mid.ru_ru_2024

Corpus based on the website of Russia's MFA (in Russian, 2003-2023)

Giorgio Comai (OBCT/CCI)

2024-09-16

Scope of this corpus

This corpus includes all news items published on the Russian-language version of the website of Russia's MFA.

Summary statistics

Dataset name: mid.ru_ru_2024

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

Start date: 2003-01-02

End date: 2023-12-31

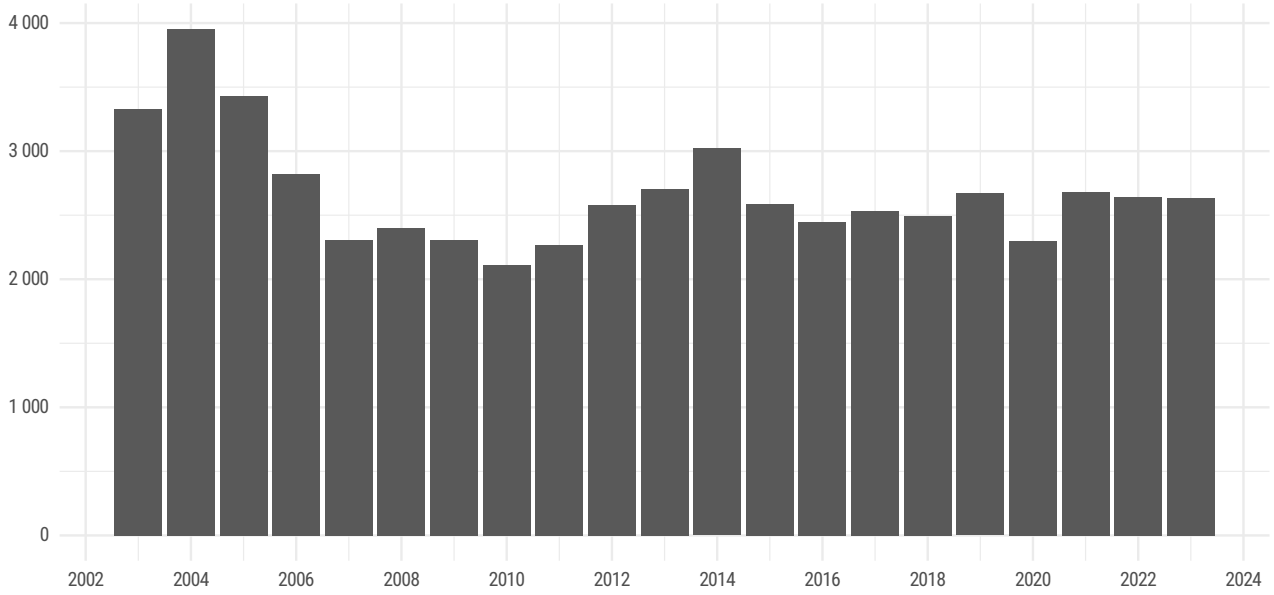
Total items: 56 203

Available columns: doc_id; text; date; datetime; title; internal_id; url_id; translations; countries; url

License: Permissive (see details)

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

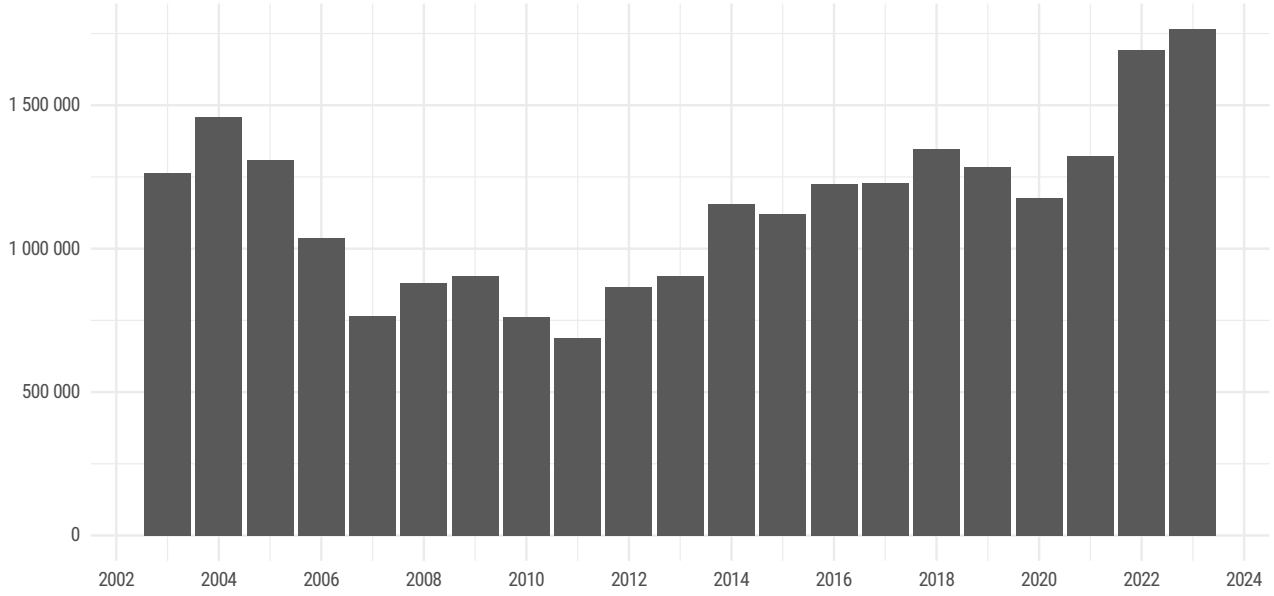
Based on 56 203 items published between 2 January 2003 and 31 December 2023



