51.0%
location_w_qid	21 282	24 256	53.3%
location_w_label_en	21 270	24 268	53.3%
location_w_description_en	21 215	24 323	53.4%
location_w_label_ru	21 259	24 279	53.3%
location_w_description_ru	20 888	24 650	54.1%
location_w_country_qid	21 250	24 288	53.3%
location_w_country_name	21 250	24 288	53.3%
location_w_country_code	21 237	24 301	53.4%
location_w_latitude	21 199	24 339	53.4%
location_w_longitude	21 199	24 339	53.4%
url_id	45 538	0	0.0%
url	45 538	0	0.0%



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://kremlin.ru/events/president/news/10235
- http://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

Available fields and data issues

Each post published on the Kremlin's website often includes additional fields, besides `title` and `date`. Whenever they are present, they are included in this dataset in a dedicated column (see above for more information about availability of such data).

These fields include:

- `text` - text is mostly available, but there are 3 966 items where only title (and occasionally other fields such as description and location) are available
- `announcement` - can either be TRUE, or FALSE. When TRUE, the relative item refers to posts marked as announcements of (usually) meetings planned to happen. They are marked with "Анонс" on the official website and indicate a future date. There are in total 613 such posts; the researcher may want to remove posts marked as "announcement" before processing the corpus. See this post as an example.
- `description` - a brief text, usually summarising the post
- `sections` - one or more website section where the post has been published. If more than one, sections are comma-separated. This is a full list of sections found in the corpus: Выступления и стенограммы, Новости, Документы, Комиссии и Советы, Администрация Президента, Поручения, Государственный Совет, Совет Безопасности, Деятельность Президента, Для СМИ, Отчёты, Поручения Президента, Аккредитация, Анонсы
- tags, including `themes` and `themes_id`,

- at the bottom of the post, the Kremlin's website often includes a few tags, separated by type. Each post can have one or more of these. These include `themes`, `persons`, and `countries`. For example, an announcement post about a forthcoming visit to Belarus president Lukashenko will have "foreign policy" marked as a theme, Lukashenko as "person", and "Belarus" as country. In the dataset, where more than one string is present, they are separated by a semi-column (;). The dataset includes also additional `themes_id`, `persons_id`, and `countries_id` column: these are unique identifiers that refer to dedicated pages on the website. For example, in the case of this post Lukashenko has `person_id` 119, corresponding to https://kremlin.ru/catalog/persons/119/, where all meetings with Lukashenko are expected to be stored; the theme, "foreign policy", has `theme_id` 82, corresponding to http://kremlin.ru/catalog/keywords/82/, and the `country_id` is "BY", corresponding to http://kremlin.ru/catalog/countries/BY/. No additional checks have been conducted to verify how complete these fields are (e.g. if effectively Lukashenko is tagged is correctly tagged in each and every meeting involving him, or whenever he is mentioned)
- `location` - next to the date, at the top of the post, often reference is made to the location from where the post is supposedly issued. For example, this post announcing that Putin arrived in Australia for a meeting, includes "Brisbane" ("Брисбен") written next to the date.

Enriching the dataset through Wikidata

The `location` field refers to locations in free text format: these include name of specific halls, of specific buildings, of villages, of cities, of regions, or generic references to countries. As such, they are difficult to parse, i.e., to identify that "Брисбен" is a city located in Australia. In order to enrich these data, each of these locations has been tentatively associated with a Wikidata identifier. Wikidata is a sort of database back-end associated with Wikipedia that enables retrieving data systematically. Based on this matching, a number of additional pieces of information have been added as separated fields. Each of these fields includes `_w_` in the column name to clarify it comes from Wikidata, not from the source website itself.

- `location_w_qid` - is the Q identifier used by Wikidata. For example, Brisbane corresponds to Q34932
- `location_w_label_en`, `location_w_description_en`: this is how Wikidata identifies and describes (in English) the given item
- `location_w_label_ru`, `location_w_description_ru`: this is how Wikidata identifies and describes (in Russian) the given item

- `location_w_country_qid`: this is the country of the given location according to Wikidata (Wikidata property "P17" of the location). `location_w_country_name` and `location_w_country_code` are derived from here.
- `location_w_latitude` and `location_w_longitude` are the coordinates of the location. Be mindful that this can be very specific (e.g. a specific building) or very generic. For example, if only "Canada" is given a location, the coordinates will be a point in the middle of Canada.

An interactive map with all geo-located post and links to the original source for each of them can be explored by following this link or by opening with the a browser the file `kremlin.ru_ru_2024_posts_by_location.html`, made available along with this release. Items including the string "Москва" ("Moscow") in the location string have been removed for clarity.

