Text as data & data in the text

Studying conflicts in post-Soviet spaces through structured analysis of textual contents available on-line

Giorgio Comai
Osservatorio Balcani e Caucaso Transeuropa
Centro per la Cooperazione Internazionale

September 2023

Table of contents

Credits and disclaimers

This project is carried out by OBCT/CCI with the support of the Italian Ministry of Foreign Affairs and International Cooperation under art. 23 bis, D.P.R. 18/1967. All opinions expressed within the scope of this project represent the opinion of their author and not those of the Ministry.

> “*Le posizioni contenute nel presente report sono espressione esclusivamente degli autori e non rappresentano necessariamente le posizioni del Ministero degli Affari Esteri e della Cooperazione Internazionale*”

Summary

This report oulines some of the key results of the project “Text as data & data in the text”, carried out by Giorgio Comai at OBCT/CCI with the support of the Italian MFA.

For a more complete experience including interactive data visualisations and more data, consider visiting the website hosting project outcomes:

https://tadadit.xyz/

Additional contents structurally not included in this report include:

- interactive interfaces enabling basic word frequency analysis:
  - https://explore.tadadit.xyz/ (the interface is expected to gain new features even after the end of the project; some of the sources included for analysis will be automatically updated)
- a series of textual datasets, some of which are available for download:
  - https://tadadit.xyz/datasets/
- slides of the paper presented at the ASN annual conference in Cluj-Napoca:
  - “Who said it first? Investigating the diffusion of the Kremlin’s buzzwords before they entered the mainstream”
  - https://tadadit.xyz/slides/2023-07-asn/
- the package `castarter` - Content Analysis Starter Toolkit for the R programming language - further developed in the course of this project:
  - source code: https://github.com/giocomai/castarter
  - documentation: https://castarter.tadadit.xyz/
- the source code used to generated the `tadadit.xyz` website is also publicly available:
  - https://github.com/giocomai/tadadit

Introducing ‘tadadit.xyz’

Originally published on tadadit.xyz on 12 March 2023

https://tadadit.xyz/posts/2023-03-tadadit/

Studying conflicts in the post-Soviet space through structured analysis of textual contents available on-line: this is the objective of “Tadadit.xyz”, which is just an acronym for “Text as data & data in the text”.

At the most basic, this is just a small contribution to the increasing amount of scholarship that relies on some form of content analysis, a trend that for better or worse is likely to increase in this space due to limited or constrained access to Russia and contested territories.

I should add from the outset that I am rather sceptic about research that relies primarily on **content analysis**. But I still feel that **structured analysis** of on-line contents has a significant role to play, including in contexts such as area studies and peace and conflict research.

This is obviously part of a big debate, so, in brief, here are the contributions I wish to make.

My starting point is that, if not for the technical hurdles, a lot more scholars working on these issue would rely on structured analysis of on-line contents.

I am not only referring to things such as word frequency analysis or more advanced techniques, but more broadly, at the idea that on-line sources can be analysed in a structured manner, rather than through serendipitous encounters via search engines or social media. Structured approaches for parsing selected on-line sources can be useful for finding specific data points, as well as information about places, institutions, individuals, amounts of resources, and governance practices. These is what I mean with “**data in the text**”, beyond the more established and generally more quantitative “**text as data**”.

Mostly, the final goal of a research endeavour does not lie in either “text as data” or “data in the text” are. These are just useful means, almost invariably complementary to other methods, for finding more evidence, developing better research questions, testing and potentially dismissing lines of inquiry, and, ultimately, giving better answers to meaningful research questions.

What to expect on *tadadit.xyz*

Based on this, here is how I hope this initiative will offer a small contribution towards mainstreaming some of these approaches.

- **Substantive articles demonstrating this approach** with concrete examples of how it can be useful, including both more classic pieces based on word frequency, as well as others showing how relevant data can usefully be found through structured approaches to text analysis
  - these will all be related to **Russia's invasion of Ukraine**, Russia-controlled territories in Ukraine, and other conflicts and contested regions in post-Soviet spaces
  - besides the classic final results, these posts will give access to **procedural steps** and will show, for example, **keywords in context**, giving the reader more opportunities to make up their own minds about the arguments proposed
  - full access to **interactive interfaces** for further exploring the datasets, empowering the reader, and giving them the possibility to test alternative hypotheses and challenge the interpretation proposed in the articles themselves (*interactive interfaces not yet publicly available, but coming soon*)
- **Increased availability to pre-processed textual datasets** relevant to scholars working on conflicts in post-Soviet spaces
  - pre-processed textual datasets, when contents are available unencumbered by copyright restrictions
  - access to pre-processed datasets through interactive interfaces, when full sharing of the data is not possible
- **Easier access to the tools needed to create similar datasets** and keep them up-to-date based on the specific interests of a researcher. Scholars working on these issues will often want a dataset based on items posted by a small local institution in a small town, or some other organisation: if this approach is to become more common, then it should be relatively easy to create new textual datasets and the keep them up to date. We need better open tools and more tutorials that explain how to use these tools in contexts such as the ones we are interested in.
  - with this in mind, I am working on a new iteration of an open source package I've been working on for a few years, `castarter` - Content Analysis Starter Toolkit for the R programming language

  - at this stage, it streamlines the text mining workflows for researchers that have some familiarity with the R programming language and how the internet works, but does little for the rest
  - the short term goal is to improve the package and make it accessible for people who have very limited programming skills and limited understanding of how the internet works, both by further simplifying the workflows and by producing a series of beginner-level tutorials
  - the long term goal it to make most or all of the process available through a web interface that does not require any programming skill, and make it much easier to collect textual datasets, explore them, and, fundamentally, share them in a readily usable format with the wider public or selected colleagues

Working in the open

Finally, a key component of this whole endeavour is **working in the open**. If you explore this website, you will see that many pages, including reviews of literature, are often in draft format, sometimes just early drafts (and are marked as such). **This is a feature, not a bug**. Significant efforts will be dedicated towards radical transparency throughout all phases of this endeavour, favouring open debates, earlier sharing of knowledge, positive feedback loops, and the advancement of open science practices in academia.

Radical transparency through all stages of the research process can be helpful in sharing ideas and methods, receiving feedback, and reducing the impact of some of the great scourges of academia: almost-finished papers that remain in some half-forgotten folder, or advanced research that is not available *for years* until publication, with the possible exception of a few individuals that attended a conference presentation.

I will consider further options for making it easier to find out about new posts as they are published (e.g. email notifications), as well as for giving feedback (I'm not a fan of blog comments, but I may consider them, or some other alternative system).

In the meantime, you can stay up-to-date by following this website's RSS feed or the project's account in the Fediverse, or consider opening an issue on the repository where this website is hosted (it is uncommon in the humanities to rely on this approach, but it's really a rather useful format for debating specific issues).

What’s needed in a content analysis toolkit for starters?

Originally published on tadadit.xyz on 7 November 2022

https://tadadit.xyz/posts/2022-10-content-analysis-toolkit-for-starters/

Context

Manuals and tutorials presenting content analysis techniques are often based on the assumption that a structured textual dataset is available to the researcher. This is convenient and understandable, as it would not make sense to start each time with technicalities about data collection methods.

And yet, very often relevant textual datasets need to be built by the researcher. For those working on current affairs or recent events, the starting point is often texts published on-line by institutions, media, NGOs, activists, etc.

As a scholar working at the intersection of area studies with peace and conflict research, the textual datasets that I felt were most useful for my research were based on contents published by institutions and media based in the area under study: not only big national media, such as e.g. national TV stations in Russia, but rather local news agencies and institutions, such as those based in contested territories. If there are various way to access data from larger media in a structured format via paid services such as LexisNexis, there is not much ready-made for the scholar working on local contexts in these areas.

More often than not, the solution to these problems comes from text mining/scraping relevant on-line sources. There is plenty of tutorials on how to do these things on the web, but in most cases I feel there is not enough attention dedicated to the workflow, and to the menial tasks that are the basis of text mining, such as managing files consistently, ensure consistency and minimise repeated processes across multiple sessions, deal with archiving and backups, and keep track of what is where, as well as the time when each content was retrieved.

When working on contemporary affairs, it is also important to be able to update a textual dataset as easily as possible.

When doing text mining as a scholar, it is also important to be able to tell how a given piece of information was found, and when a given page was visited.

In this post, I will describe how I deal with these issues in the package for the R programming language I have been developing, `castarter` - Content analysis starter toolkit for R. This post, however, does not include any code, and outlines only key concepts and steps of the workflow, so it should be relevant also to readers who are not interested in using R or my package.

Basic setup

When processing the archive of a website, there are typically two types of pages that are most relevant:

- **index pages**. These are pages that usually include some form of list of the pages with actual contents we are interested in. They are often in a format such as:

  - https://example.com/all_posts?page=1
  - https://example.com/all_posts?page=2
  - etc.

    Index pages may also not be visible to the user, and may include:

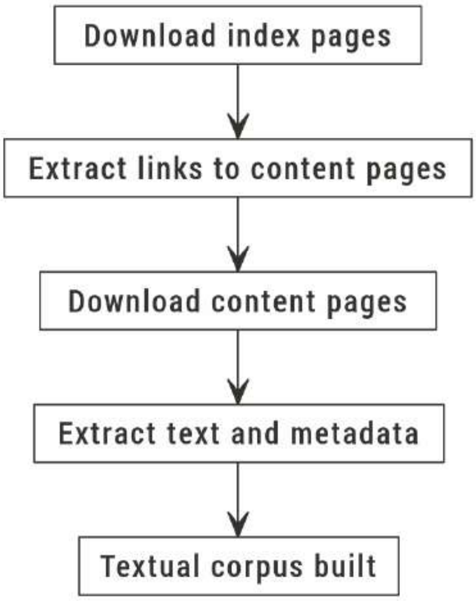
  - Sitemaps: typically in https://example.com/sitemap.xml or as defined in https://example.com/robots.txt
  - RSS feed: typically useful only to retrieve recent posts, and found e.g. in https://example.com/feed, https://example.com/feed.rss, or in some other location often shown in the header of the website

- **content pages**. These are pages that include the actual content we are interested in. These have urls such as:

  - https://example.com/node/12345 or
  - https://example.com/post/this-is-a-post

Conceptually, the contents of the **index pages** are expected to change constantly as new posts are published, so in case of an update, they will need to be downloaded again. Conceptually, **content pages** are expected to remain unchanged, at least in their core parts.

At the most basic, this is what should happen when extracting textual contents from the archive of a website.

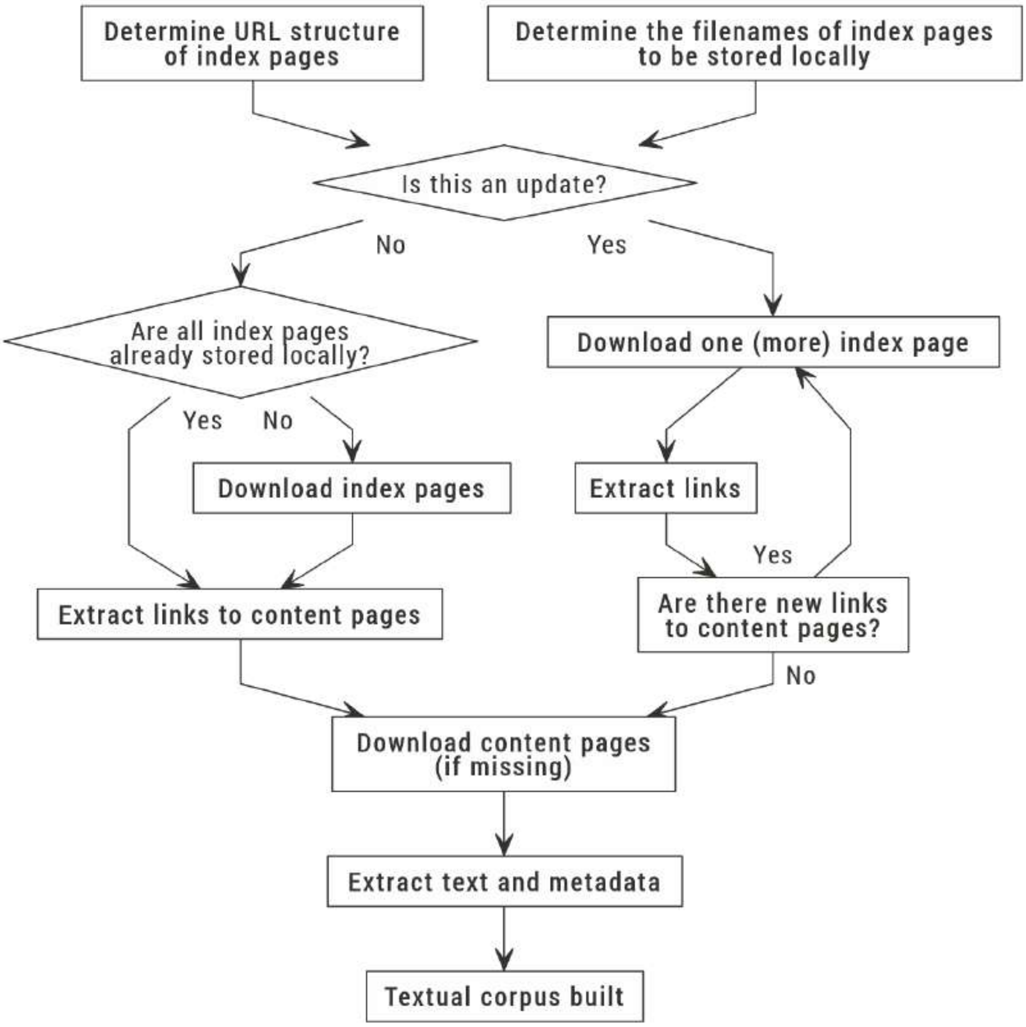


In practice, the situation is often a bit more complicated.

Firstly, there are a number of conventions to consider, such as in which folders files are to be stored? how should downloaded pages be named?

Secondly, as all files may not be downloaded in the same sessions, some checks should be in place for making sure **content pages** are downloaded only once, unless there is specific reason to download them again.

Thirdly, and somewhat more complicated, how should updates be managed? A slightly more detailed diagram would look as follows:


Determine URL structure of index pages
Determine the filenames of index pages to be stored locally
Is this an update?
No
Yes
Are all index pages already stored locally?
Download one (more) index page
Yes
No
Download index pages
Extract links
Yes
Extract links to content pages
Are there new links to content pages?
No
Download content pages (if missing)
Extract text and metadata
Textual corpus built

Memory, storage, backup, and other issues

As each of these steps requires a series of operations, the process can easily become tedious. Other menial issues likely to cause headaches at some point include:

- disk space issues when storing data - hundreds of thousands of pages can easily take many gigabytes
- backup and data management issues - you may be accustomed to keep your files synced with a cloud service such as Dropbox, Google Drive, or Nextcloud,

but this is mostly not a good idea in this scenario, since their client apps become quickly inefficient and painfully slow down start-up and syncing times when hundreds of thousands of files or more are involved (this is more an issue of file number, than total size)
- memory issues when processing the corpus - processing big datasets can crash a session due to missing ram

All of these things are substantively irrelevant, of course, but effectively end up taking up a disproportionate amount of time for the researcher.

These are all things that a content analysis starter toolkit should take care of, suggesting and implementing practical solutions by default, and letting the researcher dedicate their time to substantive issues.

Indeed, all of the issues outlined above emerge before any analysis of the data can even take place. In my understanding, these are however the issues that a *starter* toolkit should mostly focus on: once a nicely structured dataset has been gathered, there are plenty of software solutions, tutorials, and approaches that can be used. Nothing more than the most basic data exploration should be enabled by a starter toolkit.

Features needed in a useful content analysis starter toolkit

A starter toolkit should make it relatively straightforward to:

- get all relevant links from a website
- retrieve and store all files in a consistent manner
- consistently extract key textual contents and metadata from each page
- conduct some basic quality and sanity checks
- keep record of the process in both human and machine readable formats
- store the original files to a convenient backup location
- update the dataset with minimal efforts
- manage data from multiple websites or projects consistently
- make it easy to store and process data also on low-end computers
- conduct the most basic forms of analysis, such as word frequency
- export resulting datasets in standard formats, to enable more advanced techniques of analysis
- give access to the dataset to others

Finally, it should make all of the above while being robust: it should be possible to unplug the computer at any stage of the process (download, extraction, etc.) without losing any data, and proceed from where it was stopped without issues.

All of these features are already available with the R package `castarter`, or will be in the course of 2023. As this is a work in progress, each of these features will improve with time, either lowering the technical competence needed to use the package or providing more flexibility to deal with a wider set of use cases.

The long-term goal for `castarter` is to enable all of the above directly through a web interface, without requiring any familiarity with programming languages nor, at least in most cases, familiarity with how html pages are built.

Who said it first? ‘The collective West’ in Russia’s nationalist media and official statements

Originally published on tadadit.xyz on 14 March 2023

https://tadadit.xyz/posts/2023-03-who-says-it-first-nationalist-media-kremlin/

Context

Russia’s full scale invasion of Ukraine in 2022 has been accompanied by the emergence and diffusion in Russia’s public discourse of new concepts and terms needed to frame in a new light what it is that Russia is really fighting against.

In the case of Ukraine, this trend has been noticeable in mainstream Russian media since 2014 and has grown in the following years: Ukraine is ruled by a “junta”; it is controlled by nazi; its military forces are “formations of nationalists”, etc.

Russia is however also fighting a war against a larger and somewhat more sinister enemy. Its name is, increasingly, “**the collective West**”, an expression that allows to reframe “the enemy” as a single hostile entity.

There is not a single reference to “the collective West” in official declarations by the Russian president before 2021, but it has since become relatively common especially in “big speeches”, with dozens of mentions in recent months.

But where does this expression come from?

In this post, I will:

- confirm that, indeed, this is a relatively new expression
- find that before 2021 it was very rarely used in news reporting on mainstream television, and is still used only occasionally in plain news reporting
- consider the possibility that it was used earlier in fringe nationalist media:

    - find that it was sparsely used on *Tsargrad TV* before 2021, but that it became quite common starting with 2022
    - find that it was consistently used in articles published on nationalist weekly newspaper *Zavtra* starting with 2015

- observe that the expression “collective West” has been used by fringe nationalist analysts for some years, and has entered official rhetoric only recently. It has since become commonplace on nationalist tv channel *Tsargrad*, but is still used only occasionally on mainstream tv.
- notice that as “collective West” enters more common use in the presidential rhetoric, previously common expressions such as “our Western partners” - supposedly neutral, but in context often deprived of positive connotations - are used less frequently and only with qualifiers (e.g. “our so-called Western partners”) marking once more the obvious change in attitude.

Further analysis is needed to determine the origins of “collective West” as an expression and as a framing, as well as the role of nationalist media and official discourse in popularising it.

