duma.gov.ru_ru_2024

Corpus based on the Russia's Duma website (in Russian, 2006-2023)

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 duma.gov.ru as it was available online in early 2024.

Summary statistics

Dataset name: duma.gov.ru_ru_2024

Dataset description: all news items published on duma.gov.ru

Start date: 2006-04-05

End date: 2023-12-30

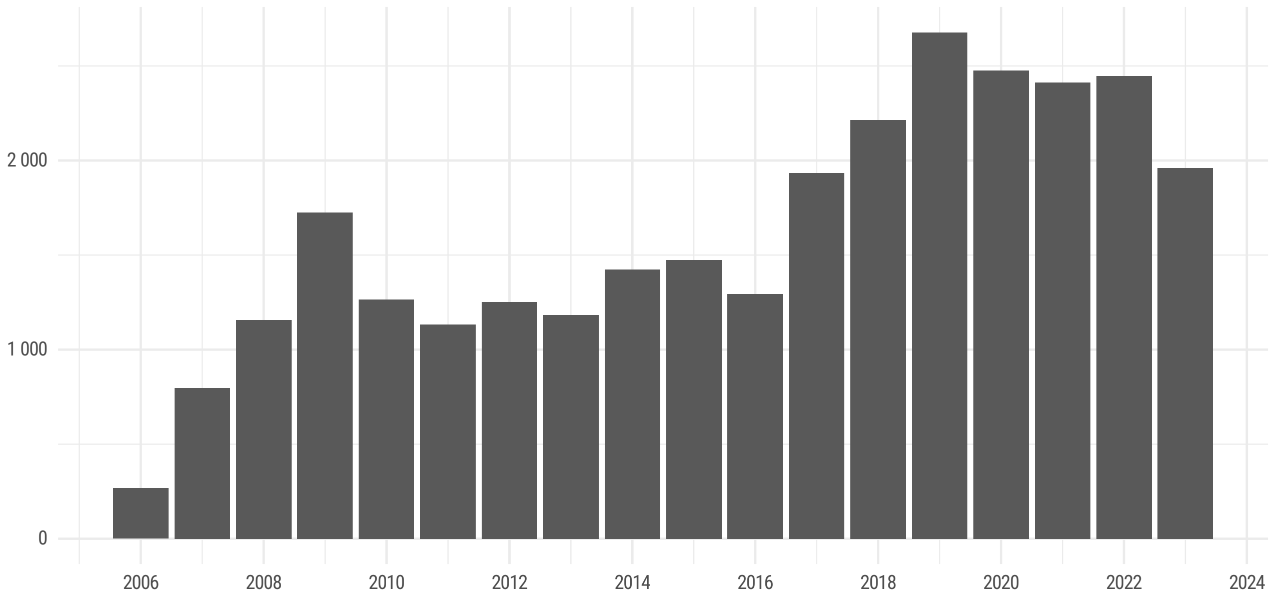
Total items: 29 094

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

License: Creative Commons Attribution 3.0 International

Number of items per year published on duma.gov.ru

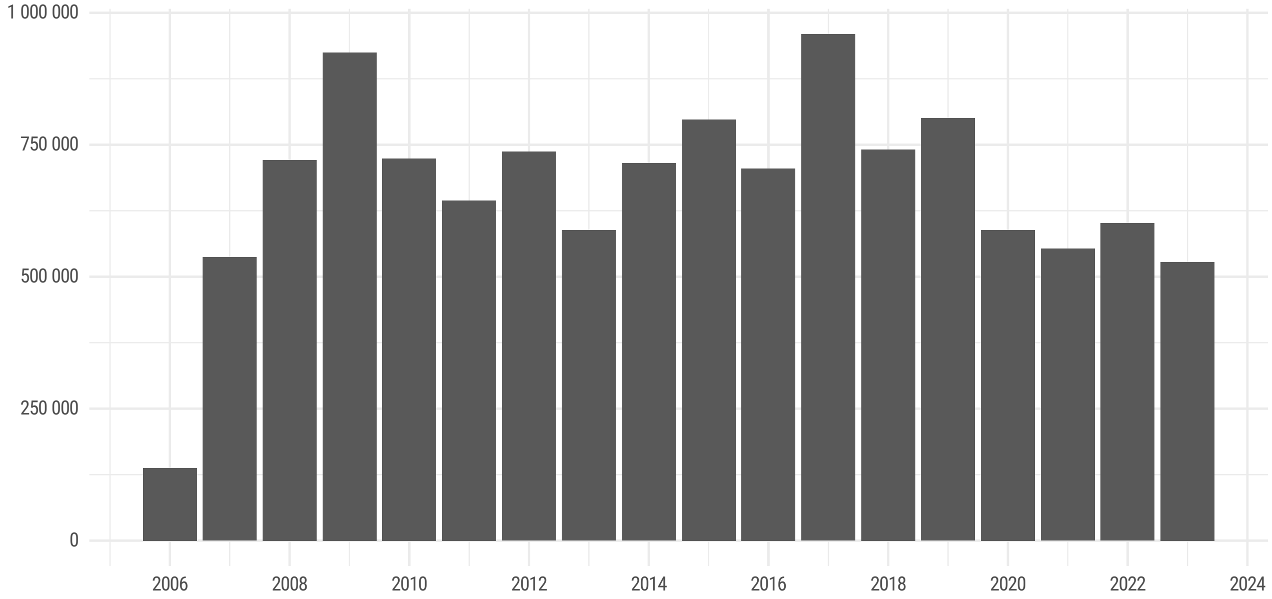
Based on 29 094 items published between 5 April 2006 and 30 December 2023



