archive.government.ru_ru_2024

Corpus based on the archived version of Russia's government website (in Russian, 2008-2013)

Giorgio Comai (OBCT/CCI)

2024-09-16

Scope of this corpus

This corpus is based on all contents published in the "news", "transcripts", and "telegrams" sections of the website archive.government.ru as it was available online in early 2024.

This website is an archived, static version of Russia's government former website, including only posts for the period May 2008 to May 2013, starting with Vladimir Putin becoming prime minister but including one year after he returned to the presidency and Dmitri Medvedev was prime minister. This is in spite of the fact that the current government website suggests this archived version should cover the period: "07.05.2008-07.05.2012".

Users should be aware that broadly for the same period (specifically, the time during which Vladimir Putin was prime minister) a separate website for the prime minister was maintained, and its archived version is still available online at archive.premier.gov.ru.

Summary statistics

Dataset name: archive.government.ru_ru_2024

Dataset description: all news items published on archive.government.ru

Start date: 2008-05-07

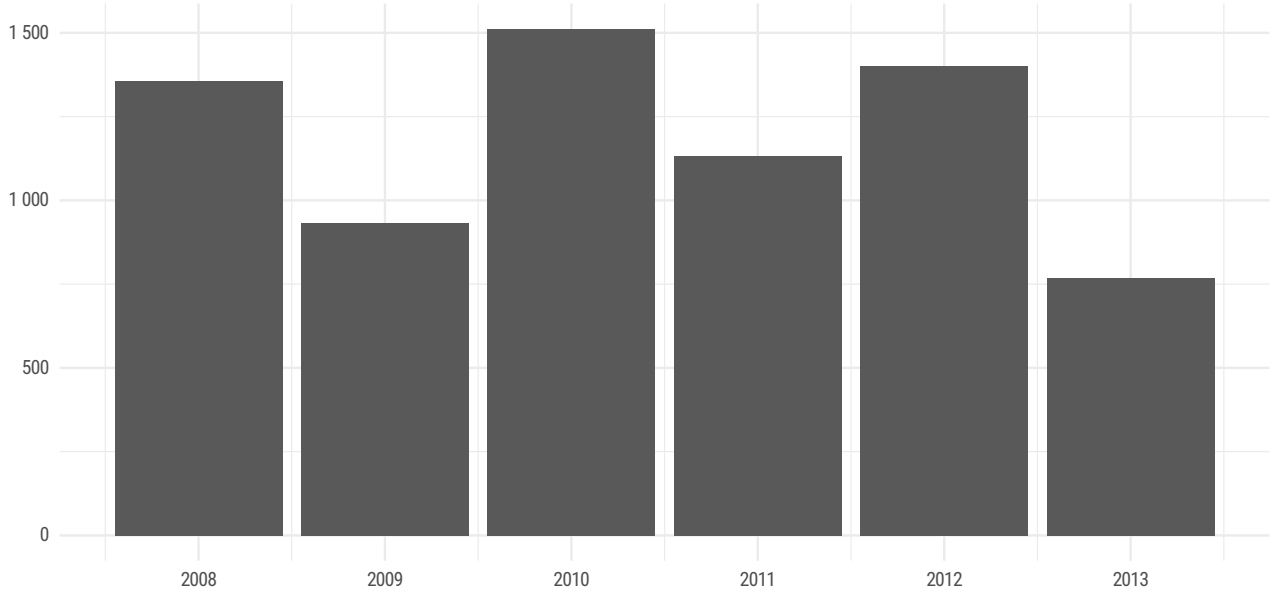
End date: 2013-05-21

Total items: 7 103

Available columns: doc_id; text; title; date; internal_id; section; participants; url

Number of items per year published on archive.government.ru

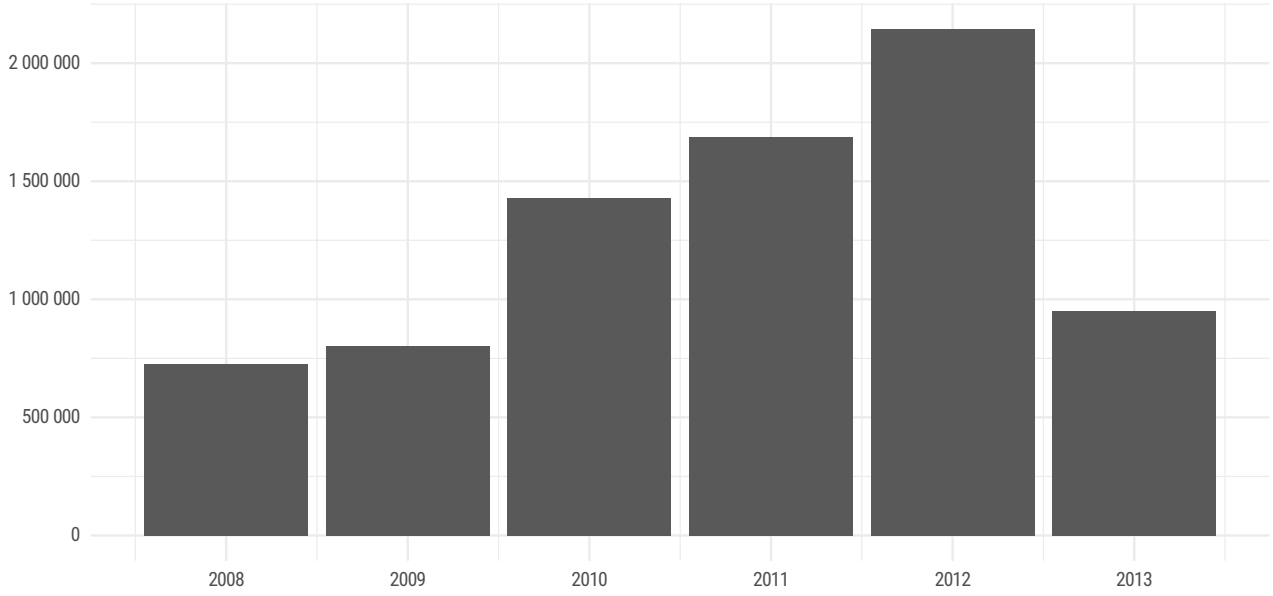
Based on 7 103 items published between 7 May 2008 and 21 May 2013



