archive.premier.gov.ru_ru_2024

Corpus based on the archived version of the website of Russia's prime minister (in Russian, 2008-2012)

Giorgio Comai (OBCT/CCI)

2024-09-16

Scope of this corpus

This corpus is based on all contents published in the "news" section of the website archive.premier.gov.ru as it was available online in early 2024.

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 government was maintained, and its archived version is still available online at archive.government.gov.ru.

Summary statistics

Dataset name: archive.premier.gov.ru_ru_2024

Dataset description: all news items published on archive.premier.gov.ru

Start date: 2008-05-07

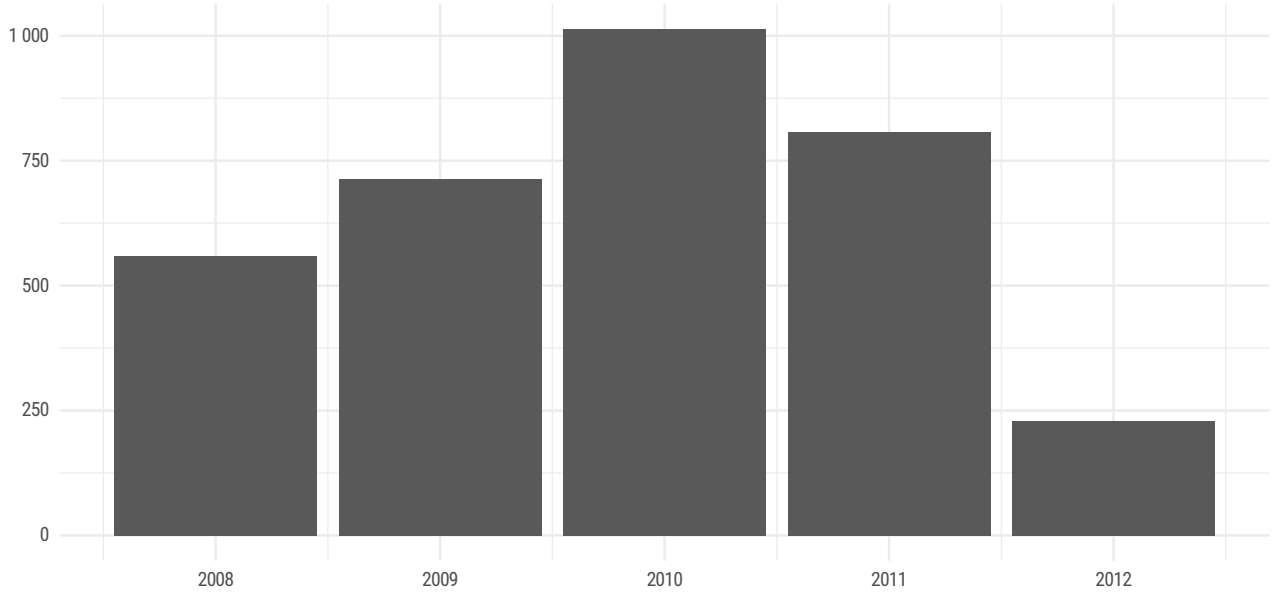
End date: 2012-05-07

Total items: 3 323

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

Number of items per year published on archive.premier.gov.ru

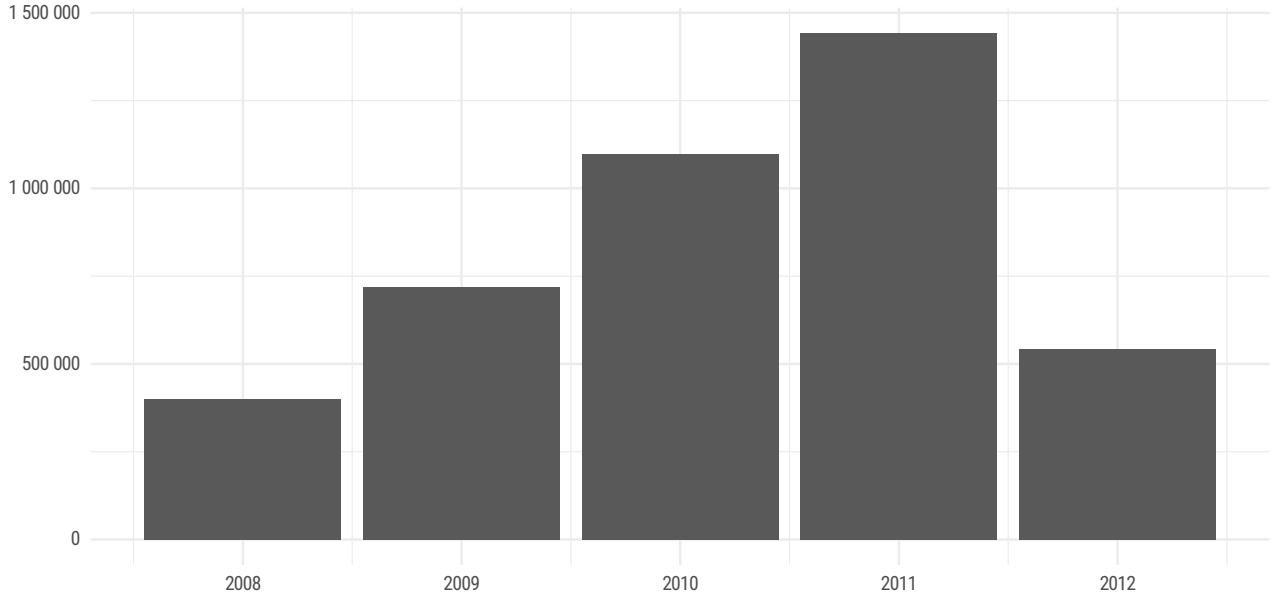
Based on 3 323 items published between 7 May 2008 and 7 May 2012