An earlier publication analysing the concept of “collective West” (Chimiris (2022)) posited that before reaching the Kremlin, the expression has occasionally been used for some time in the Ministry of Foreign Affairs, particularly by its spokesperson Maria Zakharova. Yet, all of these uses come after 2015, when the expression started to be consistently used in nationalist newspaper *Zavtra*, suggesting that fringe nationalist publications may indeed have been the breeding ground where the expression established itself before going mainstream. If confirmed, this would be a new example of how concepts and framings once popular in relatively marginal nationalist circles with limited or no access to major national media are now being used and promoted directly by the Kremlin.

For a rather similar excercise focused on the concept of “Russophobia”, you can see my 2021 post: “Russophobia in Russian official statements and media. A word frequency analysis”. Other posts on different concepts may follow.

> ⚠ Warning
>
> The datasets used in this post have last been updated at different points in time between February and March 2023. Including recent data when available has been preferred over consistency. The exact cutoff date for each dataset is always shown in the subtitle of the graph.
> Graph columns that include partial data about a given time period are shown in a lighter shade, to highlight their incompleteness.
> Graphs show absolute frequency rather than relative frequency, even if the rate of publications is not perfectly stable on some of the sources included

in this analysis. Given that the expression at the core of this article has never or almost never been used in earlier periods, this has limited substantive impact, and the absolute numbers remain intuitively easier to understand.
The number of references has been grouped either by year or by quarter, as including shorter time periods would ultimately highlight only the buzz around a single speech, rather than the overall trend. In graphs showing data grouped by quarters "2022.1" stands for the first quarter of 2022, i.e. January to March.
Full textual datasets are shared along with this post, when allowed by the license. A fullly-documented version of these dataset will be published soon. A preliminary version of the datasets is already available for download.

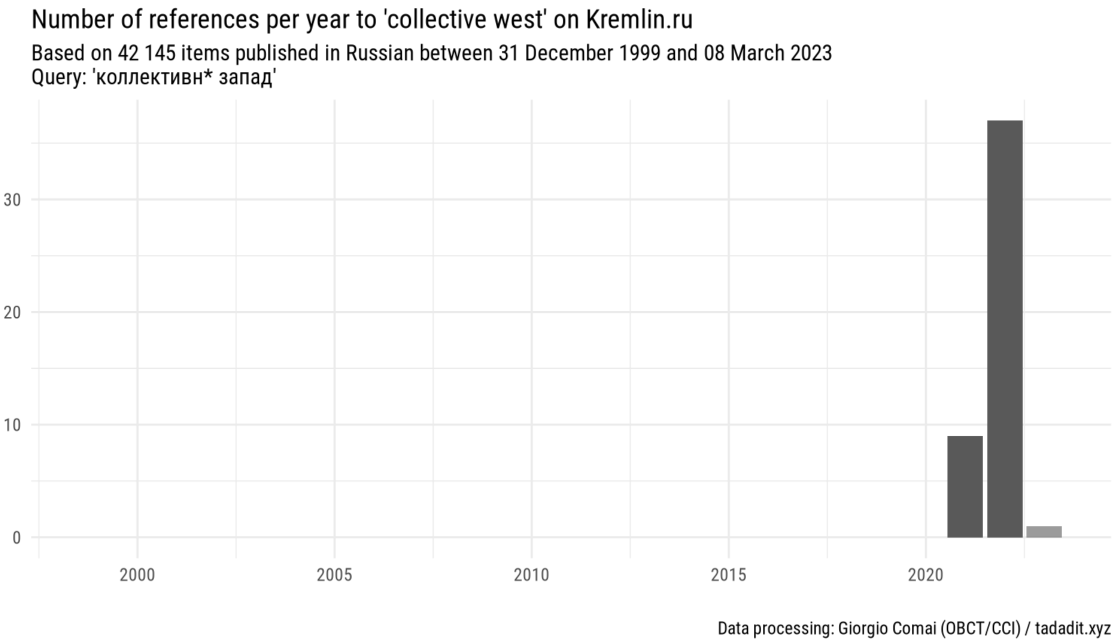
- **zavtra.ru_ru** - Russian weekly *Zavtra* (in Russian)
- **kremlin.ru_ru** - Statements from the official Kremlin website (in Russian)
- **kremlin.ru_en** - Statements from the official Kremlin website (in English)

An earlier version of this textual dataset is available online (Comai 2021).
Kremlin.ru and Zavtra.ru both publish their contents with a Creative Commons license.

The Kremlin

About the dataset

This section is based on both the Russian and English language version of the official website of the Kremlin. When the same content has been published twice on the website (e.g. as a press release and as a transcript) only one of the versions has been kept.

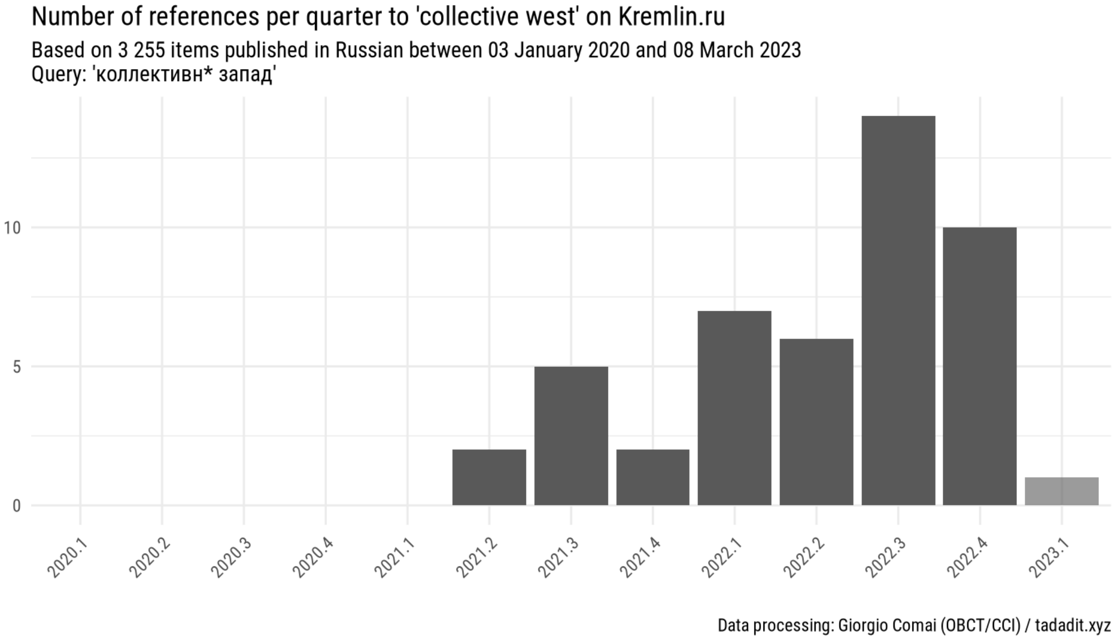


Number of references per year to 'collective west' on Kremlin.ru

Based on 42 145 items published in Russian between 31 December 1999 and 08 March 2023

Query: 'коллективн* запад'

Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz



Number of references per quarter to 'collective west' on Kremlin.ru

Based on 3 255 items published in Russian between 03 January 2020 and 08 March 2023

Query: 'коллективн* запад'

Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

Before 2021, the expression “collective West” has never been used, not a single time, in any of the tens of thousands of statements published on Kremlin’s website since 2000. Since then, the expression has featured quite frequently, in particular in occasion of “big speeches”. See below all relevant mentions, first in Russian, then in English.

On 21 April 2021, for the first time, Vladimir Putin used the expression “collective West”, or, more precisely “the so-called collective West”. But “so-called” by whom?

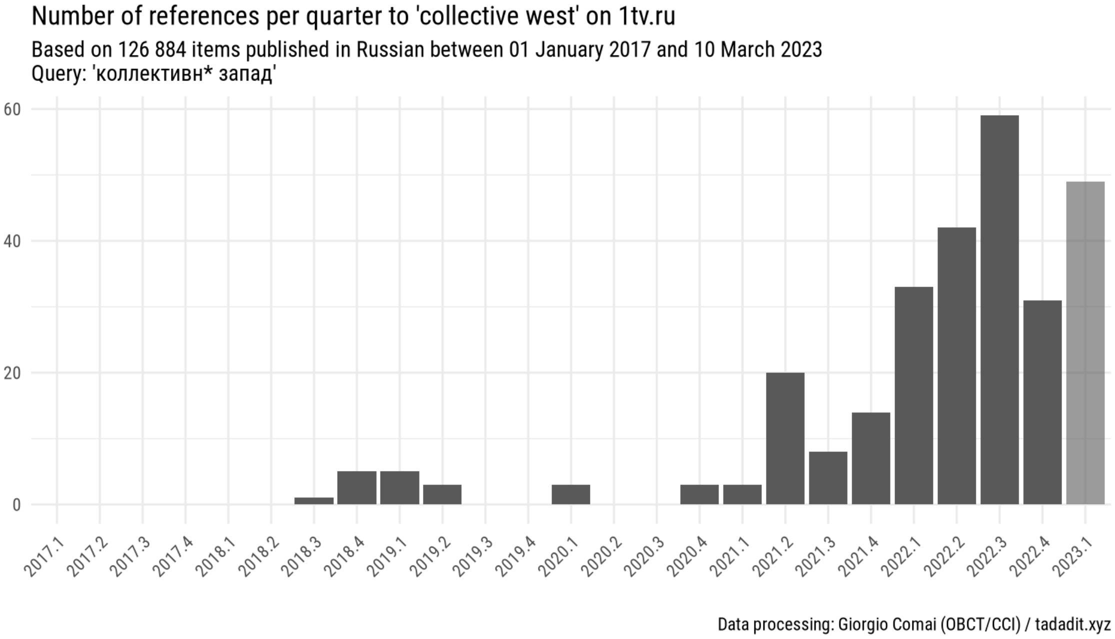
And how did this expression enter his vocabulary?

Pervy Kanal

It appears that the source of inspiration was not Russia's *First Channel*, one of the main TV channels in Russia (*Pervy Kanal* - link to Wikipedia page of *Pervy Kanal* for more context).

Indeed, the expression was barely used in mainstream news reporting before Putin's speech in April 2021.

The following graphs are based only on the transcripts of regular news segments available online. It should be noted that these do not include transcripts of the talk shows, where inflammatory or politically loaded statements are more prevalent.



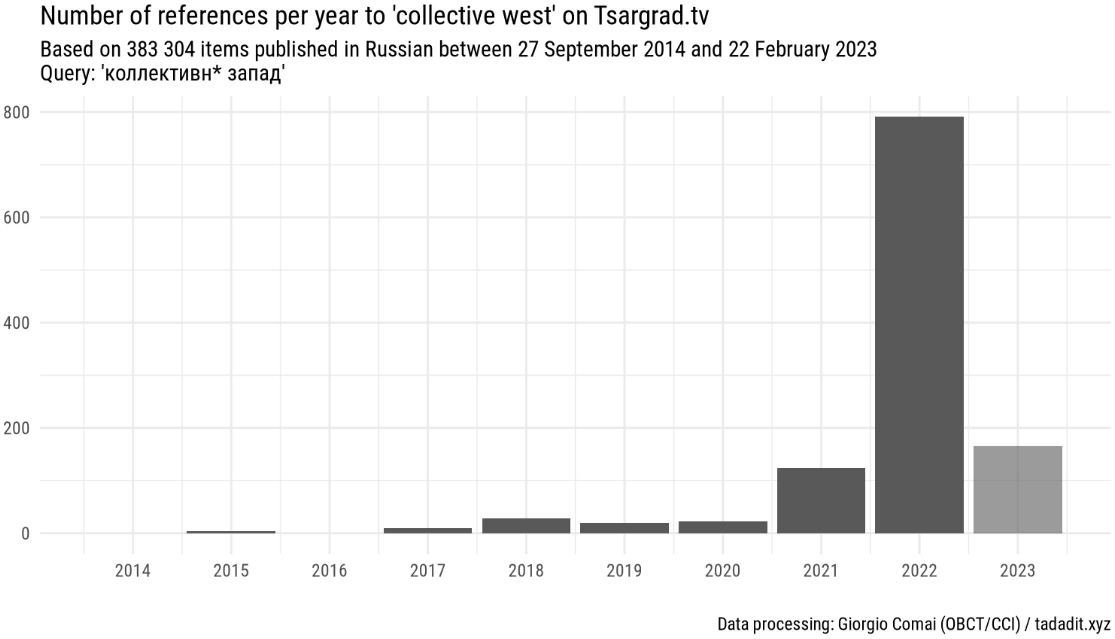
It appears that bar a few occasional earlier mentions, the expression "collective West" has effectively entered into use only starting with 2021, and even more in 2022. Given the very high number of publications (about `r per_day_1tv` items are published each day on 1tv.ru), this shows still a relatively low frequency, with the expression used on average once every couple of days in recent months. Besides, most of the early mentions are in fact repeated quotes of Putin's speeches in multiple news segments.

In this case, it appears that things go mostly as expected in the Russian context, with the media adapting to the terminology and narratives established by the Kremlin, and actually not adopting this specific expression until recently.

Nationalist media

Tsargrad

Tsargrad TV (link to Wikipedia page) is a Russian nationalist television channel launched in 2015 by Konstantin Malofeev.

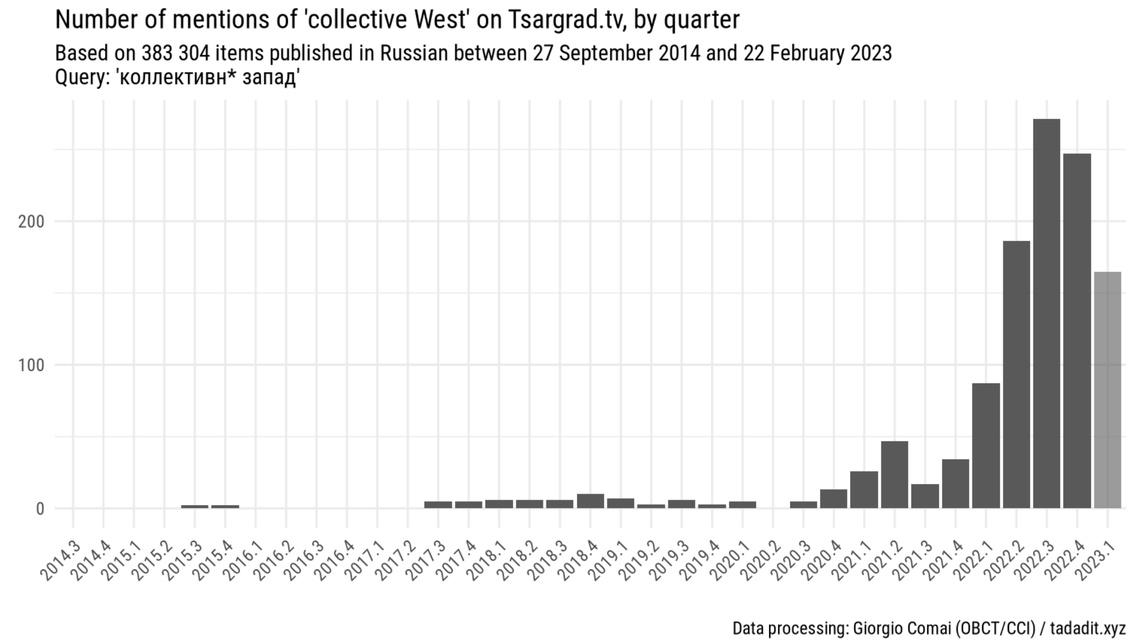


Number of references per year to 'collective west' on Tsargrad.tv

Based on 383 304 items published in Russian between 27 September 2014 and 22 February 2023
Query: 'коллективн* запад'

Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

Before 2021, references to the “collective West” are very sporadic.



Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

In particular considering the very high number of publications (about `r per_day` items are published each day on Tsargrad.tv), mentions before 2021 are very sporadic. It is only with 2022, that references to the "collective West" become commonplace, after Vladimir Putin started to use the term and in particular after the launch of Russia's invasion of Ukraine.

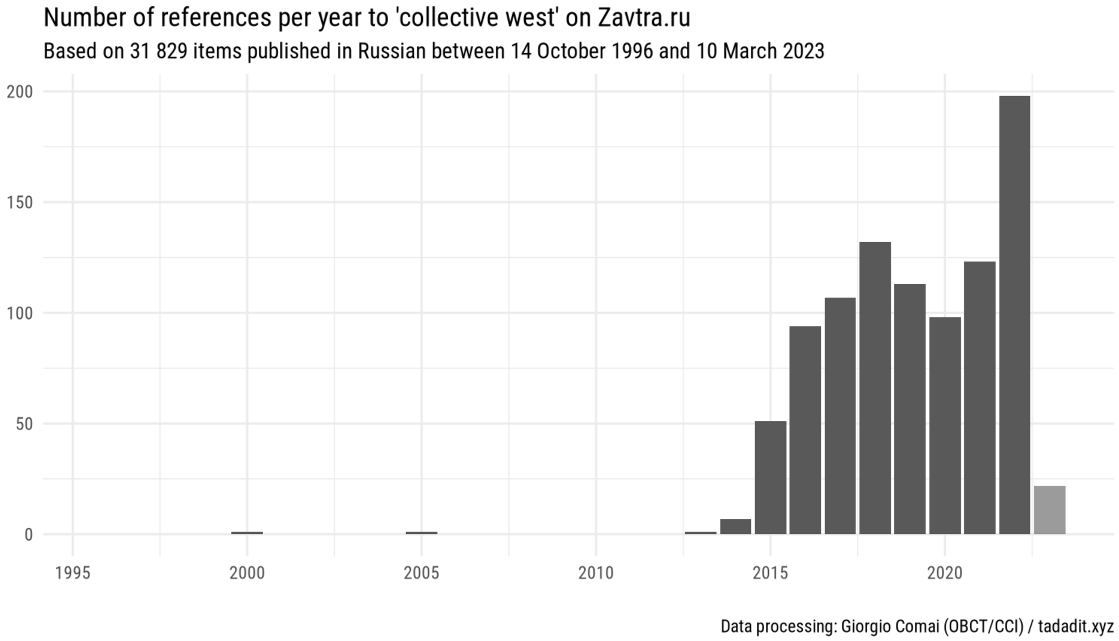
Zavtra

Zavtra has been published as a weekly printed broadsheet newspaper since the 1993. *Zavtra*'s ideology, outlined in a dedicated page on its official website, is a version of Russian nationalism that celebrates Russia's Tsarist past, its Stalinist glory, and Russia's inevitable future rise to victory.

In the same text, they also boast to have created trends that have entered the mainstream:

> "Over the years, we at *Zavtra* have created several ideologies, several powerful trends that have entered and continue to enter the public consciousness."

Could the concept of "collective West" be one such instance?