Source: Giorgio Comai / tadadit.xyz / archive.government.ru_ru_2024

Number of words per year published on archive.government.ru

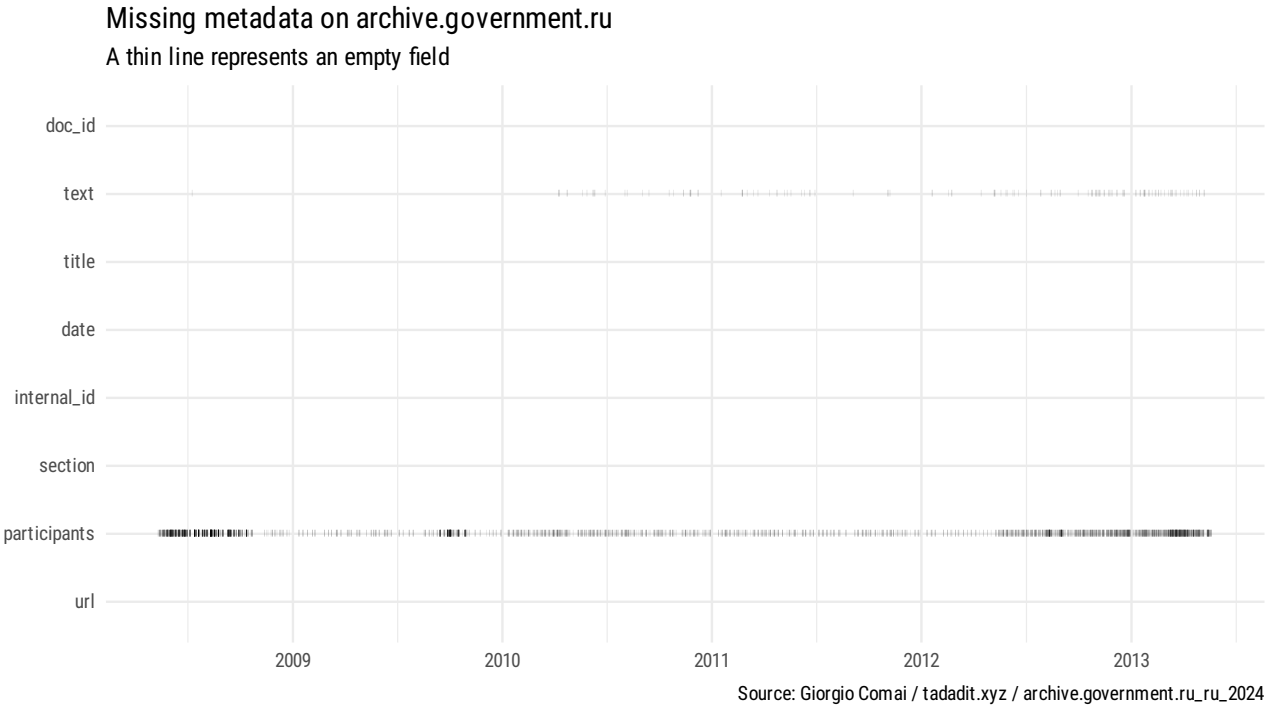
Based on 7 103 items published between 7 May 2008 and 21 May 2013



Source: Giorgio Comai / tadadit.xyz / archive.government.ru_ru_2024

Missing data

field	present	missing	missing_share
doc_id	7 103	0	0.0%
text	6 968	135	1.9%
title	7 103	0	0.0%
date	7 103	0	0.0%
internal_id	7 103	0	0.0%
section	7 103	0	0.0%
participants	4 234	2 869	40.4%
url	7 103	0	0.0%



Narrative explanation of how this corpus has been created

This corpus has been built based on index pages of the “news”, “transcripts”, and “telegrams” sections. At the bottom of each index page there is a “show more” button, which shows older posts. The script automatically parses all links for as long as the “show more” button would show older posts.

Text and metadata have been extracted from the resulting pages. Given that the format of the resulting page differed depending on context, a cascade of css extractors was used for some fields. More specifically, “date” was retrieved from either a `<p>` element of class `date`

or from a `<div>` element of class `entry__meta__date`; "text" from either a `<div>` element of class `b06-richtext` or of class `entry__content`.

Many posts include tags to individuals in a dedicated field called "participants" ("*Участники*"), which is included along with other metadata.

Duplicates

Some items have been posted on the same date, with the same title, and mostly with exactly or almost exactly the same text under different sections of the websites. In such cases (250 in total), the one categorised as "transcript" has been kept, the others discarded.

License information

The footer of the website makes clear that all contents available are published with a Creative Commons Attribution 3.0 license:

> Все материалы сайта доступны по лицензии: Creative Common Attribution 3.0"

The contents of this dataset - "archive.government.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.