Source: Giorgio Comai / tadadit.xyz / archive.premier.gov.ru_ru_2024

Number of words per year published on archive.premier.gov.ru

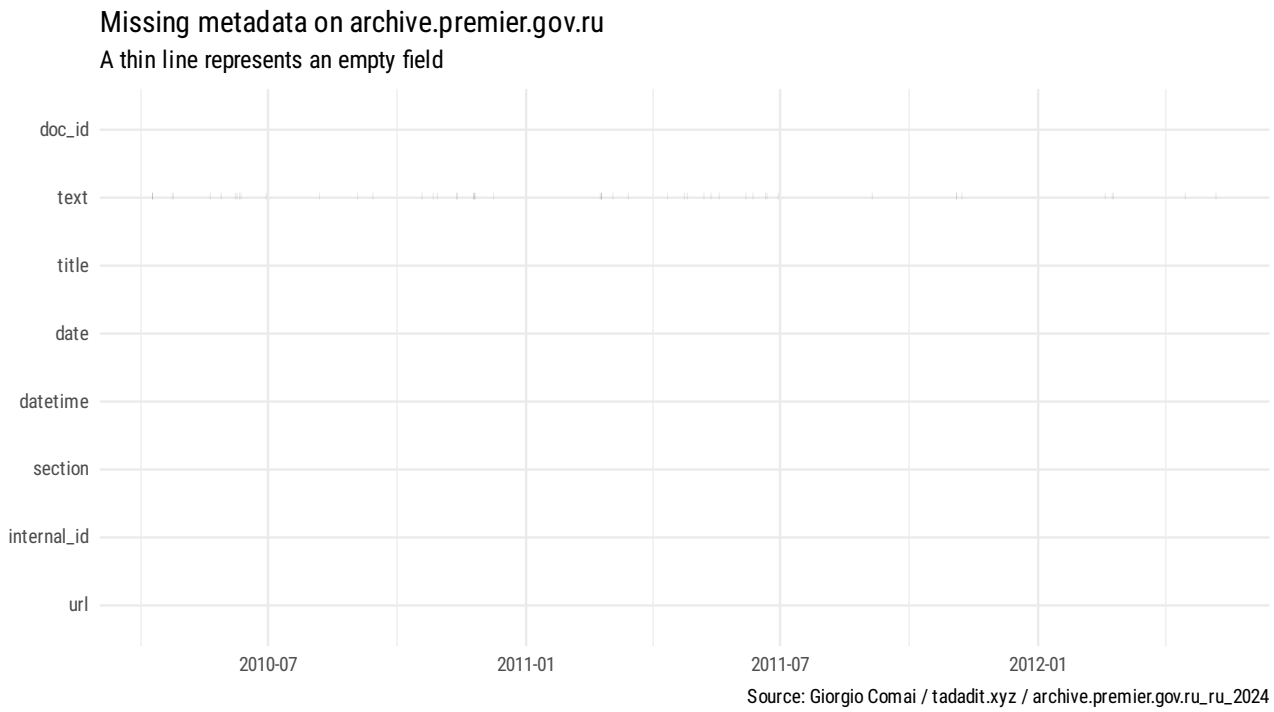
Based on 3 323 items published between 7 May 2008 and 7 May 2012



Source: Giorgio Comai / tadadit.xyz / archive.premier.gov.ru_ru_2024

Missing data

field	present	missing	missing_share
doc_id	3 323	0	0.0%
text	3 272	51	1.5%
title	3 323	0	0.0%
date	3 323	0	0.0%
datetime	3 323	0	0.0%
section	3 323	0	0.0%
internal_id	3 323	0	0.0%
url	3 323	0	0.0%



Narrative explanation of how this corpus has been created

This corpus has been built based on index pages of the event "news" section, retrieving links starting with the earliest publication.

Links to photo, video, and audio pages have been removed, only textual contents have been kept.

Text and metadata have been extracted from the resulting pages.

Duplicates

Some items have been posted on the same date, with the same title, and with the same text under different urls (but the same numeric component in the url, here recorded as `internal_id`). In such cases, duplicates have been removed.

Items with title but no text

There are 51 items with title, but no text. These are kept in the dataset, as the title may still offer relevant contents.

License information

At the time contents were retrieved, the footer of the website makes clear that all contents available are published with a Creative Commons Attribution 3.0 license:

> Creative Commons Attribution 3.0 Непортированная"

The contents of this dataset - "archive.premier.gov.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.