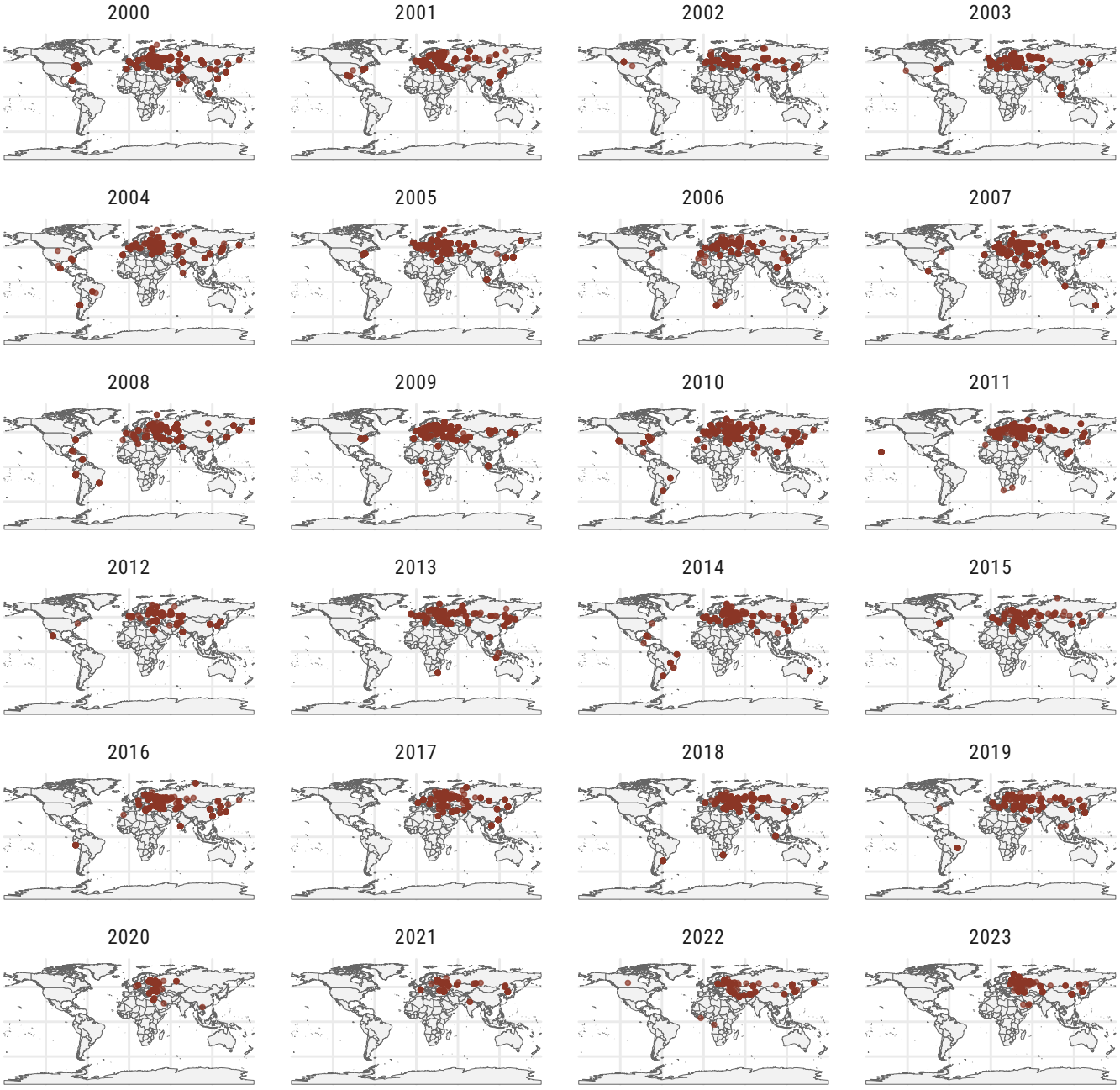
Even if the full list of locations has been manually inspected, it is highly likely that due to the large number of items involved this dataset still includes some matching errors. Additional data quality checks are recommended if these fields are central to the analysis.

The data user should also be aware that location included in the post may not necessarily refer to an official visit. No full matching with this List of international presidential trips made by Vladimir Putin is expected.

In at least some instances, presidential aids or envoys are also mentioned: for example, if you notice a meeting in Canada in December 2022, this does not refer to a Putin visit, but to the participation of a presidential advisor to a conference.

Users considering retrieval of additional properties from Wikidata may consider relying on the R package `tidywikidatar`.

Locations attached to statements on Russia's president website (Kremlin.ru, Russian Version)



Source: Giorgio Comai / tadadit.xyz / kremlin.ru_ru_2024

Useful links

- the English-language version of this corpus: kremlin.ru_en_2024
- a detailed walkthrough of the technicalities involved in creating this corpus: Extracting textual contents from the Kremlin's website with castarter
- an 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_ru" - 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.