Russian state institutions full-text datasets

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

2024 release

Giorgio Comai (OBCT/CCI)

2024-09-16

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

This is a stable release.

All datasets do not include materials 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.

Dataset availability

This release is available in full on discuss-data.net. Please use the version available on discuss-data.net, and cite your source accordingly.

A web version of the documentation is available on tadadit.xyz, at the following url:

https://tadadit.xyz/datasets/2024/russian_institutions_2024/

Dataset format

Datasets are published as compressed csv files (.csv.gz), as well as in .ods format.

In line with the `tif` standard, each corpus has a few standard columns, as well as additional metadata depending on availability:

- the first column is always `doc_id`, and is composed of the bare corpus name (based on base domain of the source and language) and a numeric id, separated by an underscore. For the Russian version of Kremlin's website, such id would look as follows: `kremlin.ru_ru_12345` (where 12345 is the numeric id associated with the given item). Numeric identifiers have no inherent meaning; their order may be substantially meaningless. If the original source website includes in the url a unique numeric id, this is maintained in the doc_id; otherwise an id is given at database creation. In such cases, the numbering may depend on the way the extraction process was implemented and may not reflect order of publication. This format allows to combine datasets, ensuring `doc_id` is still unique.
- the second column is always `text`: this is the main text included in the source page
- the third column is always `title`
- the fourth column is always `date`
- other time-related fields, such as `time` and `datetime`, may follow if available (time and date refer to the original publication timezone; in this release, this is always Moscow's time)
- additional columns include fields and metadata, depending on availability of contents: this may include substantive text contents (e.g. a separate `lede` or `description` field), categories, tags, location, author, additional identifiers, etc.
- finally, `url` is always the last column

`doc_id` and `url` are conceptually unique and always present. In all of these datasets, also `date` is always present. All other fields may be missing or empty for some of the items (e.g. there may be items with title, but no text, or vice-versa). See the documentation accompanying each dataset for more details.

List of full-text datasets included in this release

Archived versions of websites are marked with an *

institution	website	lang	corpus name
Russia's president	kremlin.ru	en	kremlin.ru_en_2024
Russia's president	kremlin.ru	ru	kremlin.ru_ru_2024
Russia's MFA	mid.ru	ru	mid.ru_ru_2024
Russia's MFA	mid.ru	en	mid.ru_en_2024
Russia's government	government.ru	ru	government.ru_ru_2024
Russia's government *	archive.government.ru	ru	archive.government.ru_ru_2024
Russia's prime minister *	archive.premier.gov.ru	ru	archive.premier.gov.ru_ru_2024
Russia's Duma	duma.gov.ru	ru	duma.gov.ru_ru_2024
Russia's Duma (transcripts)	transcript.duma.gov.ru	ru	transcript.duma.gov.ru_ru_2024

Summary statistics

corpus name	start date	end date	n_items
kremlin.ru_en_2024	1999-12-31	2023-12-31	33 165
kremlin.ru_ru_2024	1999-12-31	2023-12-31	45 538
mid.ru_ru_2024	2003-01-02	2023-12-31	56 203
mid.ru_en_2024	2003-01-04	2023-12-31	25 943
government.ru_ru_2024	2012-04-24	2023-12-30	17 135
archive.government.ru_ru_2024	2008-05-07	2013-05-21	7 103
archive.premier.gov.ru_ru_2024	2008-05-07	2012-05-07	3 323
duma.gov.ru_ru_2024	2006-04-05	2023-12-30	29 094
transcript.duma.gov.ru_ru_2024	1994-01-11	2023-12-15	6 032

N.B. `end date` reflects the date of the last item included in the dataset. In all cases when the end date falls in December 2023, the dataset effectively includes all items posted until 31 December 2023.

Explore full-text datasets in an interactive web interface

All of the corpora included in this release can be explored through an interactive web interface, available at the following url:

https://explore.tadadit.xyz/2024/ru_institutions_2024/

The interface is very basic, but it allows to explore the corpora based on word frequency, and shows keywords in context, always including links to the original source. It also allows to export subsets of the dataset (following the "export" link in the header at the top of the screen), based on date and pattern matching. It is possible, for example, to export only sentences where a given string is present. Enhancements to the interface are planned, and will be available at this link. Some functions, such as showing keywords in context, may take a few seconds, and up to a minute if there are many matches.

Should the online web interface become unavailable, or should users prefer to run it locally from their own computer, they can do so through the R package `castarter`.

The following lines of code should install all needed dependencies and open the interface in the browser, assuming the datasets are stored in the current working directory.

```
remotes::install_github("giocomai/castarter")

castarter::cas_explorer(
  corpus = readr::read_csv(file = "mid.ru_en_2024.csv.gz"),
  default_pattern = "ukrain"
)
```

It is possible to increase performance of the interface and to reduce the RAM memory needed to run it by first exporting the corpus in the `parquet` format.

```
remotes::install_github("giocomai/castarter")
corpus_df <- readr::read_csv(file = "mid.ru_en_2024.csv.gz")

castarter::cas_write_corpus(corpus = corpus_df,
                            partition = "year",
                            path = "mid.ru_en_2024")

castarter::cas_explorer(
  corpus = arrow::open_dataset(sources = "mid.ru_en_2024"),
  default_pattern = "ukrain"
)
```

As the full text datasets are distributed in standard formats, they can of course be processed with any compatible software.

License

Details about licensing are includeded along with the documentation of each corpus. The specifics vary slightly, but all of the source websites used to create this collection explicitly allowed for re-publication of contents a under a Creative Commons (CC-BY) license or similar. To the extent that it is possible, the datasets themselves are also distributed by its creator, Giorgio Comai, under the Open Data Commons Attribution license (ODC-BY).

Contact details

If you find data quality or other issues, or if you would like to receive more information or clarifications about this collection, please get in touch by email: g@giorgiocomai.eu

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.