Source: Giorgio Comai / tadadit.xyz / duma.gov.ru_ru_2024

Number of words per year published on duma.gov.ru

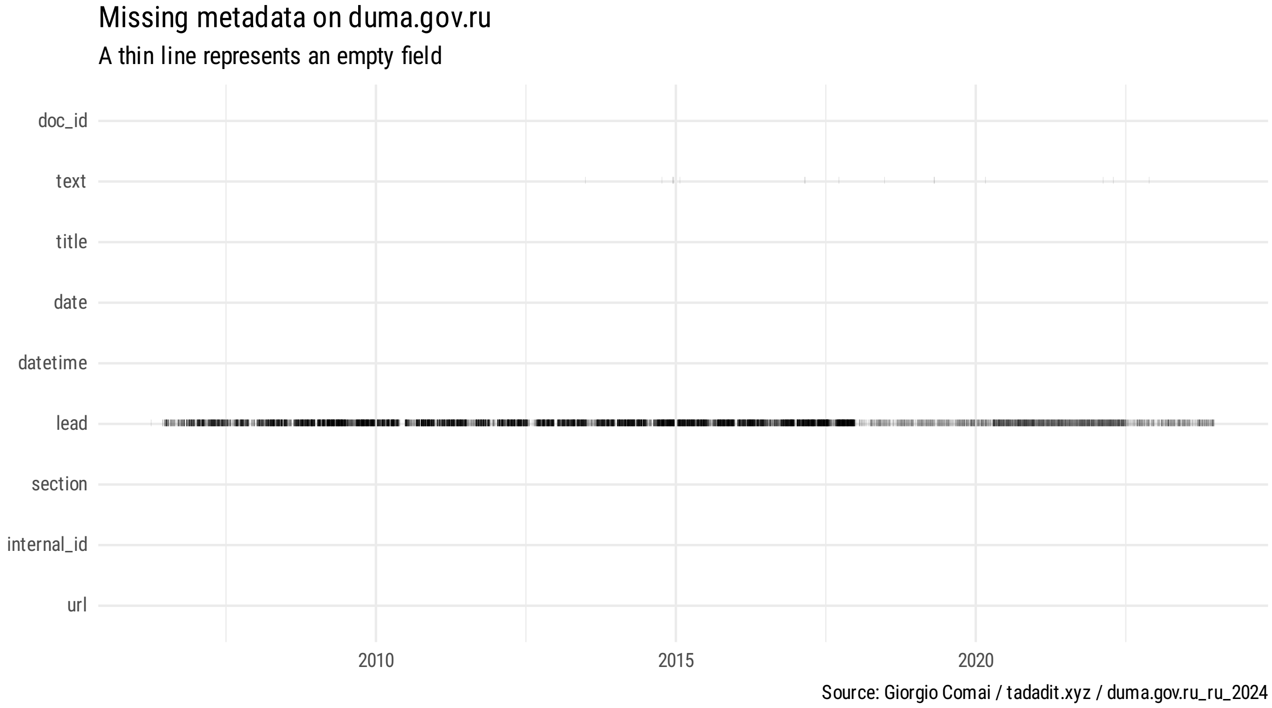
Based on 29 094 items published between 5 April 2006 and 30 December 2023



Source: Giorgio Comai / tadadit.xyz / duma.gov.ru_ru_2024

Missing data

field	present	missing	missing_share
doc_id	29 094	0	0.0%
text	29 078	16	0.1%
title	29 094	0	0.0%
date	29 094	0	0.0%
datetime	29 094	0	0.0%
lead	11 744	17 350	59.6%
section	29 094	0	0.0%
internal_id	29 094	0	0.0%
url	29 094	0	0.0%



Narrative explanation of how this corpus has been created

This corpus has been built based on index pages of the “news” section of the website, parsing older posts as they would appear when clicking on the “Загрузить предыдущие материалы” button.

Text and metadata have been extracted from the resulting pages, relying on the well structured format of the news pages, presenting each element in a dedicated element:

- the title is always included in a `<h1>` element of class `article__title`
- the date and datetime are retrieved from `time` container, `datetime` attribute, `datePublished` item proposition
- the lead is included (when available) in a `<div>` element of class `article__lead`
- the section is included in a `<a>` element of class `article__caption`
- the main text is always included in a `<div>` element of class `article__content`

Data cleaning

All items published on the website include a date of publication. The `lead` is quite often missing, as appears from the summary information above. There are, in total, 16 items with an empty text field; 9 of them have also an empty `lead` field. This is not due to data retrieval issues, but rather, to the original contents themselves, which often expect the title - perhaps accompanied by a picture - to be self-explanatory. See an example.

License information

The about page of the website includes a section "On the use of information" ("Об использовании информации"), which clarifies the permissive conditions for re-publishing contents used on the website. Even if it does not include reference to specific license, it unambiguously states that contents can be published anywhere, without any sort of limitation, with the only condition being that a link the original source must be included.

> Все материалы официального сайта Государственной Думы Федерального Собрания Российской Федерации могут быть воспроизведены в любых средствах массовой информации, на серверах сети Интернет или на любых иных носителях без каких-либо ограничений по объему и срокам публикации. Это разрешение в равной степени распространяется на газеты, журналы, радиостанции, телеканалы, сайты и страницы сети Интернет. Единственным условием перепечатки и ретрансляции является ссылка на первоисточник. Никакого предварительного согласия на перепечатку со стороны Аппарата Государственной Думы не требуется.

The contents of this dataset - "duma.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 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.