# government.ru_ru_2024

**Corpus based on the Russia's government website (in Russian, 2013-2023)**

**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 government.ru as it was available online in early 2024.

## Summary statistics

**Dataset name**: government.ru_ru_2024

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

**Start date**: 2012-04-24

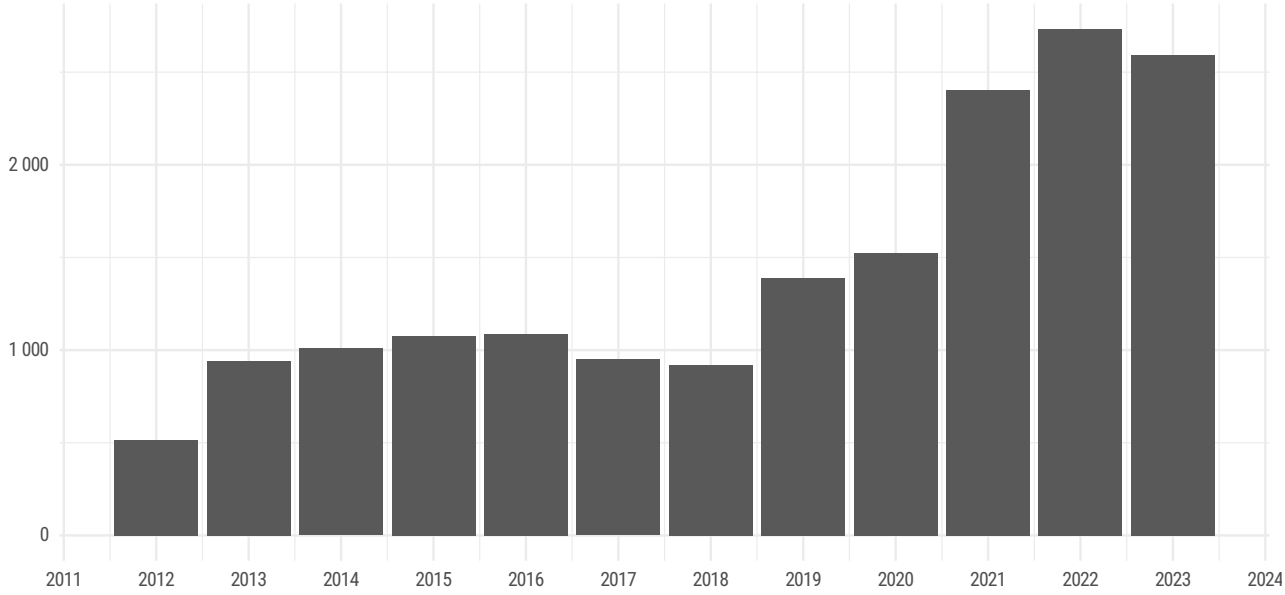
**End date**: 2023-12-30

**Total items**: 17 135

**Available columns**: doc_id; text; title; date; internal_id; time; place; tags; url

Number of items per year published on government.ru

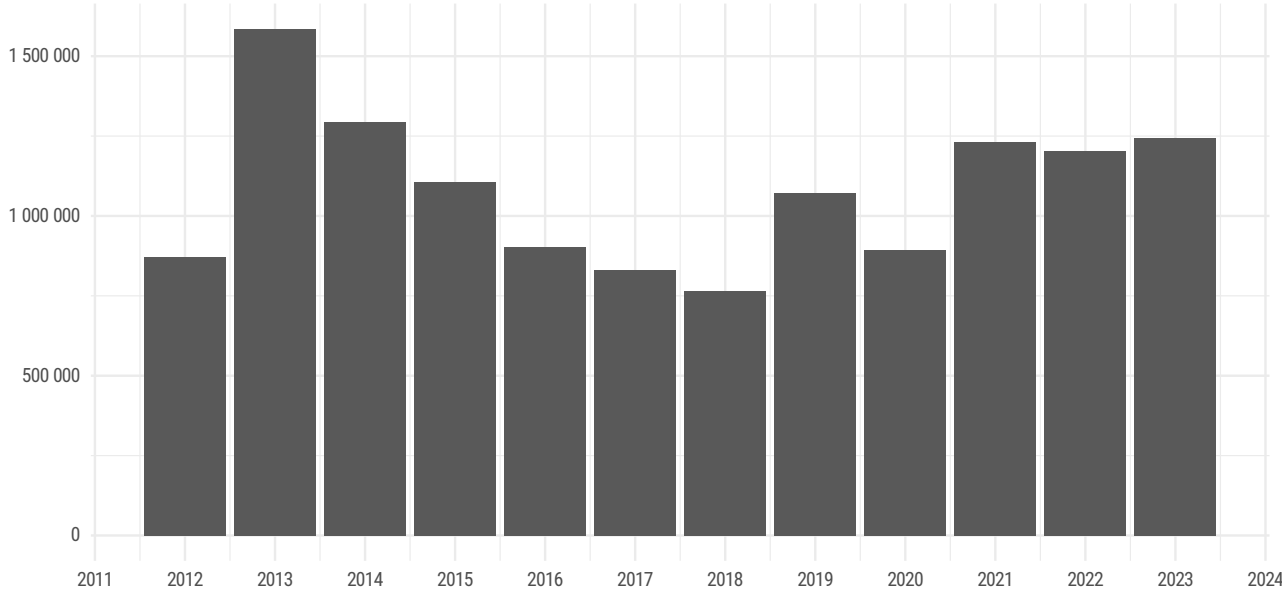
Based on 17 135 items published between 24 April 2012 and 30 December 2023