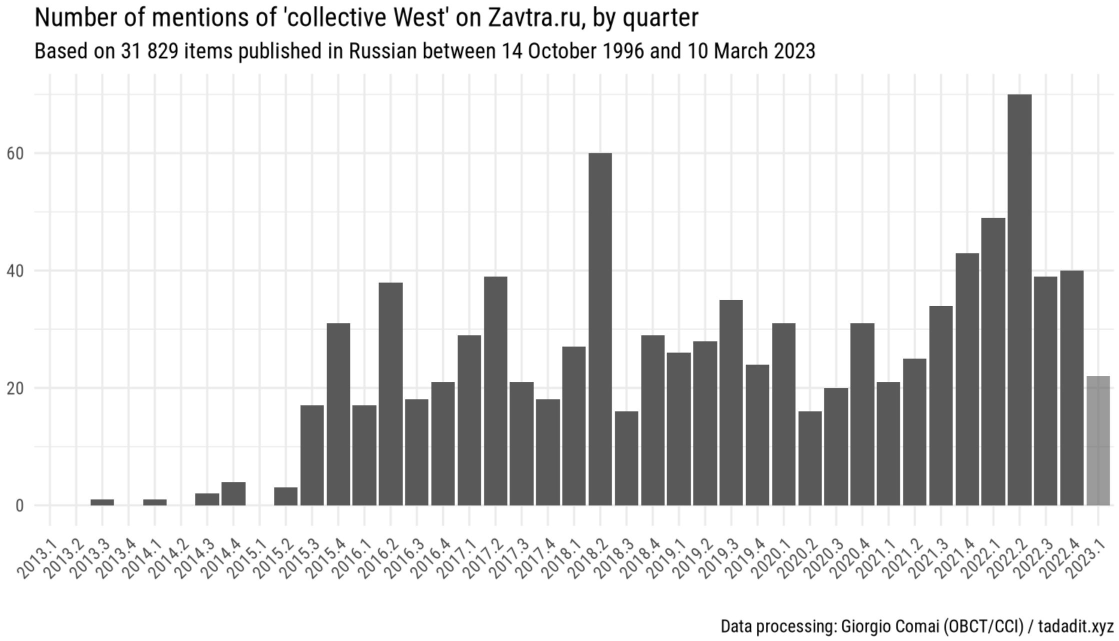


Number of references per year to 'collective west' on Zavtra.ru

Based on 31 829 items published in Russian between 14 October 1996 and 10 March 2023

Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

There are two early mentions in 2000 and 2005, but the expressions starts to feature routinely only starting with the summer of 2015, as appears more clearly from the following graph.



Number of mentions of 'collective West' on Zavtra.ru, by quarter

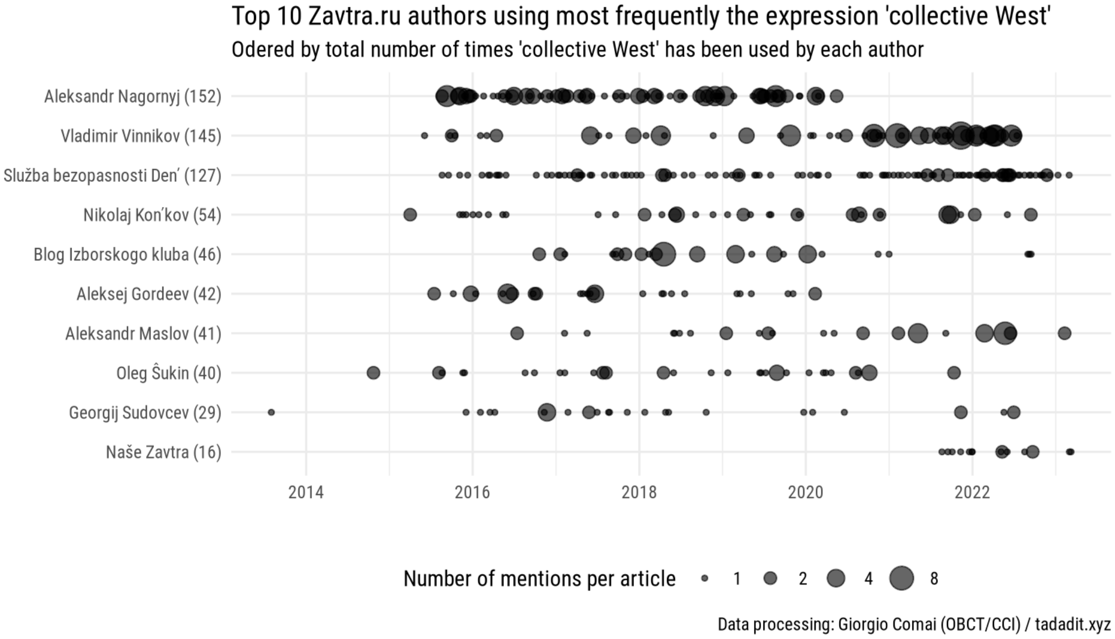
Based on 31 829 items published in Russian between 14 October 1996 and 10 March 2023

Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

Even if it is only starting with 2015 that the expression becomes commonplace, the first mention appearing in 2014 summarises in a single sentence many of the themes that would feature prominently in the official rhetoric in later years.

> "On 19-21 February in Kyiv took place a neo-nazi-Banderite coup, inspired by the collective West, and, first of all, the United States." — "After the coup)", Andrey Fursov, 13 March 2014

Even if the expression is used by a wide range of authors, almost half of the mentions can be tracked to three authors (the third of them is actually an editorial collective). Their names are listed below:



Even if there is not a distinct "manifesto" piece that explicitly introduces the term (or if there is, I haven't yet found it), it seems clear that is is starting with the late summer of 2015 that the expression really becomes widespread on *Zavtra*. More than anybody else, it appears that it is Zavtra's deputy director Alexander Nagorny's insistent references to "collective West" that really set the trend. As the deputy director of *Zavtra* between 1997 and his death in 2020, Nagorny has been a regular contributor to the newspaper for years, and yet, only starting with 2015 he started to use the expression routinely.

Without further qualitative analysis, it is difficult to gauge what is really the starting point that made the expression go mainstream, but this initial quantitative analysis offers useful hints.

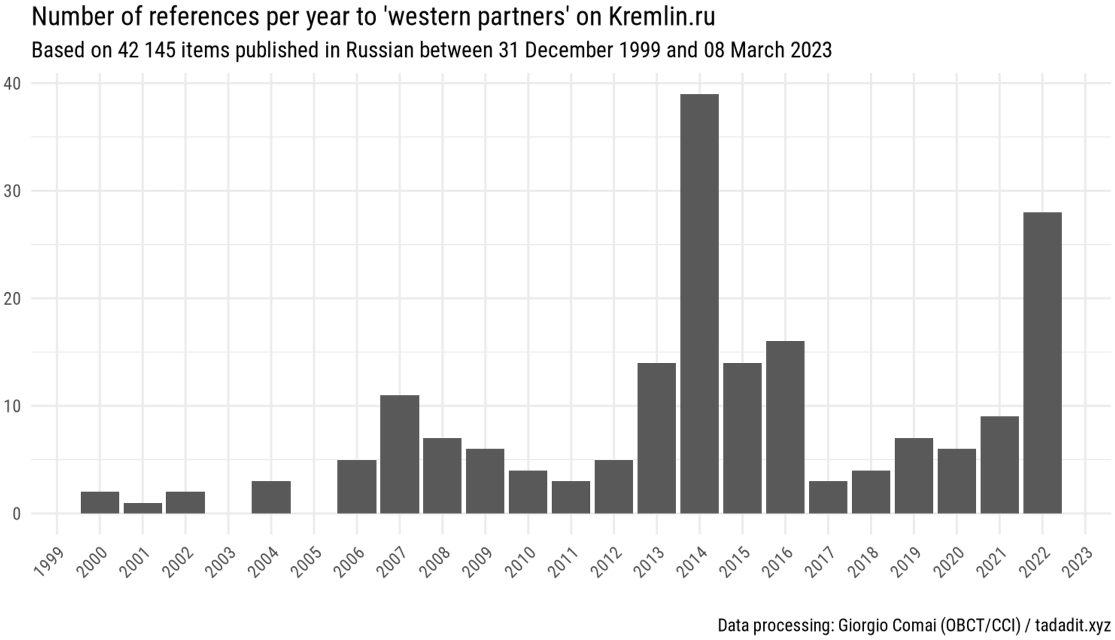
What about “our Western partners”?

Before the “collective West” entered official discourse, how did the Kremlin refer to these countries?

Other phrases may have been in use, but “Western partners” has long been a standard expression, even if seldom used in the “good times” of the early 2000s. Indeed, it is in the “bad years” of 2014 and 2022 that the expression has been used most frequently by Russia’s president.

As you look at the these graphs below and scroll down the use of the expression in context, you may want to pay attention in particular to recent months:

- there have been zero references to “Western partners” in the first two months of 2023, which are part of the dataset
- recent mentions in 2022 show how the expression “Western partners” is becoming explicitly problematised, and cannot be used seriously any more. We see more references to our “so-called Western partners” and even, on 7 December 2022, “our – we should put it in quotation marks – ‘Western partners.’ ”

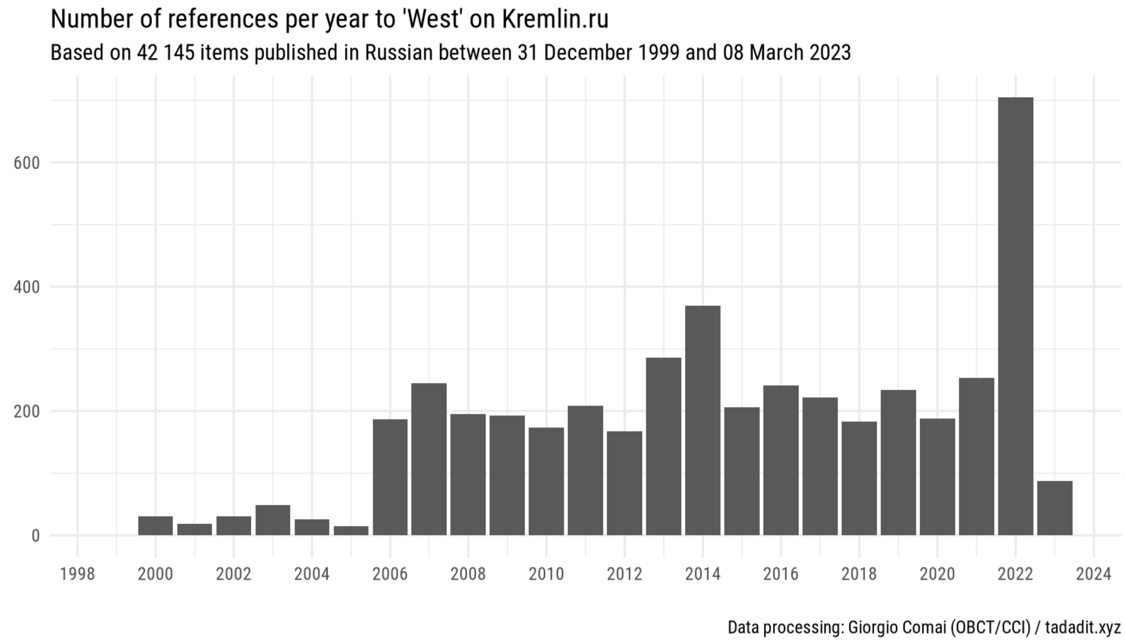

Number of references per year to 'western partners' on Kremlin.ru

Based on 42 145 items published in Russian between 31 December 1999 and 08 March 2023

Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

As you can see from the keywords in context, this catches all the Russian forms of both “Western partners” and “partners in the West”.

Finally, I should add that there is some indication that the expression “West” itself has become more common. The following graph is based on a simple word count of references to “West”, and does not differentiate between its use in generic expressions and the geo-politicised ones. Yet, the booming number of references in 2022 gives a clear hint that a new phase has started.



Traditional, conservative, Christian, distinct… how supposedly old values emerged in the official and media discourse in Russia since 2012

Originally published on tadadit.xyz on 16 May 2023

https://tadadit.xyz/posts/2023-05-conservative-traditional-values-russia/

ℹ Note

N.B. This post is best viewed online, as it includes explorable datasets and multiple versions of graphs

Context

"Traditional values" have emerged in Russia's public discourse with new vigour starting with 2011-2012 (see e.g. Quenoy e Dubrovskiy 2018, 94.) and have soon become prevalent in Russian official and public discourse. With Putin's return to the presidency in 2012, conservatism has become "official state posture" (Laruelle 2016, 287) in Russia. Increased reliance on conservative ideology may have been conceived largely instrumentally (Rodkiewicz e Rogoża 2015), as a discursive turn useful to reconsolidate after the wave of protests of 2011-2012 (Sharafutdinova 2014), but in the following years it has solidified as a central component of the Kremlin's narrative about Russia itself, as well as its contrapposition with the West. These values have been called "traditional", "conservative", "Christian" - all expressions which would suggest they may not be specifically Russian. And yet, they were also unique, something that made Russia a "distinct civilisation" (Tsygankov 2016).

Well before Russia's fulls scale invasion of Ukraine in 2022, "spiritual-moral values" became an issue of national security in Russia, and a form of defence against foreign threats and influence (Østbø 2017). Indeed, “traditional Russian spiritual-moral values” (“традиционные российские духовно-нравственные ценности”) explicitly and repeatedly featured in the national security strategy of the Russian Federation introduced in 2015 (and then again in the more recent strategy introduced in 2021), in contrast to previous versions of the same document President of the Russian Federation (2021). “Traditional spiritual-moral values” are mentioned 10 times in the Foreign Policy Concept of the Russian Federation approved on 31 March 2023 (President of the Russian Federation 2023).

In 2022, they arguably became an important part of the rhetoric repertoir used to legitimise Russia’s military actions, and have become even more clearly “a focal point in the domestic crackdown on the liberal opposition, and in the standoff between Russia and the West in international affairs” (Østbø 2017, 202).

In this post, I will explore references to “values” in Russian official discourse and media in Russia. More specifically, I will:

- identify which qualifiers have most commonly been added to “values” in the last two decades
- check the relative frequency of various expressions used to refer to these values (e.g. traditional VS conservative VS Christian), and how they have evolved
- check if indeed references to “traditional values” (or similar expressions) has increased in 2012 compared to previous years
- check if there is a change in connection to Russia’s invasion of Ukraine and in more recent months.

As will become apparent, some of the data sources used in this document have some features that make it more difficut to conduct consistent longitudinal analyses. Such limitations will be outlined clearly, as they are likely to have substantive impact on the results; they include issues of data availability, change in the volume of text available, and the same expressions used in different contexts (e.g. Western values were positive in the early 2000s, but negative in the early 2020s). Indeed, in this respect this post serves also as an exploration of the many challenges that come with longitudinal data analysis of textual corprora generated from online sources.

The unit of analysis

Throughout this post, the main unit of analysis will be the *sentence*. As will become soon apparent, qualifiers that clarify which “values” are being talked about often do not come immediately before or after the word “values” itself, but are part of longer expressions. Also, quite often more than qualifier is included (e.g. “spiritual-moral values”); some analyses will consider all qualifiers found within the same sentence, while others, for comparison, will just consider the qualifier that immediately precedes the word “values”. In such cases, the unit of analysis will only be mentions of “values” preceded by a relevant qualifier (full list below).

How frequent are references to “values”?

Let’s start with the English-language version of the Kremlin’s website, which should give some context, and then switch to a wider range of sources which are available only in Russian.

English-language version of Kremlin.ru has 32 102 items in total published between 31 December 1999 and 27 February 2023. If we count sentences, rather than items published, we have 821 552 sentences in total in our corpus. Among them, 2 120 make reference to “values”.[1]

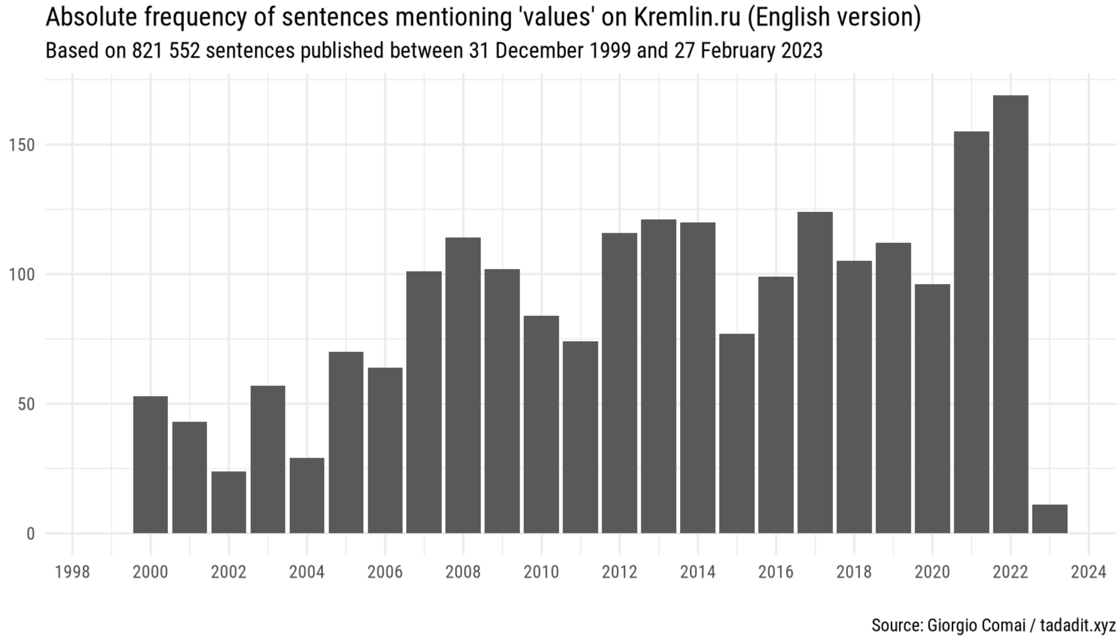
Let’s look at their distribution.

While there is an increase in the absolute number of references to “values”, this is overall balanced out once we account for the increasing number of text published by the Kremlin each year.

See the data in both relative and absolute terms.

[1] This applies only to the online news archive of 1tv.ru, which does not include full transcripts of all broadcasts. Even if Prigozhin’s role may have emerged in debates during talk shows, his complete absence from standard news reporting remains telling.

