DISCUSS DATA

Archiving, sharing and discussing research data on Eastern Europe, South Caucasus and Central Asia

Russian state institutions full-text datasets

A collection of corpora based on contents extracted from the websites of Russian state institutions

Version 1.0, published: 2024-10-29 15:03:13.665362+00:00

https://discuss-data.net/dataset/0578d7fe-35f7-4e9e-a29d-926618a5c6bd/

Data Citation

Giorgio Comai (2024): Russian state institutions full-text datasets – A collection of corpora based on contents extracted from the websites of Russian state institutions, v. 1.0, Discuss Data, https://doi.org/10.48320/0578D7FE-35F7-4E9E-A29D-926618A5C6BD

Description

This is a collection of full-text datasets based on contents extracted from the websites of Russian state institutions.

All datasets do not include items published after 31 December 2023.

These datasets have been introduced in the following book chapter, which offers additional context:

> Comai, Giorgio (2025, forthcoming), “Text-mining on-line sources from Russia openly”, in **Autocracy, Influence, War: Russian Propaganda Today**, edited by Paul Goode

The name of each corpus is composed of the bare domain name, a two letter code of the main language of the contents, and the year of release of the dataset, separated by an underscore, e.g. kremlin.ru_ru_2024 for the Russian-language version of Kremlin.ru.

This release includes the following websites:

- Russia’s president, kremlin.ru, in English, filename: kremlin.ru_en_2024, from 1999-12-31 to 2023-12-31. Items included: 33 165
- Russia’s president, kremlin.ru, in Russian, filename: kremlin.ru_ru_2024, from 1999-12-31 to 2023-12-31. Items included: 45 538

- Russia’s MFA, mid.ru, in English, filename: mid.ru_en_2024, from 2003-01-04 to 2023-12-31. Items included: 25 943
- Russia’s MFA, mid.ru, in Russian, filename: mid.ru_ru_2024, from 2003-01-02 to 2023-12-31. Items included: 56 203
- Russia’s government, government.ru, in Russian, filename: government.ru_ru_2024, from 2006-06-22 to 2023-12-30. Items included: 17 135
- Russia’s government (archived version), archive.government.ru, in Russian, filename: archive.government.ru_ru_2024, from 2008-05-07 to 2013-05-21. Items included: 7 103
- Russia’s prime minister (archived version), archive.premier.gov.ru, in Russian, filename: archive.premier.gov.ru_ru_2024, from 2008-05-07 to 2012-05-07. Items included: 3 323
- Russia’s Duma, duma.gov.ru, in Russian, filename: duma.gov.ru_ru_2024, from 2006-04-05 to 2023-12-30. Items included: 29 094
- Russia’s Duma (transcripts), transcript.duma.gov.ru, in Russian, filename: transcript.duma.gov.ru_ru_2024, from 1994-01-11 to 2023-12-15. Items included: 6 032

File formats: compressed csv files (.csv.gz); Open Document Spreadsheets (.ods)

A web version of the documentation accompanying this release is available online: https://tadadit.xyz/datasets/2024/russian_institutions_2024/

Explore through a basic web interface: https://explore.tadadit.xyz/2024/ru_institutions_2024/

Files

- 0-about_this_release.pdf, 42981 bytes
- archive.government.ru_ru_2024.csv.gz, 23408143 bytes
- archive.government.ru_ru_2024.ods, 23714516 bytes
- archive.government.ru_ru_2024.pdf, 59528 bytes
- archive.premier.gov.ru_ru_2024.csv.gz, 13170076 bytes
- archive.premier.gov.ru_ru_2024.ods, 13329233 bytes
- archive.premier.gov.ru_ru_2024.pdf, 46455 bytes
- duma.gov.ru_ru_2024.csv.gz, 35266299 bytes
- duma.gov.ru_ru_2024.ods, 36387387 bytes
- duma.gov.ru_ru_2024.pdf, 78682 bytes
- government.ru_ru_2024.csv.gz, 41292047 bytes
- government.ru_ru_2024.ods, 42106390 bytes
- government.ru_ru_2024.pdf, 71005 bytes
- kremlin.ru_en_2024.csv.gz, 36907849 bytes
- kremlin.ru_en_2024.ods, 39315208 bytes