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

Number of words per year published on government.ru

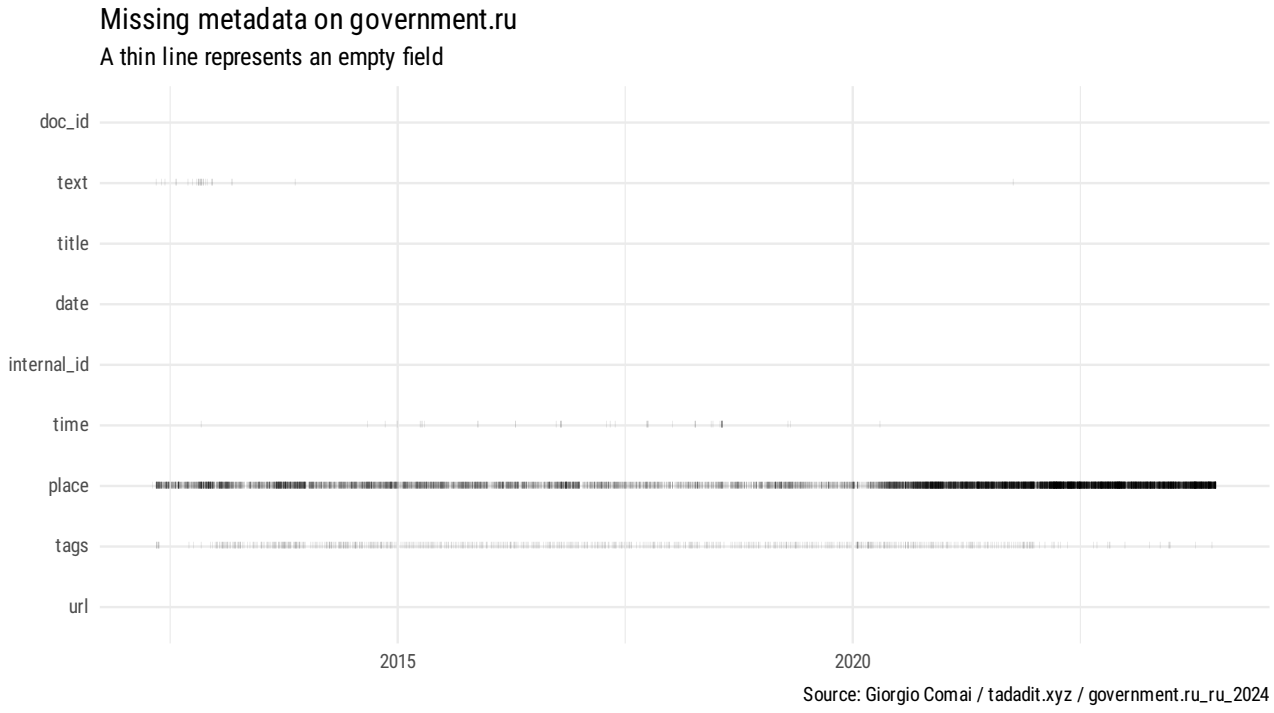
Based on 17 135 items published between 24 April 2012 and 30 December 2023




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

## Missing data

| field | present | missing | missing_share |
|---|---|---|---|
| doc_id | 17 135 | 0 | 0.0% |
| text | 17 107 | 28 | 0.2% |
| title | 17 135 | 0 | 0.0% |
| date | 17 135 | 0 | 0.0% |
| internal_id | 17 135 | 0 | 0.0% |
| time | 17 091 | 44 | 0.3% |
| place | 6 299 | 10 836 | 63.2% |
| tags | 16 439 | 696 | 4.1% |
| url | 17 135 | 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 be auto-loaded when scrolling the news index page.

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 `<h3>` element of class `reader_article_headline`

- the date is always included in a `<span>` element of class `reader_article_dateline__date`
- the time is always included in a `<span>` element of class `reader_article_dateline__time`
- the place is always included in a `<span>` element of class `entry__meta__date__place`
- the place is always included in a `<li>` element of class `reader_article_tags_item`
- the main text is always included in a `<div>` element of class `reader_article_body`

#### Data cleaning

Among all news items extracted, only two do not have a date (one and two), seemingly because they effectively link to other contents. They have been removed from the dataset.

Besides, two items have a date of publication set many years before all other contents available on the website (one and two). They have also been removed for clarity.

There are 28 items with date, title, and tags, but an empty text field. They mostly refer to meetings such as this one; the titles have a format similar to "Medvedev met gubernor of X", and such, with no additional content shared. They are maintained in the dataset, as title and tags may still contain useful information.

## 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 4.0"

The contents of this dataset - "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.