Relative frequency

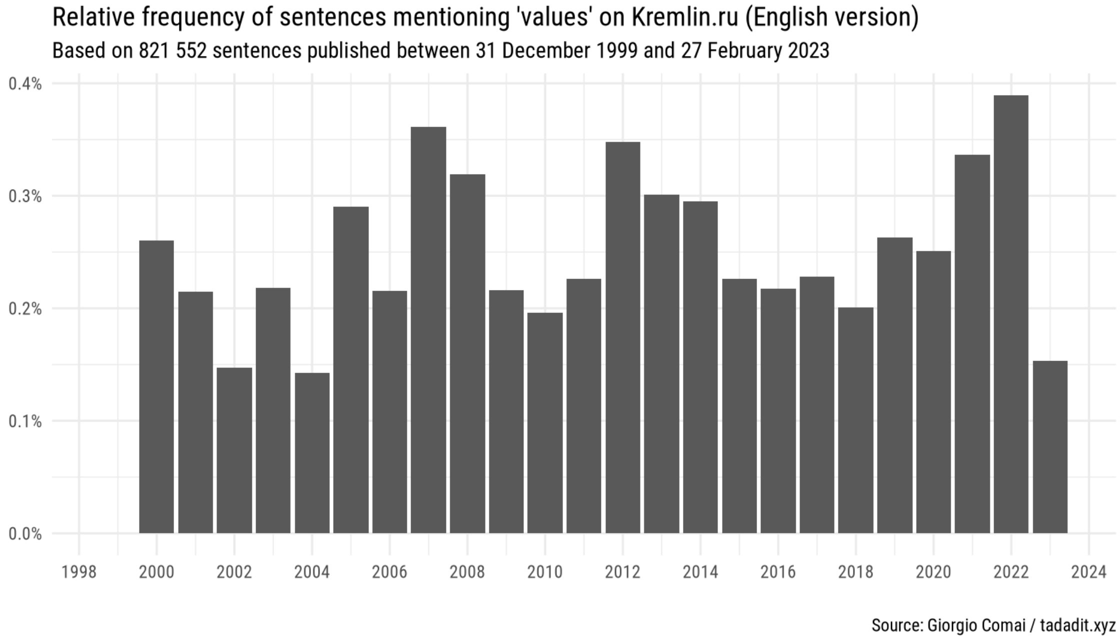


Absolute frequency of sentences mentioning 'values' on Kremlin.ru (English version)

Based on 821 552 sentences published between 31 December 1999 and 27 February 2023

Source: Giorgio Comai / tadadit.xyz

Absolute frequency



Relative frequency of sentences mentioning 'values' on Kremlin.ru (English version)

Based on 821 552 sentences published between 31 December 1999 and 27 February 2023

Source: Giorgio Comai / tadadit.xyz

Tabular data

An argument could possibly be made that the increased number of publications on the Kremlin's website reflects the fact that it has become more common to cover minor events, but that the overall number of "big speeches" where one would talk about values has effectively remained constant, and hence the absolute figures are what really matters. But even looking at absolute numbers, we can see that the "moral turn" that according to the literature has started in 2012 is not really reflected in a sudden spike of references to "values" in the presidential rhetoric. Indeed, even if the figure for 2012 is much higher than that for the last year of Medvedev's presidency in 2011, it is very much comparable to the the number of mentions recorded in 2007 and 2008. If we focus on absolute numbers, only beginning with 2021 there is an effective increase in comparison with 2007-2008.[2] If instead we focus on relative numbers, we'll notice 2007, 2012, and 2022 as somewhat exceptional years, when about 0.4 per cent of all sentences included a reference to "values".

Ultimately, however, it is not the number of references to "values" per se that is most relevant: what type of values have been talked about at different points in time may be more revealing.

First, what "values" are we talking about?

There are many different types of "values" that appear in the Kremlin's statements. Here is a rather comprehensive list of the most common qualifiers or expressions used to characterise the values. They are categorised and colour-coded in order to make it easier to spot them in the quotes below. The categories are indicative and may be adjusted (suggestions welcome).

Table 1: Tipology of qualifers used to characterise values

pattern_ru	pattern_en	type
духовн*	spiritua*	traditional
традиц*	traditio*	traditional
историческ*	histor*	traditional
христианск*	christian	traditional
религиозн*	religious	traditional
библейск*	Biblical	traditional

[2] For those unfamiliar with Telegram channels and curious about where to start, the website tgstat.ru collects statistics about popular Telegram channels in each language.

pattern_ru	pattern_en	type
моральн*	moral	traditional
нравственн*	moral	traditional
семейн*	family	traditional
семья	family	traditional
патриот*	patrioti*	traditional
чужды*	alien	negative
ложн*	false	negative
искаженн*	warped	negative
тоталитарн*	totalitarian	negative
прав* человек*	human rights	Western
верховенств* закон*	rule of law	Western
свобод	freedom	Western
демократи*	democra*	Western
запад*	West*	Western
запад*	West	Western
европейск	Europ*	Western
либеральн*	liberal	Western
рыночн*	market	Western
рынка	market	Western
современн*	modern	Western
универсальн*	univeral	universal
общемиров*	global	universal
гуманистическ*	humanis*	universal
гуманиз*	humanis*	universal
гуманитарн*	humanitarian	universal
общечеловеческ*	human	universal
культурны*	cultur*	universal
цивилизац*	civiliz*	universal
цивилизац*	civilis*	universal
базов*	basic	universal
базов*	essential	universal
фундаментальн*	fundamental	universal
основны*	central	universal
общи*	common	common
общи*	shared	common
ценност*	values	values

Different types of values in the Kremlin's official statements

I will now proceed to different ways of looking at these data. First, here are all references to values found in transcripts, statements, and press releases published on the Kremlin's website between 31 December 1999 and 27 February 2023. Look at both tabs to see the quotes in either English or Russian.

For easier reference, keywords such as "values", and different types of values are highlighted and colour-coded - traditional, Western, universal, common, and negative.

> ℹ Note
>
> **N.B.** This is a static preview. See the online version for exploring all matches.

As will be immediately evident, different types of qualifiers can be present in the same sentence. Some of them are very generic; for example, talk of common values is common when meeting foreign delegations, no matter the country involved ("based on our common values…"). In a few instances, they are not even used as qualifiers to values, but in the overwhelming majority of cases there is a clear connection.

I will now proceed to count the frequency of different types of "values" at different point in times. The reader is however advised to have a quick look in the table above at some of the references in the early years as well as to those in recent months. Even if references to "spiritual values" are present in both, I would argue that a change in tone is evident to the human reader. This, of course, will not emerge from the numbers.

For example, on 11 June 2000, president Putin made reference to Russia within:

> "the unity of Europe on the basis of common values of progress, democracy and freedom" (source)

Reference to the values of democracy and freedom, are quite different in recent documents. For example, in a 21 February 2023 Presidential Address to the Federal Assembly, president Putin made reference to freedom and democracy along with "values", but this is made with scorn with reference to the West:

> "They can also continue to rob everyone under the guise of democracy and freedoms, to impose neoliberal and essentially totalitarian values…" (source)

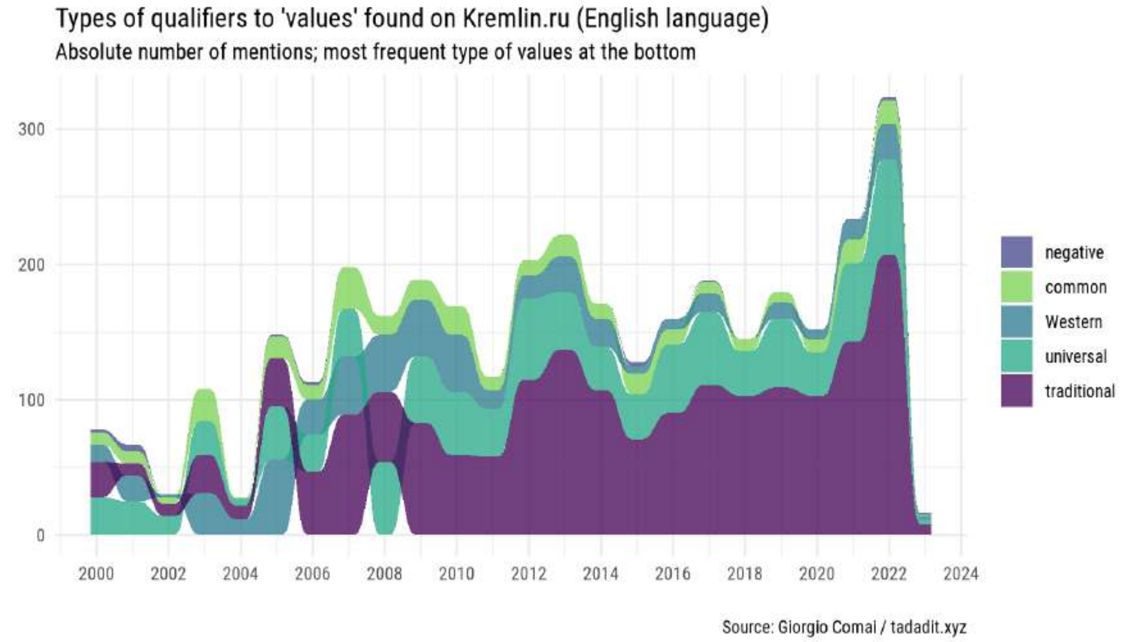
year	Western	traditional	universal	common	negative
2000	13	26	28	9	2
2001	19	9	25	9	5
2002	2	9	14	5	0
2003	31	28	25	24	0
2004	12	10	5	1	0
2005	56	36	39	16	1
2006	26	47	27	11	2
2007	43	89	35	31	0
2008	42	52	54	14	0
2009	42	83	49	15	0
2010	42	59	47	21	0
2011	14	58	35	10	0
2012	17	115	60	11	0
2013	26	137	43	16	0
2014	21	107	32	11	0
2015	6	71	33	15	3
2016	8	90	51	11	0
2017	14	111	54	8	1
2018	0	103	33	9	0
2019	12	109	51	8	0
2020	8	103	32	9	0
2021	15	143	58	17	1
2022	26	207	71	17	3
2023	3	8	3	1	1

Both of these quotes are counted as intances of “values” associated with multiple “Western” qualifiers.

Counting different types of values by the Kremlin

Here are a few ways to look at these data visually. The first type of plot allows to see both absolute number and ranking of the most common types of “values” found in this dataset. Looking at the graph with “Most frequent at the bottom”, it appears clearly how “traditional values” became the most common type of values since the mid-2000s, but also how the number of mentions has substantially increased in the following years. Switching to “Most frequent on the top” it emerges more clearly how the number of references to other types of values, taken together, have not really decreased… the total number has gone up due to increased references to “traditional values”. Finally, the faceted barchart allows to see better the developments within each category. For example, it appears more clearly how references to “Western values” have decreased even in absolute numbers (and in spite of the fact, which does emerge from this graph, that early mentions were almost exclusively positive, while many more of the recent mentions are effectively negative in sentiment); indeed, “Western values” were the type most commonly mentioned in both 2003 and 2004.

Most frequent at the bottom

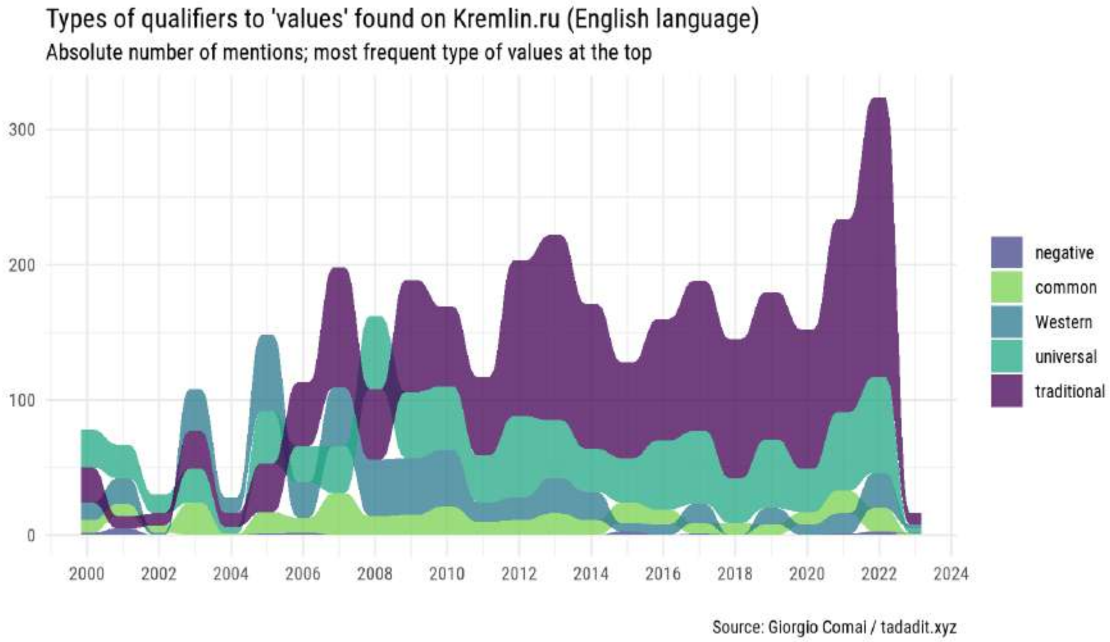


Types of qualifiers to 'values' found on Kremlin.ru (English language)

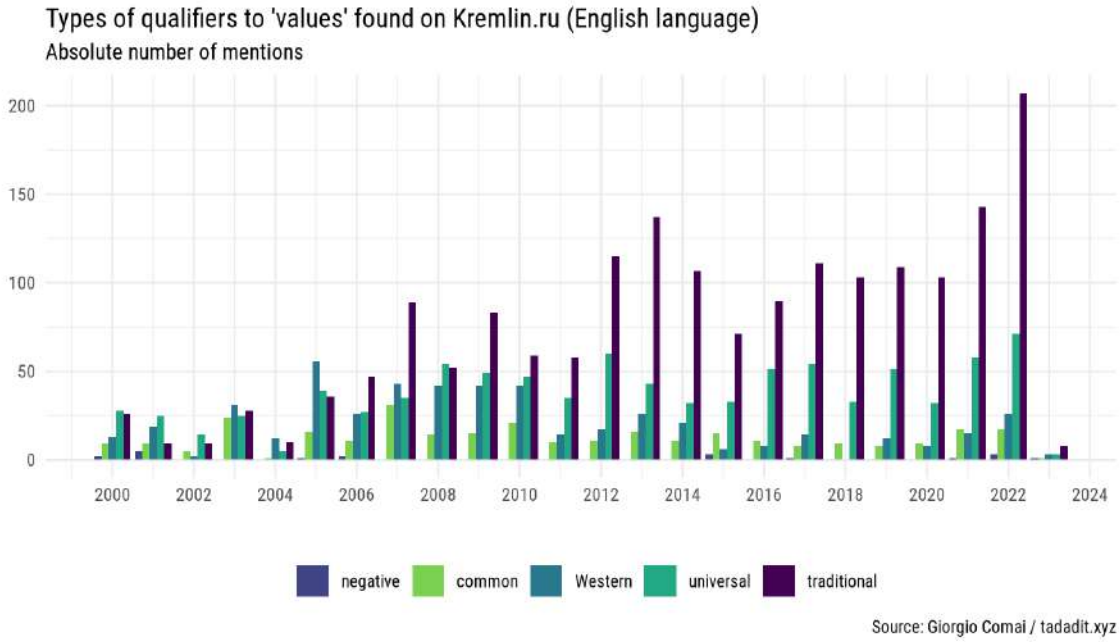
Absolute number of mentions; most frequent type of values at the bottom

Source: Giorgio Comal / tadadit.xyz

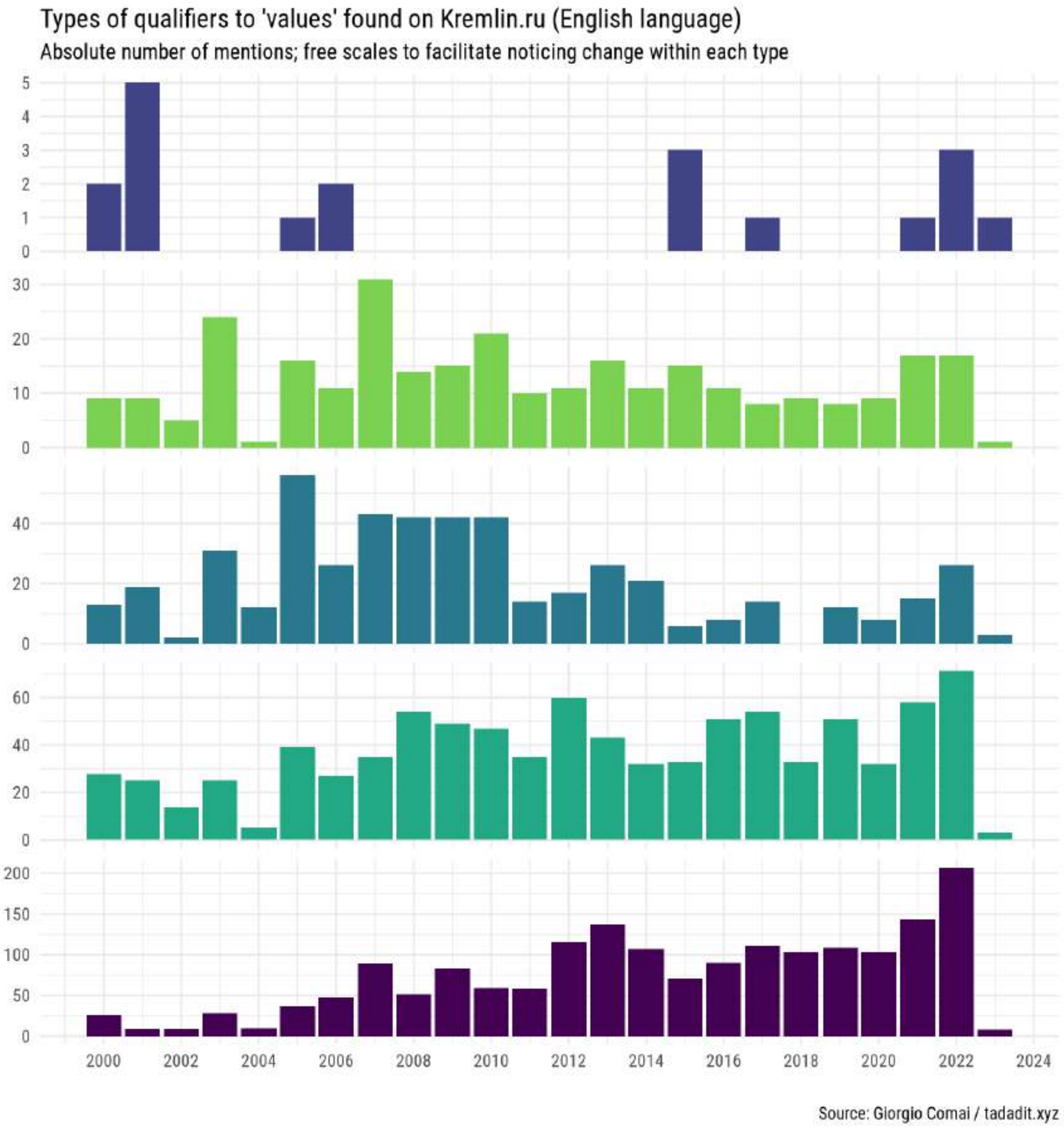
Most frequent on top



Dodged barchart



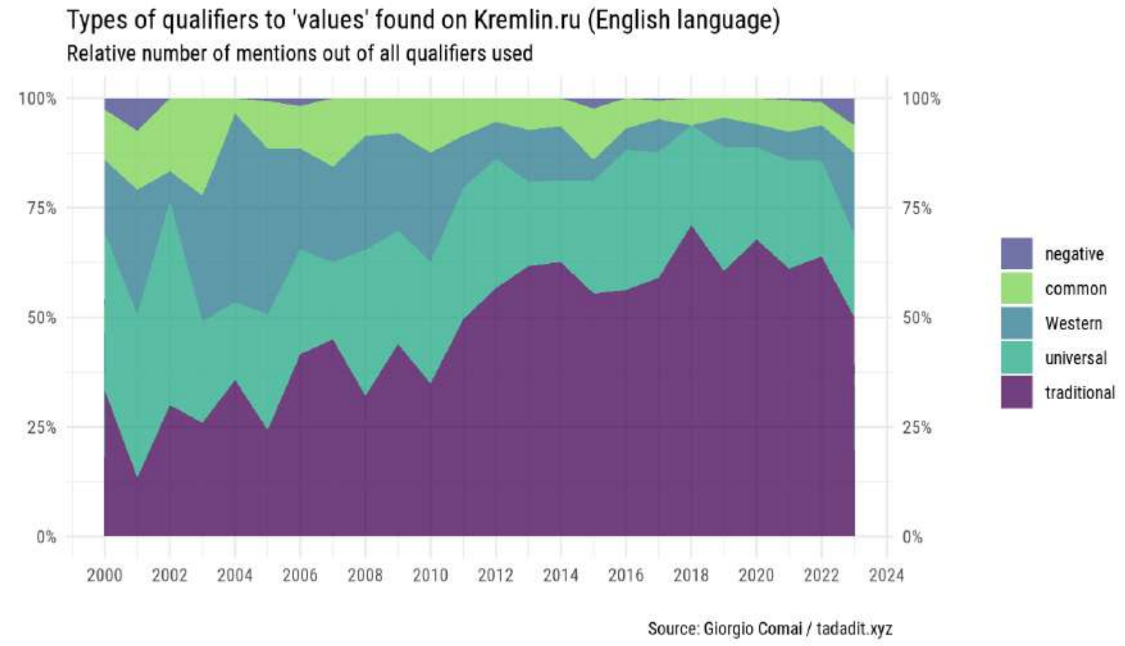
Faceted barchart



All of these graphs are in absolute, not relative number of mentions, which means that showed values are impacted by both the increased number of publications and the increased number of references to “values”.