- kremlin.ru_en_2024.pdf, 235555 bytes
- kremlin.ru_ru_2024.csv.gz, 88245548 bytes
- kremlin.ru_ru_2024.ods, 93176233 bytes
- kremlin.ru_ru_2024.pdf, 26097934 bytes
- kremlin.ru_ru_2024_posts_by_location.html, 4437202 bytes
- mid.ru_en_2024.csv.gz, 36739157 bytes
- mid.ru_en_2024.ods, 38111971 bytes
- mid.ru_en_2024.pdf, 48614 bytes
- mid.ru_ru_2024.csv.gz, 84517556 bytes
- mid.ru_ru_2024.ods, 87299527 bytes
- mid.ru_ru_2024.pdf, 142945 bytes
- transcript.duma.gov.ru_ru_2024.csv.gz, 231778177 bytes
- transcript.duma.gov.ru_ru_2024.ods, 231712405 bytes
- transcript.duma.gov.ru_ru_2024.pdf, 56631 bytes

Metadata

Title	Russian state institutions full-text datasets
Subtitle	–
Version	1.0
Creators	Comai
Uploaded by	Giorgio Comai
Main Category	Miscellaneous
Additional Categories	–
Institutional Affiliation	Osservatorio Balcani Caucaso Transeuropa / Centro per la Cooperazione Internazionale (OBCT/CCI). Trento, Italy.
Publication date	Oct. 29, 2024, 4:03 p.m.
Period of data creation/gathering	from Jan. 1, 2024 to Sept. 16, 2024
Date of data creation (text)	2024-09-16
Time period covered	from Jan. 11, 1994 to Dec. 31, 2023
Time period covered (text)	–
Sources of data	Websites of Russian state institutions
Archival Record IDs	–

Data types	text document
Data type (text)	–
Countries	Russia
Languages	Russian, English
Disciplines	Communication Studies, Political Science
Keywords	Government, Parliament, Russian Institutions, Russian President, Text Mining
Related datasets	–
Related datasets (text)	A previous version of one of the datasets included in this release was published on Discuss Data in 2021 - https://doi.org/10.48320/5EB1481E-AE89-45BF-9C88-03574910730A
Related publications	Giorgio Comai, Text-mining on-line sources from Russia openly, in: **Autocracy, Influence, War: Russian Propaganda Today**, edited by Paul Goode, 2025, forthcoming Giorgio Comai, Who said it first? Investigating the diffusion of the Kremlin's buzzwords before they entered the mainstream, in: ASN Europe 2023, 2023, https://tadadit.xyz/slides/2023-07-asn/ Giorgio Comai, Who said it first? 'The collective West' in Russia's nationalist media and official statements, tadadit.xyz, 2024, https://tadadit.xyz/posts/2023-03-who-says-it-first-nationalist-media-kremlin/
Related projects	Text as data and data in the text - tadadit.xyz
Institutional affiliation	Osservatorio Balcani Caucaso Transeuropa / Centro per la Cooperazione Internazionale (OBCT/CCI). Trento, Italy.
Funding	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.
Methods of data collection	Text Mining

Methods of data analysis	Word Frequency

Version History

Data License

Attribution: Open Data Commons Attribution License (ODC-By) v1.0

Data Collections with the Open Data Commons Attribution License (ODC-BY) can be freely assessed and used reused and redistributed if proper attribution is provided. Any public use of the database, or works produced from the database, must be attributed in the manner specified in the license. For any use or redistribution of the database, or works produced from it, the license of the database must be clear and any notices on the original database must be kept intact.For more information, please read the Open Data Commons Attribution License (ODC-BY) Full Legal Text.