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

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

Based on 56 203 items published between 2 January 2003 and 31 December 2023



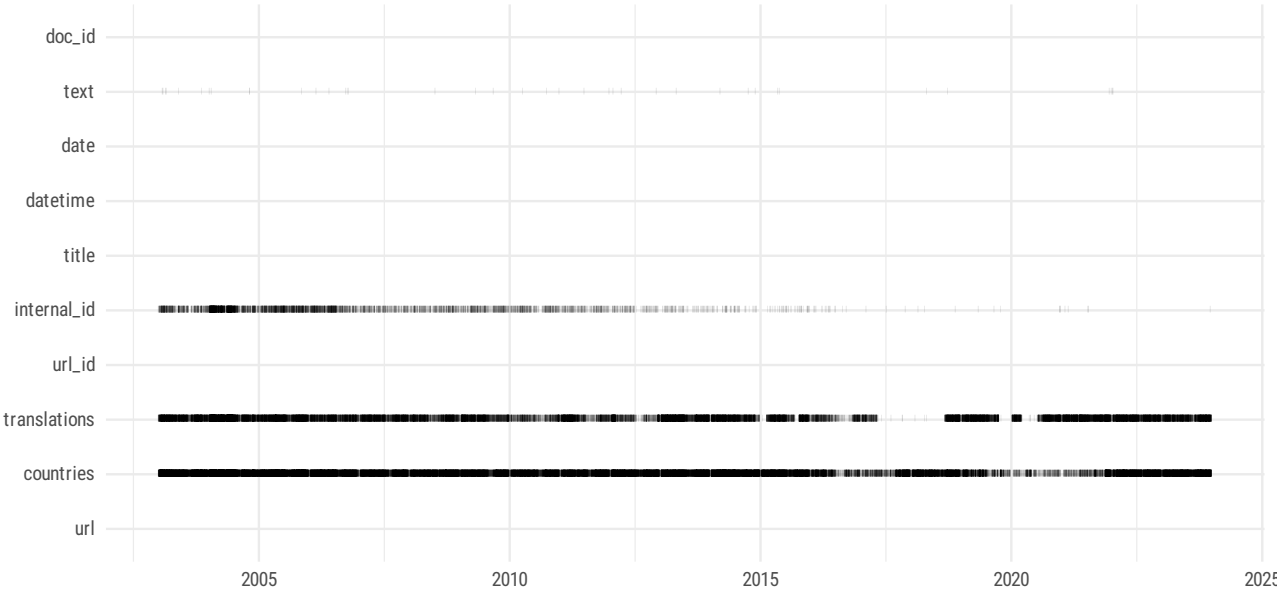
Source: Giorgio Comai / tadadit.xyz / mid.ru_ru_2024

Missing data

field	present	missing	missing_share
doc_id	56 203	0	0.0%
text	56 164	39	0.1%
date	56 203	0	0.0%
datetime	56 203	0	0.0%
title	56 203	0	0.0%
internal_id	51 330	4 873	8.7%
url_id	56 203	0	0.0%
translations	27 589	28 614	50.9%
countries	8 888	47 315	84.2%
url	56 203	0	0.0%

Missing metadata on the Russian-language version of mid.ru

A thin line represents an empty field



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

Narrative explanation of how this textual dataset was built

The website of Russia's MFA makes it possible to search in its news section by date. All index pages for each date starting with earliest publications have been retrieved. In the few occasions when more than 20 items were published on the same day, a second page for the relevant day was also retrieved. Here is an example of such an index page:

- https://www.mid.ru/ru/foreign_policy/news/?activeFrom=22.09.2011&activeTo=22.09.2011

Direct links to news items were extracted from these pages.

The corpus includes the limited metadata available through the website, namely:

- title
- date and time of publication
- an internal id which is included in almost all posts (see note below)
- a list of the languages in which a given post has been published

Notes

This section lists some issues that may be of interest to users of this corpus

- Along with news, the MFA publishes items that detail the timing and accreditation rules for press briefings, see for example: https://mid.ru/ru/foreign_policy/news/1927386/. As these do no not include substantive contents, they are not included in the dataset.
- Almost all news items are published with an identifier, e.g. "1383-22-09-2011" for this item. In many instances, in particular in the earlier years, the identifier is missing, and in a handful it is not unique. As a consequence, the numeric component of the url is likely preferable as the main unique identifier.
- The Russian-language version of this corpus has a significantly larger number of publications.
- There are 39 items with empty text fields. Indeed, they simply include no text besides the title or include just a link to an external file (not included in this corpus).

License information

At the time contents were retrieved, the page on the conditions for the use of website contents makes clear that contents can be used for research purposes and can be republished, as long as reference is always made to the website of the MFA.

> Materials on the website of the Russian Ministry of Foreign Affairs are generally accessible and open for non-commercial use (personal, family, education, research, etc.).
>
> Their reprinting, as well as any quoting in the mass media is allowed only with a reference to the website of the Russian Ministry of Foreign Affairs as a source of the information.

No specific license is however mentioned.

The contents of this dataset - “mid.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, at the same conditions, 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.