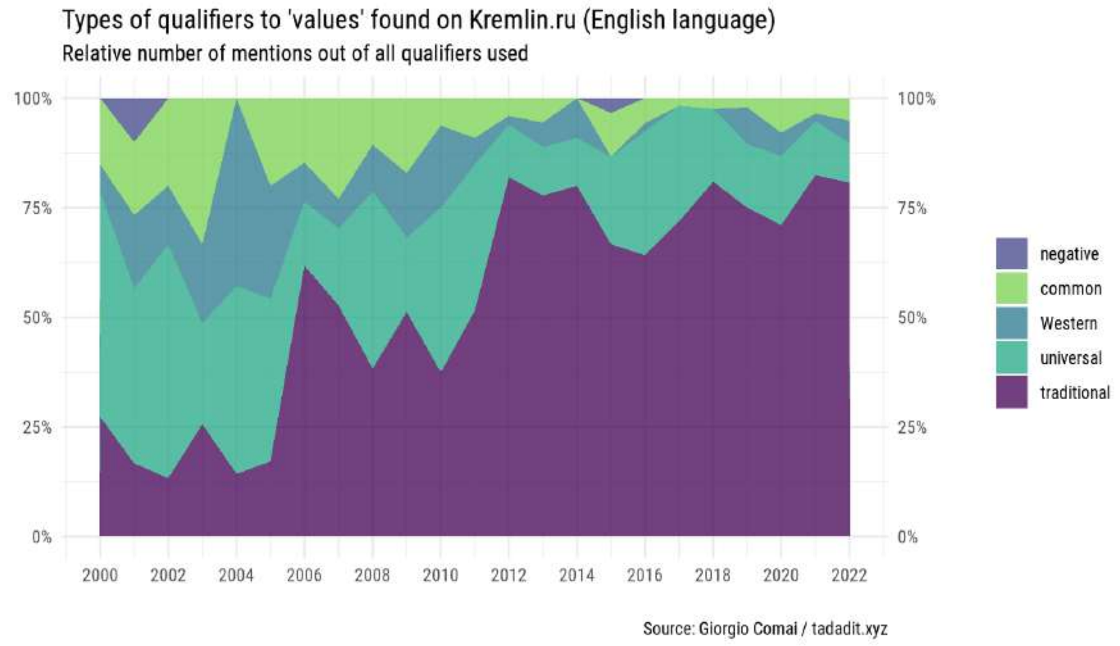
Here are two alternative ways to look at these data.

First, let’s show the relative frequency of each type of qualifier, out of all qualifiers found in the same sentence as “values”.



This shows a more clearly the increase in qualifiers of the "traditional" type. Considering that, as shown above, more than one qualifier can be attributed to a single reference to "values", this can partly be attributed to the increased frequency of multiple qualifiers associated with traditional values: in a sentence such as "traditional spiritual, moral and family values", the current method counts four instances of qualifiers of the traditional type.

An additional approach is to look specifically at the last word before "values", and drop further context. In this case, each instance of value can be associated with either one or zero categories (if the word preceding "values" is not in the list).



Types of qualifiers to 'values' found on Kremlin.ru (English language)

Relative number of mentions out of all qualifiers used

Source: Giorgio Comai / tadadit.xyz

This shows even more clearly the increase in references to traditional values. Among all references to values, only about 20 per cent were references to traditional values in the early 2000s. This grew to about 75% starting with 2012.

Looking specifically at traditional values

Finally, as we are specifically looking at “traditional values”, we will be separately singling out a variety of expressions used to refer to such values to make sure we are only capturing expressions that are very clearly identifiable as “traditional values”. Here is a table with the most common expressions used:

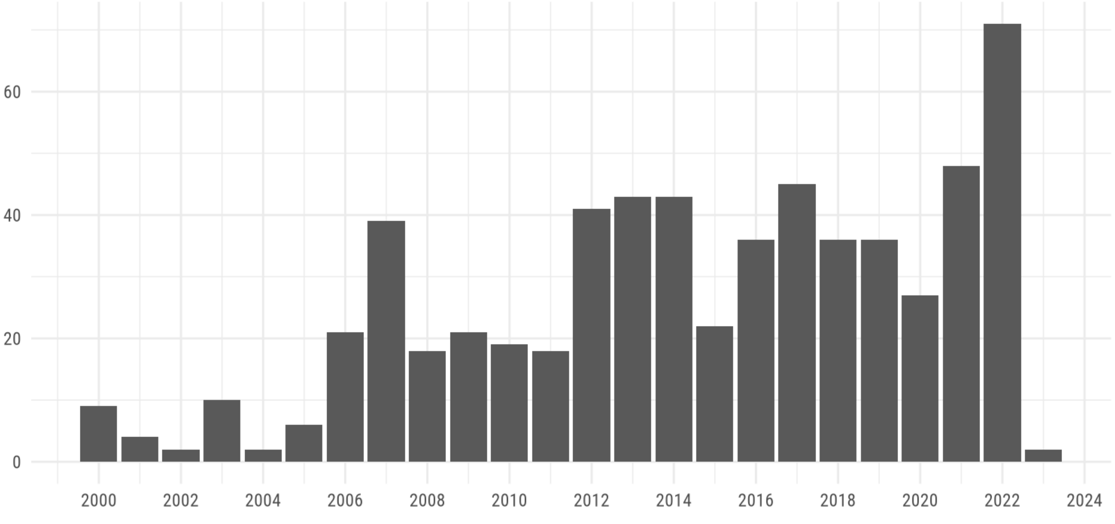
Table 3: Expressions used to characterise traditional values

Traditional values
traditional spiritual and moral values
traditional Russian spiritual and moral values
traditional spiritual, moral and family values
traditional family values
traditional values
traditional national values
traditional Christian values
traditional Orthodox Christian values
traditional ethical values

Traditional values
traditional cultural values
Christian values
Christian traditions and values
values of Christianity
traditional, conservative values
traditional, primarily Christian values
spiritual and moral values
spiritual, moral and family values
family values
spiritual values
moral values
moral and spiritual values
moral, spiritual and patriotic values
spiritual, moral and patriotic values
patriotic values
national values
values of patriotism
spiritual and patriotic values
patriotic, spiritual and moral values

How has the use of these specific expressions evolved through time?

Here is a graph showing the absolute number of mentions.

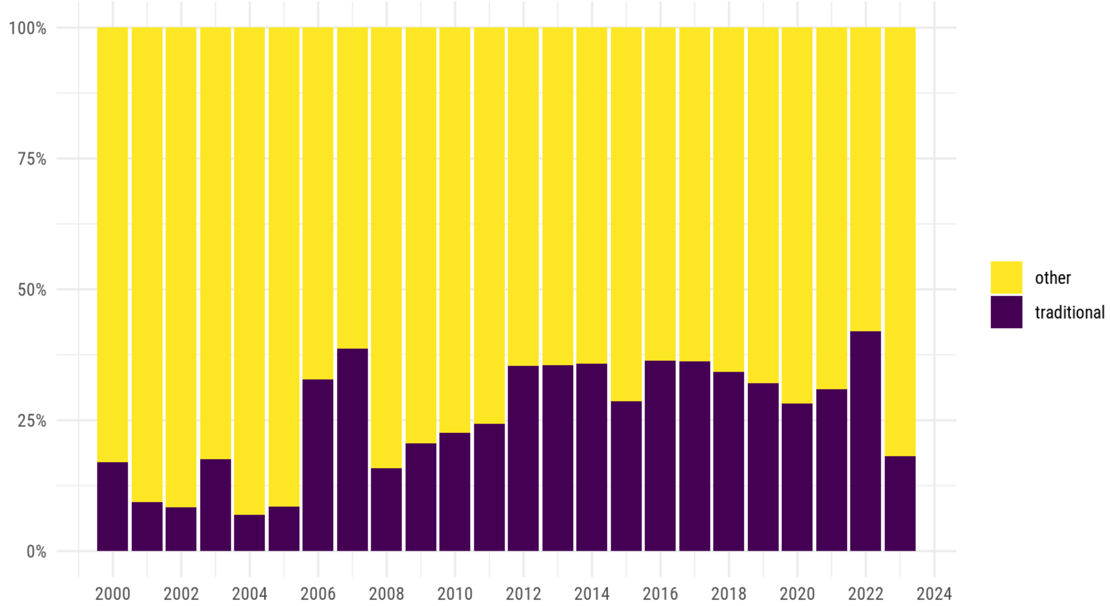


Absolute frequency of sentences mentioning traditional values on Kremlin.ru (English version)

Based on 821 552 sentences published between 31 December 1999 and 27 February 2023

Source: Giorgio Comai / tadadit.xyz

Here is a graph showing the relative number of mentions of the specific list of expressions associated with traditional values listed above, compared with all other references to "values".



It appears that indeed references to such expressions characterising traditional values were little used in the early 2000s; even if they did not fully enter the Kremlin's vocabulary in the Medvedev years, thay have been used rather frequently both in 2006 and 2007, and then more consistently (and in line with expectations stemming from the literature) starting with 2012, after Putin's return to the presidency, with the partial exceptions of 2015 and 2020 (in the latter case, perhaps due to more restrained public discourse in the context of the Covid pandemic). Finally, in 2022, references to "traditional values" increased very significantly in absolute terms, and quite noticeably in relative terms compared to other expressions referring to values (in a context in which total number of references to "values" has anyway been increasing).

What about mainstream media?

Pervy Kanal

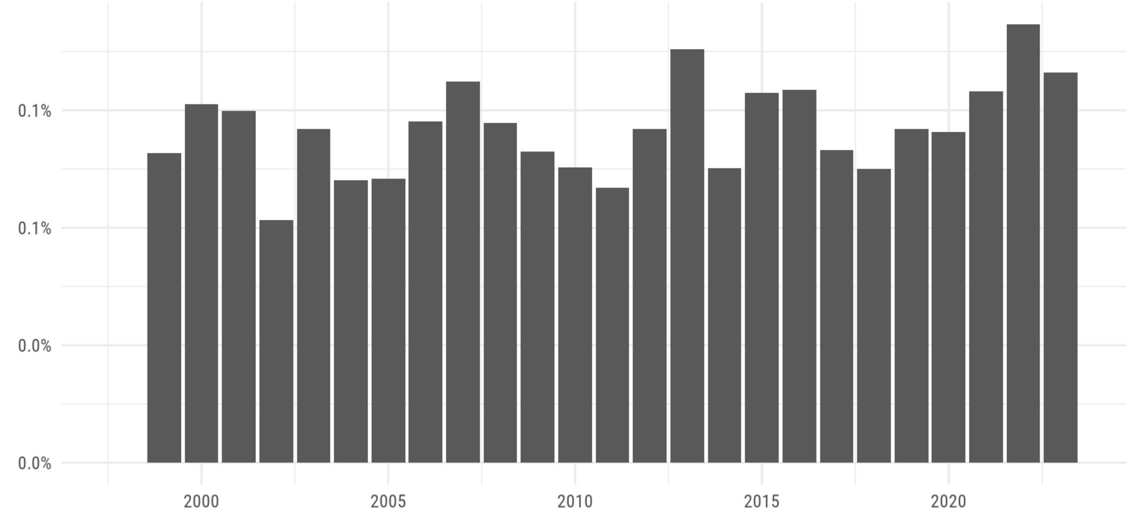
As appears from the the page describing the dataset, a very noticeable change in the format of web publications on 1tv.ru in 2022 makes it difficult to compare data starting with 2022 with earlier periods (in brief, it became common practice to include just a brief summary of news items rather than including full transcripts). There should still be value in the long term analysis, as *Pervy Kanal*'s website includes news starting from 1998.

Secondly, in Russian it is more difficult to single out references to “values” (“ценности”) in their moral sense, from other references to “value”, such as “the value of something”.

Relative frequency

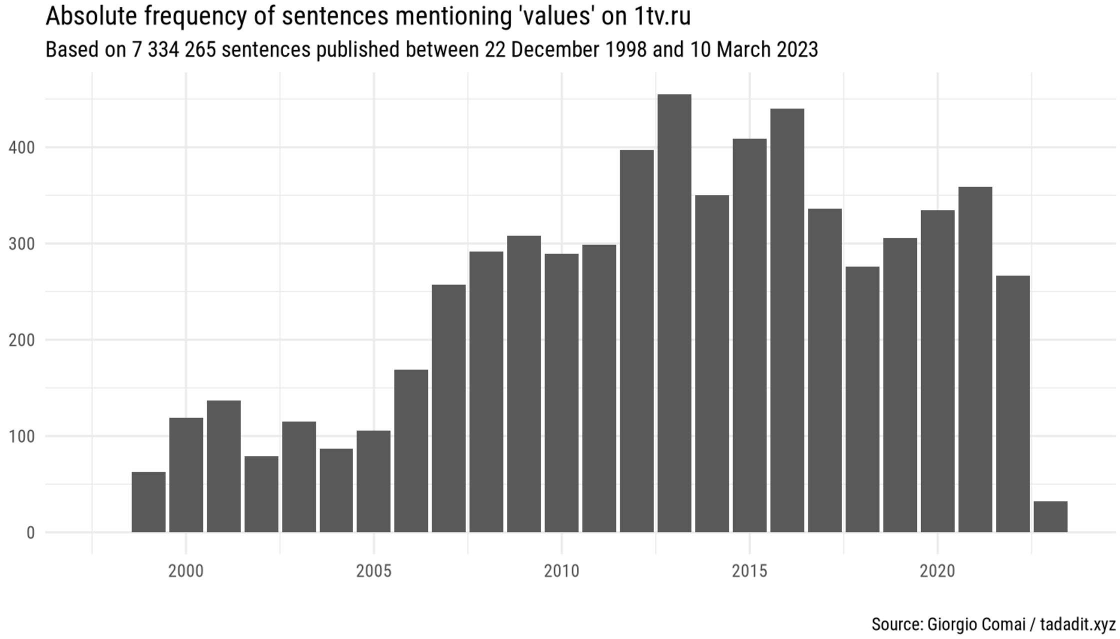
Relative frequency of sentences mentioning 'values' on 1tv.ru

Based on 7 334 265 sentences published between 22 December 1998 and 10 March 2023



Source: Giorgio Comai / tadadit.xyz

Absolute frequency



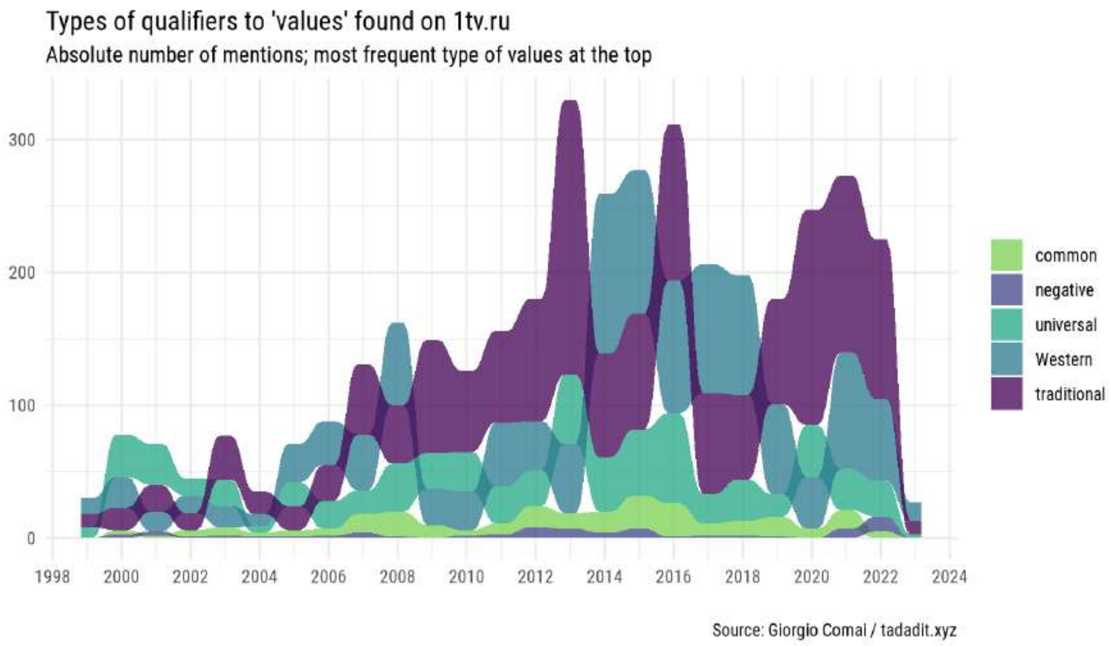
Tabular data

Table 4: Number of sentences that do or do not include references to ‘values’ per year on the English language version of the Pervy Kanal’s website (1tv.ru)

year	all	with_values	without_values
1998	434	0	434
1999	79 615	63	79 552
2000	130 001	119	129 882
2001	152 488	137	152 351
2002	127 686	79	127 607
2003	135 061	115	134 946
2004	120 577	87	120 490
2005	146 143	106	146 037
2006	193 819	169	193 650
2007	264 106	257	263 849
2008	336 668	292	336 376
2009	388 110	308	387 802
2010	382 989	289	382 700

year	all	with_values	without_values
2011	425 266	299	424 967
2012	465 610	397	465 213
2013	431 217	455	430 762
2014	465 516	350	465 166
2015	432 624	409	432 215
2016	462 492	440	462 052
2017	421 170	336	420 834
2018	367 524	276	367 248
2019	359 233	306	358 927
2020	396 872	335	396 537
2021	378 470	359	378 111
2022	238 480	267	238 213
2023	32 094	32	32 062

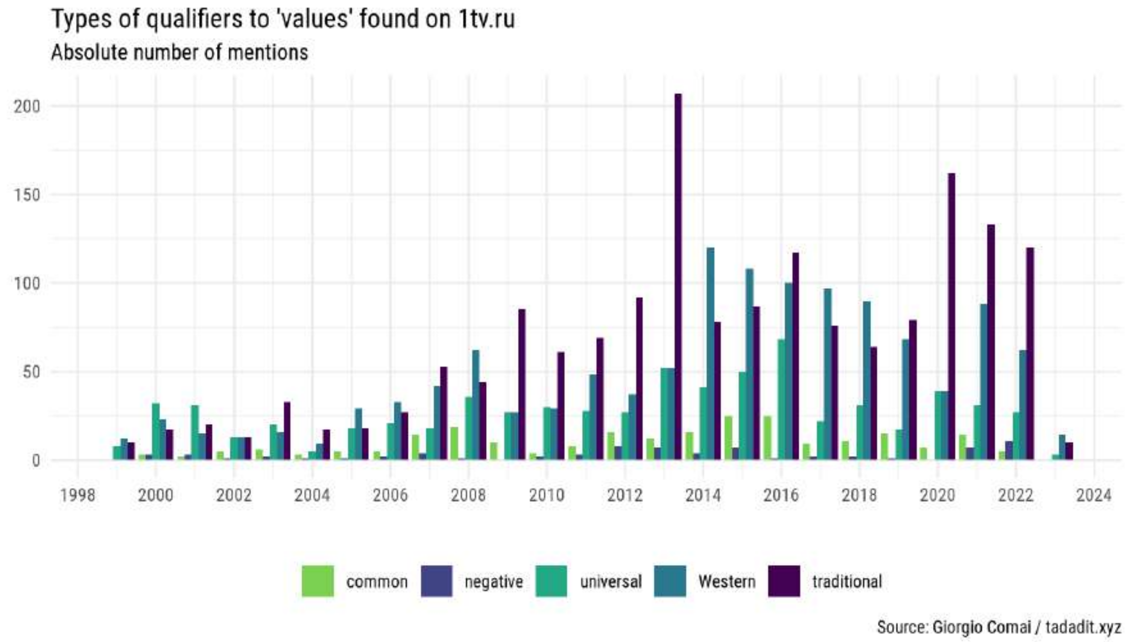
Most frequent on top



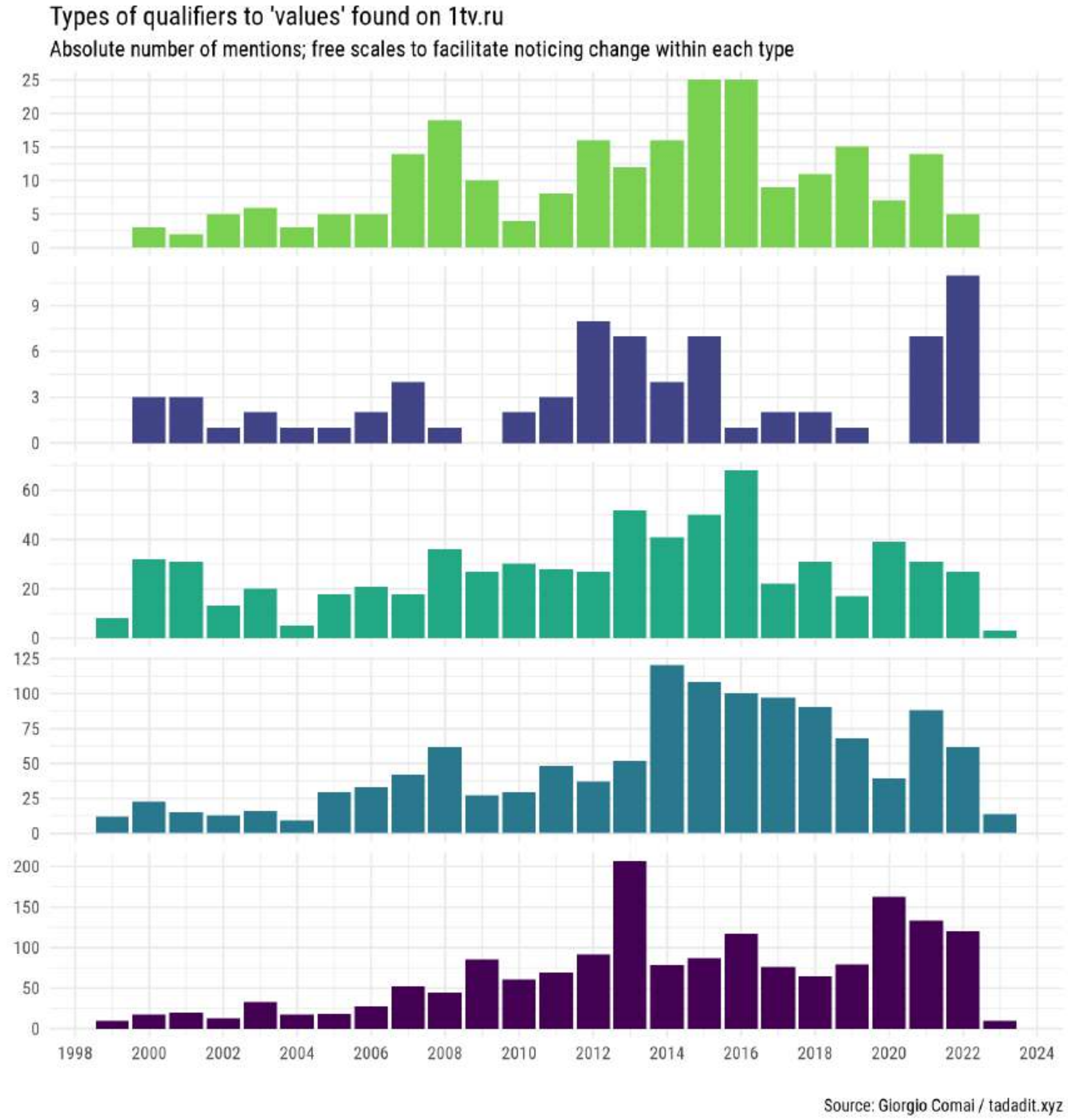
Source: Giorgio Comai / tadadit.xyz

year	Western	traditional	universal	common	negative
1999	12	10	8	0	0
2000	23	17	32	3	3
2001	15	20	31	2	3
2002	13	13	13	5	1
2003	16	33	20	6	2
2004	9	17	5	3	1
2005	29	18	18	5	1
2006	33	27	21	5	2
2007	42	53	18	14	4
2008	62	44	36	19	1
2009	27	85	27	10	0
2010	29	61	30	4	2
2011	48	69	28	8	3
2012	37	92	27	16	8
2013	52	207	52	12	7
2014	120	78	41	16	4
2015	108	87	50	25	7
2016	100	117	68	25	1
2017	97	76	22	9	2
2018	90	64	31	11	2
2019	68	79	17	15	1
2020	39	162	39	7	0
2021	88	133	31	14	7
2022	62	120	27	5	11
2023	14	10	3	0	0

Dodged barchart

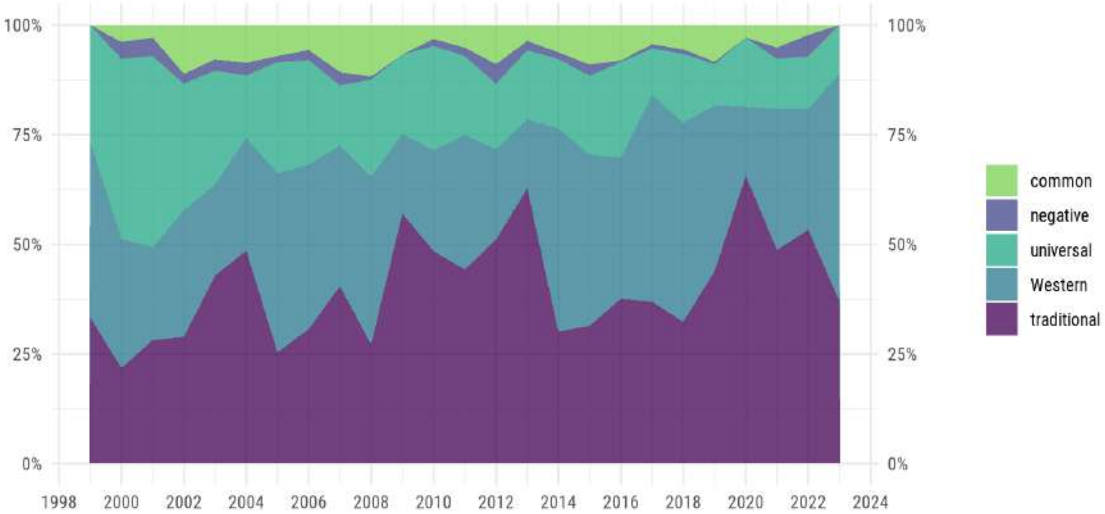


Faceted barchart



Types of qualifiers to 'values' found on 1tv.ru

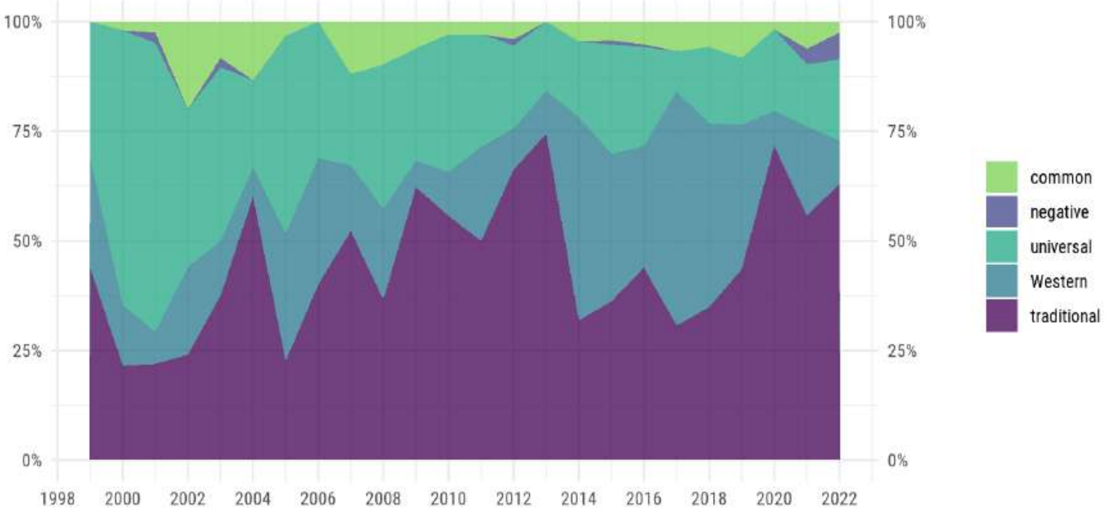
Relative number of mentions out of all qualifiers used



Source: Giorgio Comai / tadadit.xyz

Types of qualifiers to 'values' found on 1tv.ru

Relative number of mentions out of all qualifiers used immediated before 'values'



Source: Giorgio Comai / tadadit.xyz

From the ‘battle of Bakhmut’ to the ‘march of justice’: Prigozhin’s audio files, transcribed

Originally published on tadadit.xyz on 22 August 2023

https://tadadit.xyz/posts/2023-08-telegram_prigozhin/

ℹ Summary of key results

- it is possible to use open source tools for transcribing audio messages locally and the quality is reasonably good also in the case of messages such as Prigozhin’s filled with slang and expletives (adding automatic translation, the contents are still mostly readable, but the quality is noticeably degraded)
- exploring mixed-media contents published on Telegram channels may be challenging, but feasible
- just looking at changes in the frequency of posting it is possible to observe Prigozhin’s radicalisation journey
- a full dataset with all of Prigozhin’s audio messages transcribed is available for download, or to consult in a single page in Russian or with automatic translation in English

Context

As the Kremlin tightened its control of narratives and news that feature in mainstream media, Telegram has gained a significant role as the venue where Russian citizens of different persuasions look for information and opinions. Indeed, Telegram has remained one of the few uncensored on-line spaces (another one being YouTube) that can be freely accessed from Russia without having to rely on VPNs or other censorship circumvention techniques.

In many ways, mainstream media and Telegram channels seem to be two parallel information spaces, with debates and news that are dominant on Telegram (where re-posts among popular channels are common) may be completely ignored by major broadcast media. Indeed, the invisible line tightly separating these spaces is punctured only occasionally, and figures that are prominent on Telegram or even Western discourse about the war would be almost unknown to people who relied strictly on federal TV channels to get their news.

In the full web archive of news of Russia's *Pervy Kanal*, there is literally only a handful of mentions of Evgeny Prigozhin until June 2023, all of them related to questions Putin has received in interviews in earlier years and that refer to Prigozhin's involvement with the so-called "troll factory" based in Saint Petersburg. But there is no reference to his role in Ukraine, not even during the months-long battle of Bakhmut; not even a hint or passing reference to the growing tensions between Prigozhin and the Ministry of Defence that marked the months preceding Prigozhin's mutiny.[1] And yet, most respondents to opinion polls seemed to know enough about Prigozhin to have an opinion about him. For a brief period before the mutiny, he was one of the public figures most frequently mentioned approvingly by survey respondents, at one point even the most frequently mentioned after president Putin, even if this is likely more the result of a relatively small number of strong supporters rather than of widespread support.

Either way, it seems clear that contents spread through Telegram reach a substantial part of the Russian population. Telegram channels is also the primary way used by figures such as Prigozhin to share their opinions and messages. In brief, there is plenty of good reasons for scholars interested in the spread of information and narratives related to Russia's invasion of Ukraine to dedicate some attention to Russian-language Telegram channels. Indeed, there have been some efforts in this space that outline the prevalence of pro-Kremlin channels on Telegram.[2]

Rather than dealing with a large number of Telegram channels and their interactions, this post focuses on the task of analysing the contents published by a single figure - Evgeny Prigozhin. It is an interesting case not only because of its obvious relevance in relation to the war, but also technically, because of the variety of formats it employs as well as the peculiarity of each format for conveying different messages. Indeed, as will become apparent by the end of this post, in the case of Prigozhin the switch from written text to audio messages

[1] This applies only to the online news archive of 1tv.ru, which does not include full transcripts of all broadcasts. Even if Prigozhin's role may have emerged in debates during talk shows, his complete absence from standard news reporting remains telling.

[2] For those unfamiliar with Telegram channels and curious about where to start, the website tgstat.ru collects statistics about popular Telegram channels in each language.

has effectively characterised the radicalisation journey of Prigozhin's public persona.

Step 0: Understanding Prigozhin's presence on Telegram

How does Prigozhin's communication work? In brief, Prigozhin's press service actively responds via Telegram to questions received by journalists via email. Questions are mostly posted as screenshots, while responses have been increasingly posted as audio messages. Other posts may be in video format, including clips with prisoners, combat images, or video clips with Prigozhin's voice.

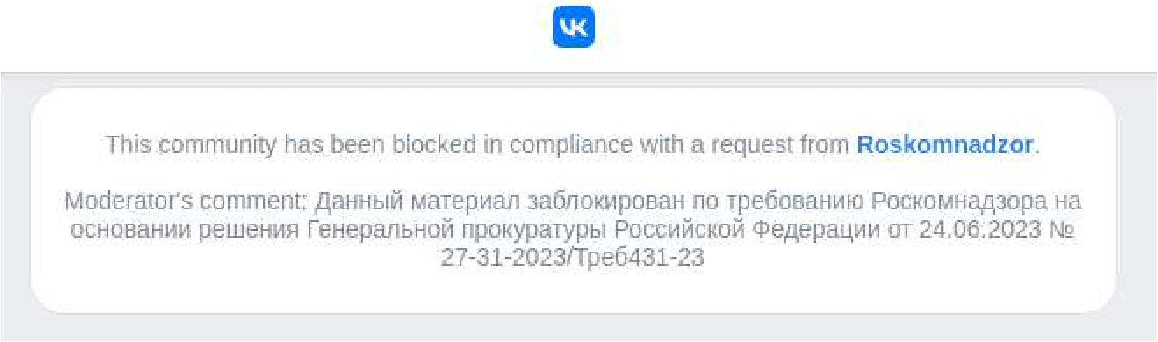
There is an additional difficulty: Prigozhin's communication is surprisingly orderered in some respects, messy in others.

Prigozhin has one main official Telegram channel that he uses for "official" communication, which is called "Prigozhin's press office", and has the Telegram handle @concordgroup_official, "Concord Group" being the name of the holding company that controls Prigozhin's various businesses. At the time of this writing in August 2023, the channel has almost 1 million 250 thousands subscribers.

Let's start with the surprisingly ordered part: (almost) each message starts with a numeric identifier, in a format such as the following: "#903 Запрос от редакции газеты…" ("#903 Question from…"). The post on 26 June in which Prigozhin offers some "clarifications" on the mutiny which has by now over 4 million views starts with "#1851 We publish the response of Evgeny Prigozhin…". As of August 2023, no other message has been posted on this channel. So in principle, everything looks nice and clear: there are so far 1851 statements by Prigozhin.

Now, let's move on to the messy part.

First, there is no official website to take as a point of reference. The official Telegram channel's bio includes a link to the official page on *Vkontakte*, a popular Russian service similar to Facebook. However, the page on VKontakte has been blocked after the mutiny by request of the Russian authorities and appears to be still blocked with the following notice:



This community has been blocked in compliance with a request from Roskomnadzor.

Moderator's comment: Данный материал заблокирован по требованию Роскомнадзора на основании решения Генеральной прокуратуры Российской Федерации от 24.06.2023 № 27-31-2023/Треб431-23

The Telegram channel itself was opened only in November 2022, with message #897. Previous messages were probably published first on VKontakte (now not reachable), and presumably reposted from there in other Telegram channels.

Telegram channel "Prigozhin's hat", for example, has over 500.000 subscribers and was opened in February 2019. Even earlier posts look exactly the same as ones published on the official channel, and it appears they are effectively just (partially automated) re-posts. However, unfortunately, they do not include the same numeric id used in official posts. Besides, there are some more contents published on *Prigozhin's hat* that do not feature on the official channel, including a few posts published *after* the mutiny. These are mostly forwarded from other Prigozhin-related channels such as "Razgruzka Vagnera" or "SOMB - 'Tourists in Africa" (related to Wagner's presence in Africa). On 21 August 2023, for example, a new video post by Prigozhin aimed at recruiting personnel for Wagner missions in Africa appeared; an audio message posted as a response to a question supposedly by an African media posted as a screenshot in French has appeared on the same day. A new chapter in Prigozhin's communication efforts may be beginning.

"Prigozhin's hat" may at this stage be the easiest source for earlier posts issued by Prigozhin (as well as possibly for the post-mutiny period)^[Others sources may well be available; I welcome suggestions about full archives if available.], while @concordgroup_official is the most consistent source for recent months. As will appear from the following sections, it is really starting from early 2023 the Prigozhin stepped-up his virulent rhetoric expressed through audio messages, so in many respects it makes sense to focus on more recent contents and look at previous contents only as a term of reference.

To summarise, contents posted by Prigozhin's press service are a combination of:

- text messages
- text included as screenshots of emails (and, occasionally, documents)
- audio messages
- video clips of different length and format

How do we turn these into something that can be searched and analysed?

Step 1: Get the data out of Telegram

From Telegram Desktop, it easy to export the full archive of Telegram channel in machine-readable format, exporting all posts with metadata as a single .json file, as well as all images and files in dedicated subfolders.

First, let's have a look at some basic information about the dataset we have:

Earliest post: 2022-11-05

Latest post: 2023-06-26

Total number of posts: 1 242

Total number of audio files: 408

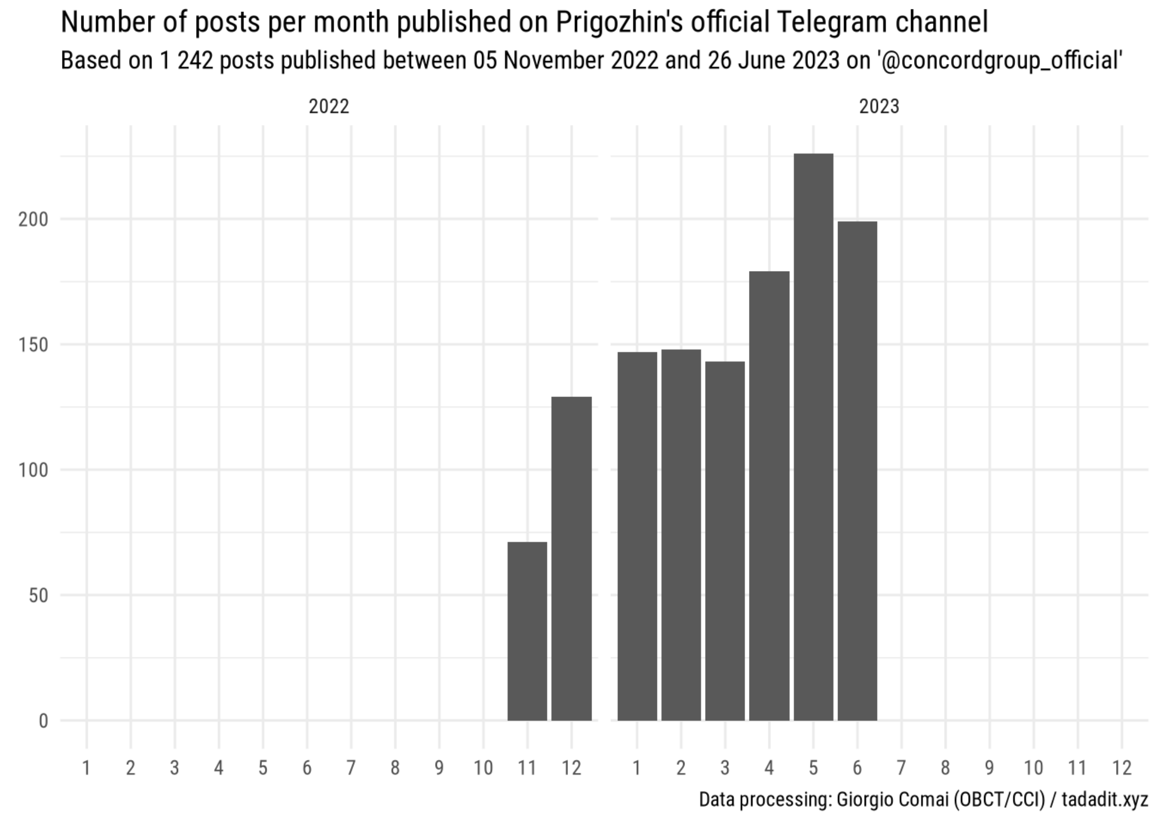
Earliest post with audio file: 2022-12-26

Latest post with audio file: 2023-06-26

Total duration of audio files: 29628s (~8.23 hours)

Average duration of audio files (in seconds): 73

Median duration of audio files (in seconds): 51

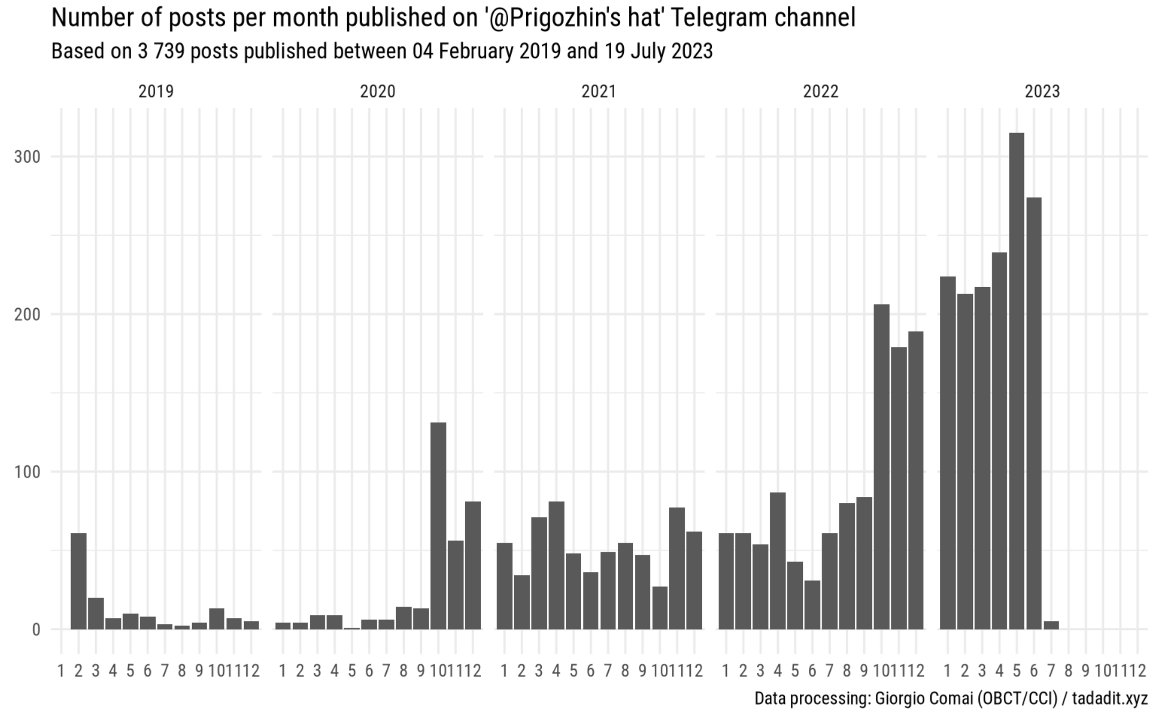


Number of posts per month published on Prigozhin's official Telegram channel

Based on 1 242 posts published between 05 November 2022 and 26 June 2023 on '@concordgroup_official'

Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

For reference, it may be useful to look at “Prigozhin’s hat” to look at the frequency of posts in earlier months.



Number of posts per month published on '@Prigozhin's hat' Telegram channel

Based on 3 739 posts published between 04 February 2019 and 19 July 2023

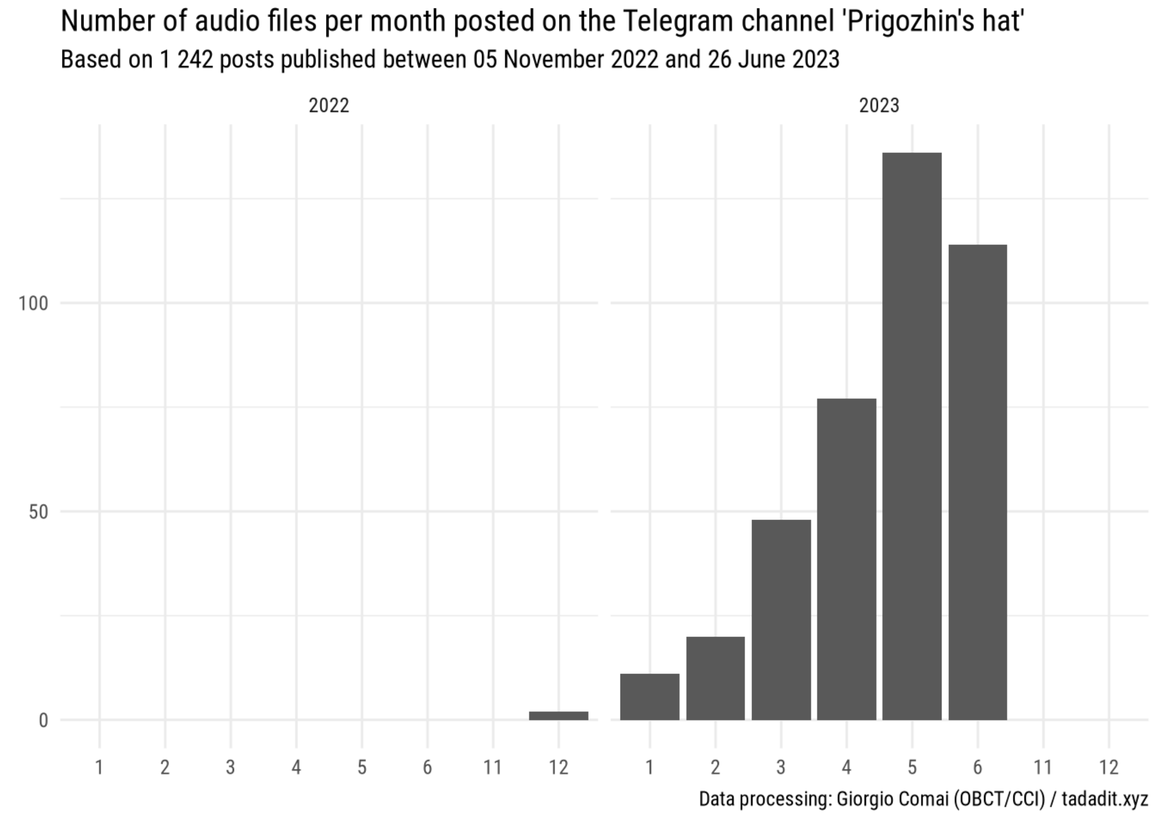
Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

It appears there is a distinct *crescendo* in the number of posts published by this channel (presumably reflecting Prigozhin’s overall post frequency also on its currently unavailable official channels), from just a handful of posts per month until September 2020, then mostly between 40 and 80 monthly posts until September 2022, going up to more than 200 post per month until the end of June 2023, when the channel fell silent post-Mutiny, after averaging close to 10 posts per day in the previous weeks

Even these basic descriptive statistics reflect some of the things we know about Prigozhin: the big increase in posts in October 2022 can easily be explained by the fact that it is only then, more precisely on 26 September 2022, that Prigozhin publicly admitted its ties to Wagner. Two days earlier, on 23 September 2020, the US treasury significantly expanded its sanctions against entities linked to Prigozhin, which may be related to him taking a more public role.

Since the very beginning of its online presence, Prigozhin’s press team published the questions it received as a screenshot, and added Prigozhin’s own reply either in the text of the message or as an additional screenshot with text. As emerges from the following graph, it is basically only starting with 2023 that Prigozhin started to respond with audio messages - often, angry rants - that quickly became a trademark element of its communication.

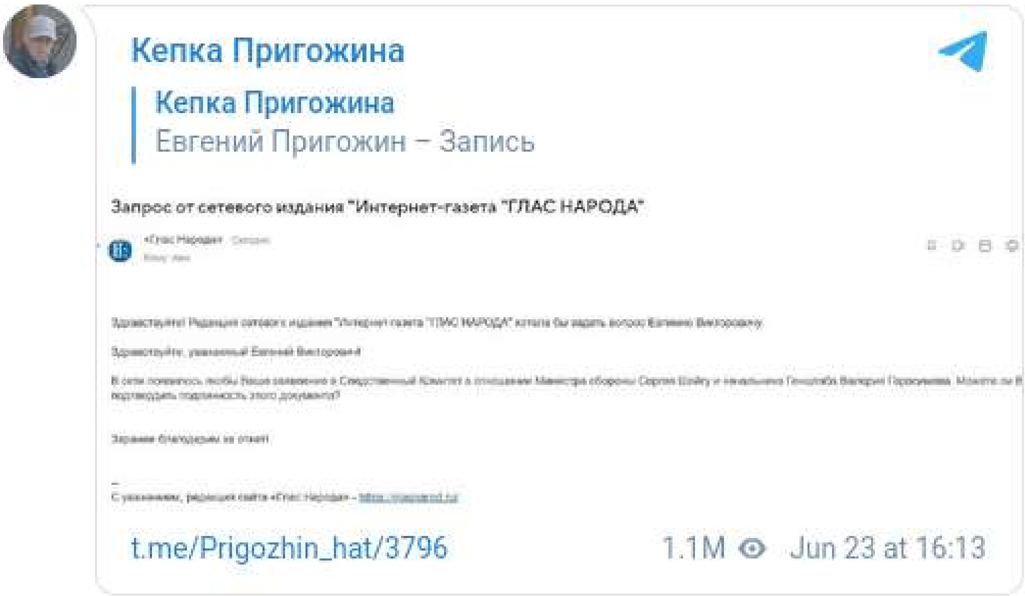
As a consequence, since the focus of this post is Prigozhin's audio messages and its rhetoric escalation in recent months, for the rest of this post I will mostly stick to Prigozhin's official channel: as no audio message was posted before the new channel has been opened, key contents should all be there.



Step 2: An overview of the kind of posts published

Posts published on the official Telegram channel of the Prigozhin's press service as well as on "Prigozhin's hat" Telegram channel including older posts by Prigozhin are mostly based on a combination of formats; sometimes the contents are repeated in more than one format, sometimes they are not.

For example, this post shows a question asked by a media organisation as a screenshot:



Кепка Пригожина

Кепка Пригожина
Евгений Пригожин – Запись

Запрос от сетевого издания "Интернет-газета "ГЛАС НАРОДА"

«Глас Народа» Сегодня
Кому: мне

Здравствуйте! Редакция сетевого издания "Интернет газета "ГЛАС НАРОДА" хотела бы задать вопрос Евгению Викторовичу.

Здравствуйте, уважаемый Евгений Викторович!

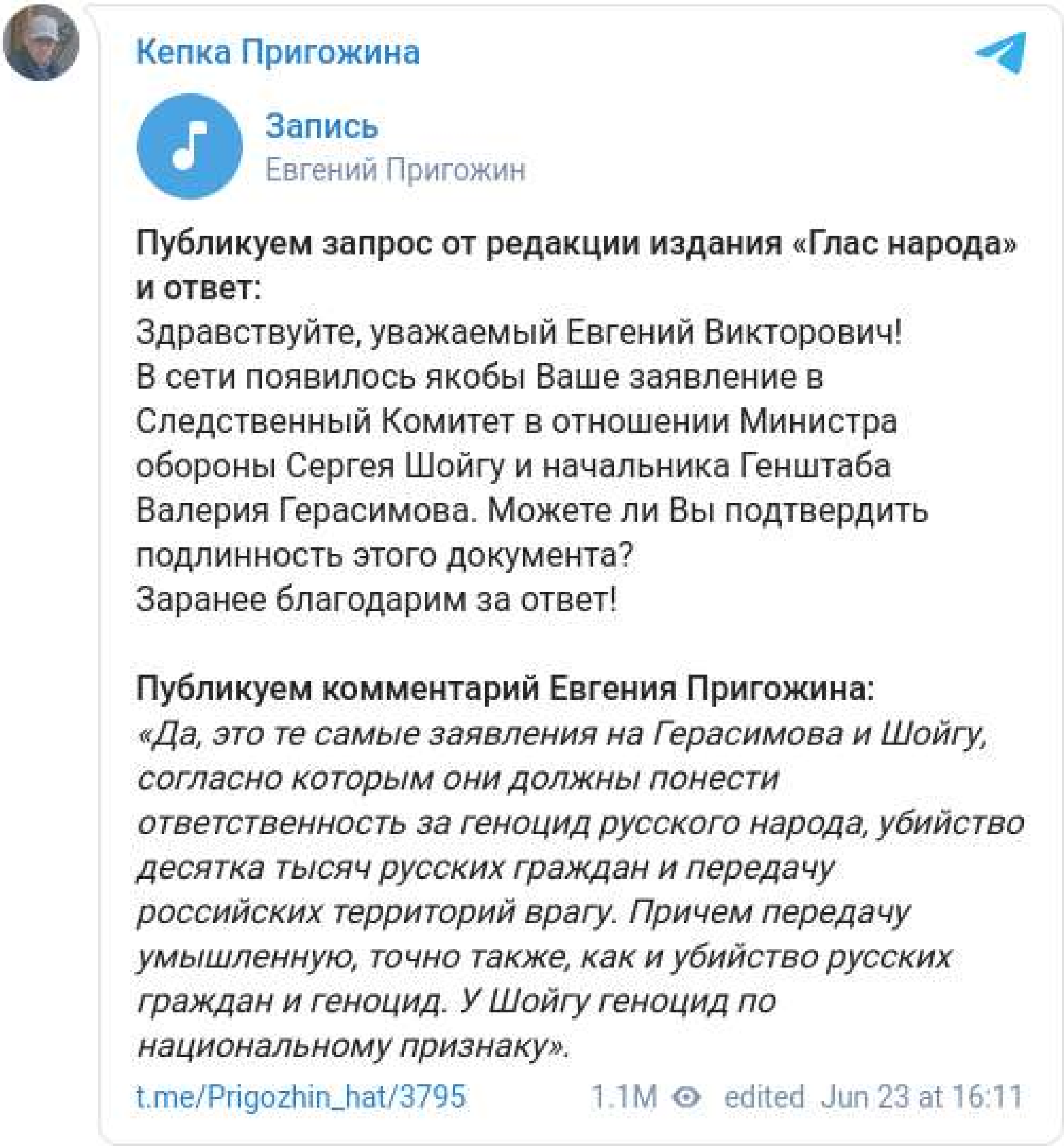
В сети появилось якобы Ваше заявление в Следственный Комитет в отношении Министра обороны Сергея Шойгу и начальника Генштаба Валерия Герасимова. Можете ли В подтвердить подлинность этого документа?

Заранее благодарим за ответ!

--
С уважением, редакция сайта «Глас Народа» – https://glasnarod.ru

t.me/Prigozhin_hat/3796 1.1M Jun 23 at 16:13

Conveniently, this is accompanied by another post that includes both the question and the answer given in both textual and audio format:



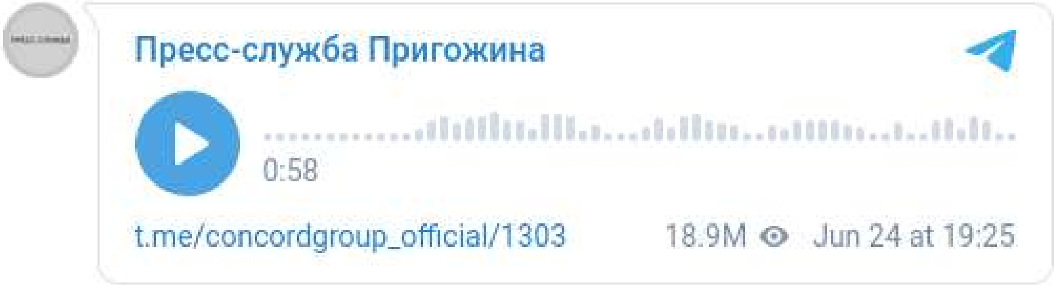
Публикуем запрос от редакции издания «Глас народа» и ответ:
Здравствуйте, уважаемый Евгений Викторович!
В сети появилось якобы Ваше заявление в Следственный Комитет в отношении Министра обороны Сергея Шойгу и начальника Генштаба Валерия Герасимова. Можете ли Вы подтвердить подлинность этого документа?
Заранее благодарим за ответ!

Публикуем комментарий Евгения Пригожина:
«Да, это те самые заявления на Герасимова и Шойгу, согласно которым они должны понести ответственность за геноцид русского народа, убийство десятка тысяч русских граждан и передачу российских территорий врагу. Причем передачу умышленную, точно также, как и убийство русских граждан и геноцид. У Шойгу геноцид по национальному признаку».

t.me/Prigozhin_hat/3795 1.1M edited Jun 23 at 16:11

In this case, everything seems easy: we can in principle ignore both the screenshotted picture and the audio-file, as the very same contents are presented also in textual format.

But then, in other occasions there are only audio-files or voice messages with no context whatsoever. This was the case, for example, for most messages posted during the mutiny on 24 June, including the one that announced its end:



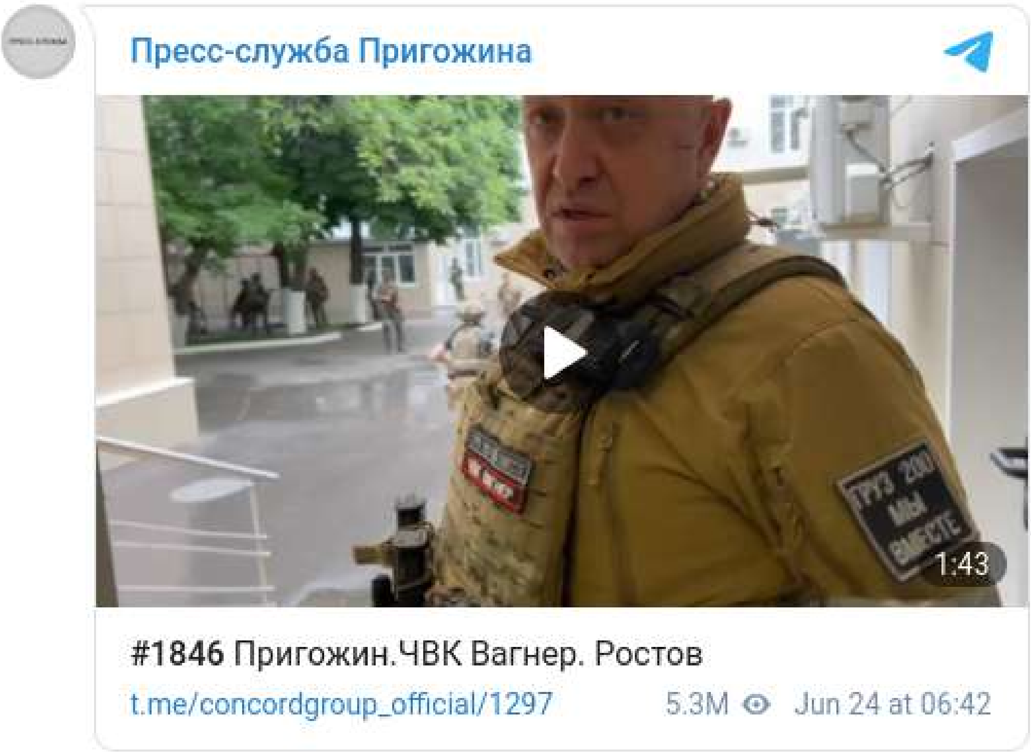
Пресс-служба Пригожина

0:58

t.me/concordgroup_official/1303 18.9M Jun 24 at 19:25

In others still, the content of the question previously-screenshotted is transcribed, but Prigozhin's comments are conveyed only in audio format.

Finally, there are occasional posts including some documents or video files:



Пресс-служба Пригожина

#1846 Пригожин.ЧВК Вагнер. Ростов

t.me/concordgroup_official/1297 5.3M Jun 24 at 06:42

Video files often include spoken comments, or depict meetings. They are only very occasionally central to Prigozhin's communication, and even when they are, mostly not for the spoken content. Video files should not be dismissed, however, and they may actually be an important part of the communication of other Telegram channels, all the way from Strelkov to the "military bloggers" who produce video contents. In the rest of this post, I will leave out video clips, both for facilitating consistency in the processing of results and because including them introduces further ethical questions (they include, among other things, the voice of Ukrainians prisoners of war).

In the following steps, I will proceed with turning images into text (only briefly) and then really focus on turning audio messages into text format that can be searched and processed further.

Step 3: OCR images

As screenshots of text have become less common on Prigozhin's official channel in recent months, and as they are broadly a format less frequently found on Telegram in comparison to video and audio files, I will go through the image part quickly, and then focus more on audio messages. Given the prevalence of textual screenshots in earlier posts, this section will take messages from "Prigozhin's hat" Telegram channel, rather than the official one; the vast majority of posts are exactly the same on both channels, but, as mentioned before, "Prigozhin's hat" has a lot more of the early posts.

OCR techniques to recognised text from images are well established. In this specific case, the quality of results is hindered mainly by two aspects:

- low resolution of the images
- the fact that many of these are screenshots of emails, and they may include some email metadata at the top, or some signature text at the bottom of the email
- the fact that there are sometimes more than one language in the same image, either because there's some clutter in the email screenshots, or because questions are asked in English and the response given in Russian (in the vast majority of cases, however, both questions and answers are given in Russian)

The following is a quick attempt to extract the text of the images via OCR, with no particular effort dedicated to polishing the results. Even so, the process allows to conduct quick searches among transcribed text. For example, if you look for "Wagner" ("Вагнер") in the search box for the `text_photo` column, only posts where "Wagner" is mentioned in the screenshotted text will be kept. For the records, this shows that out of 1 839 posts with valid text extracted from the images, 491 mention "Wagner", all the way from the early days of the channel back in 2019 when Prigozhin was still vehemently denying any association with it.

Some information about the following table:

- the table includes all posts that have attached a photo from where seemingly meaningful text could be automatically extracted
- the text has been automatically extracted with OCR with `tesseract`, setting the language as Russian (hence, the glaring inaccuracies where the images include contents in other languages)
- if the post has attached more than one image, the text for each image is included in a separate row; the embedded post is always the same and it may not be immediately obvious that it includes more than one picture
- very often, the response to the question is given in a separate post: clicking through the embedded post, and then clicking on "context" may be helpful in finding more details in the posts immediately preceding or following any given post.

> ℹ Note
>
> **N.B.** This is a static preview. See the online version for exploring the data.

For more detailed analysis, and depending on the type of analysis, this would likely require some more polishing efforts. Also, as the same textual content is often repeated both in the screenshotted image and as text in the original post, this may lead to extensive duplication of contents. On the other hand, if one is not into word frequency analysis but just into more effective ways to search through all contents of the channel, this may well already be of use.

Step 4: Speech-to-text of audio and video attachments

One of the most distinguishable features of Prigozhin's mutiny for external observers was just how much it was communicated through Telegram posts, mostly bare audio messages of Prigozhin's raucous voice: the mutiny was launched with an audio message and its end was declared in the same way. Indeed, audio messages had become an increasingly central component of Prigozhin's approach to communication, as it was perhaps most fitting to his harsh and increasingly unhinged comments towards Russia's military leadership and the Ministry of Defence.

Many of these audio messages have been transcribed and posted as text in the same post that has the audio file, or in a separate message posted immediately sooner or later. However, others were not: these may be longer audio messages, messages with more vulgar or explicit expressions, as well as most of the messages posted during the mutiny on 24 June 2023.

Due to the molteplicity of formats employed, systematically harmonising these contents through consistent deduplication may be challenging. But the first step in that direction surely involves transcribing these audio messages.

There are a variety of online services that offer speech-to-text as a service for a fee. Depending on the use case, the budget available, and the amount of audio to be transcribed they may be a valid option. Processing data locally introduces other constraints - mostly, computing time and resources - but allows for reproducibility, and makes irrelevant a set of additional concerns that may emerge by relying on third parties, including questions such as:

- “are third-party vendors fine with me using their services for transcribing profanities by an alleged war criminal?”
- especially for those working on violent extremism or terrorism, “will sending a bunch of extremist materials get me into trouble”?
- if I am transcribing non-public and potentially sensitive contents, is it fine ethically to send them through a third-party, perhaps one based in different jurisdiction?

Terms of service may offer some assurances, but processing data offline makes such points moot.

Speech-to-text: some details on the technicalities

Some details about the technicalities: uninterested readers may skip to the next section.

To transcribe these audio-messages I used OpenAI’s “Whisper” Automatic Speech Recognition model. For ease of use and consistency with my R-based workflow, I did this through the `audio.whisper` R package, which is a wrapper around the whisper.cpp C++ library. This may sound convoluted, but it is rather straightforward in practice, as it deals with many of that complexities that users of libraries based on machine-learning models often encounter. Especially for users with poweful GPUs, this may lead to significant inefficiency. For others, this should really be just about as efficient as running the original library if `audio.whisper` is installed with the right flags (see the package’s readme) and if the software is run naively, with no specific customisation or optimisations. Notice that `whisper` can process audio using one of a set of models ranging from “tiny” to “large”, with “tiny” being the quickest and least accurate and “large” the slowest but more accurate. Notice that if you use the original library, you have a decent GPU and set it up correctly, you can have *very* noticeable speed boosts, *if* the model fits into the VRAM of your GPU. Unless you really care about these things, even higher-end laptops mostly have GPUs with 4GB of VRAM or less, meaning you would probably be able to run only the “small” model (“large” requires ~10 GB of VRAM, see the project’s repository for details).

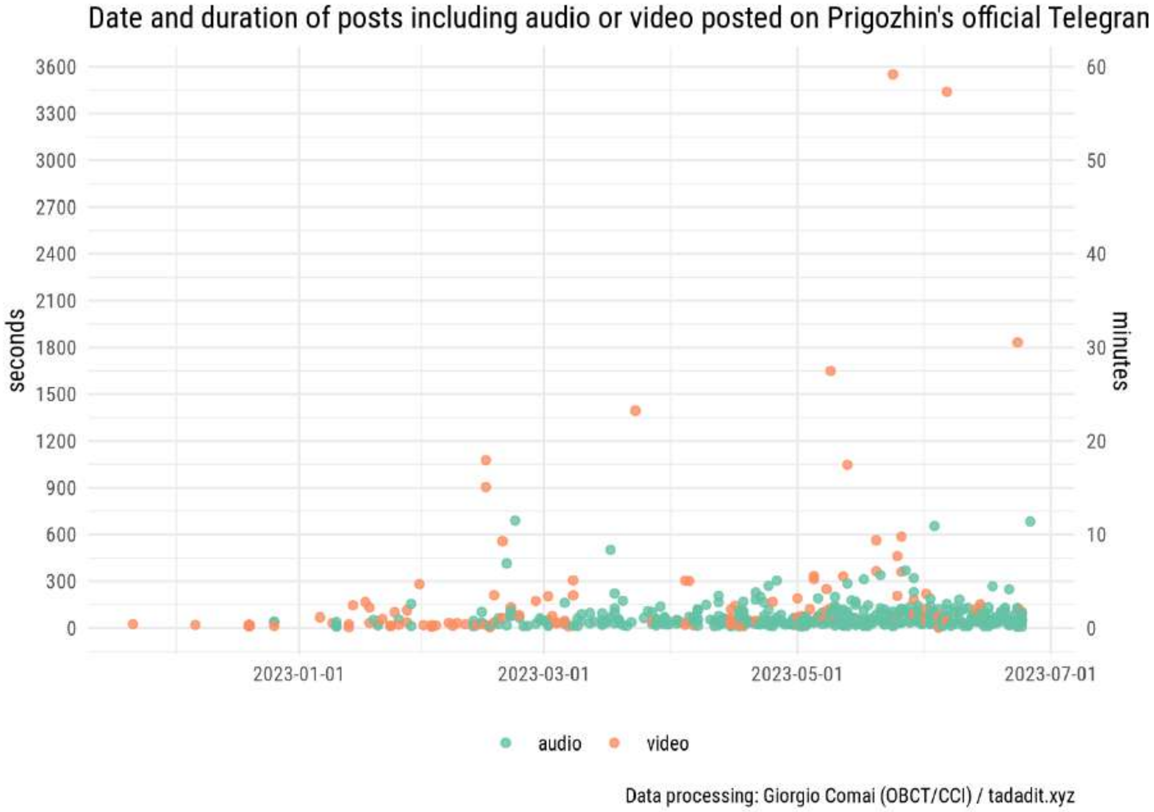
After some testing, I found that when the quality of the audio is very good, e.g. a TV news segment where words are spelled out clearly, even smaller models perform relatively well, and “medium” is already almost flawless. However, with audio messages filled with slang and often with less than ideal audio quality such as the one posted by Prigozhin, the “large” model seems to be really needed to get reasonably reliable results. It is however, time consuming, as transcribing a minute of audio using the `large` language model on a modern laptop takes a few minutes (it would be much quicker with enough GPU vRAM for the `large` model, but, again, most consumer laptops won’t have the right hardware); transcribing all of Prigozhin’s audio files implies quite a few days of processing.

Before proceeding with transcribing, here are some summary statistics about what we’re looking at.

Summary statistics about Prigozhin’s audio and video posts

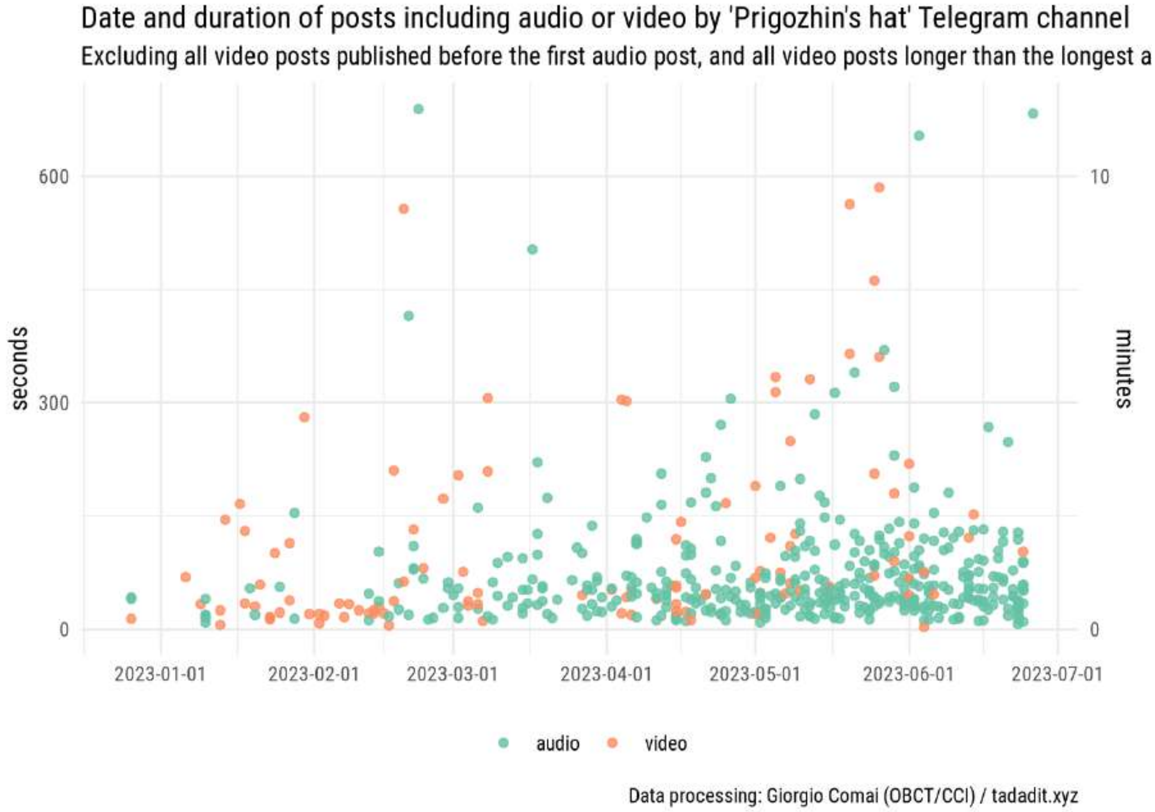
There are both audio and video messages that can be transcribed. Occasional video messages had been posted on the “Prigozhin’s hat” channel for a long time, but audio messages are a relative novelty, with the first one posted on 26 December 2022. In the rest of this post, I will focus on audio messages and Prigozhin’s official channel, as it includes all of the audio posts available.

If we plot the date when each post with either audio or video was posted, and the length of each of these post (i.e. its duration), it’s easy to notice that, a small number of video posts are quite long, up to almost one hour in length, but audio messages are almost invariably shorter than 10 minutes and mostly much shorter.



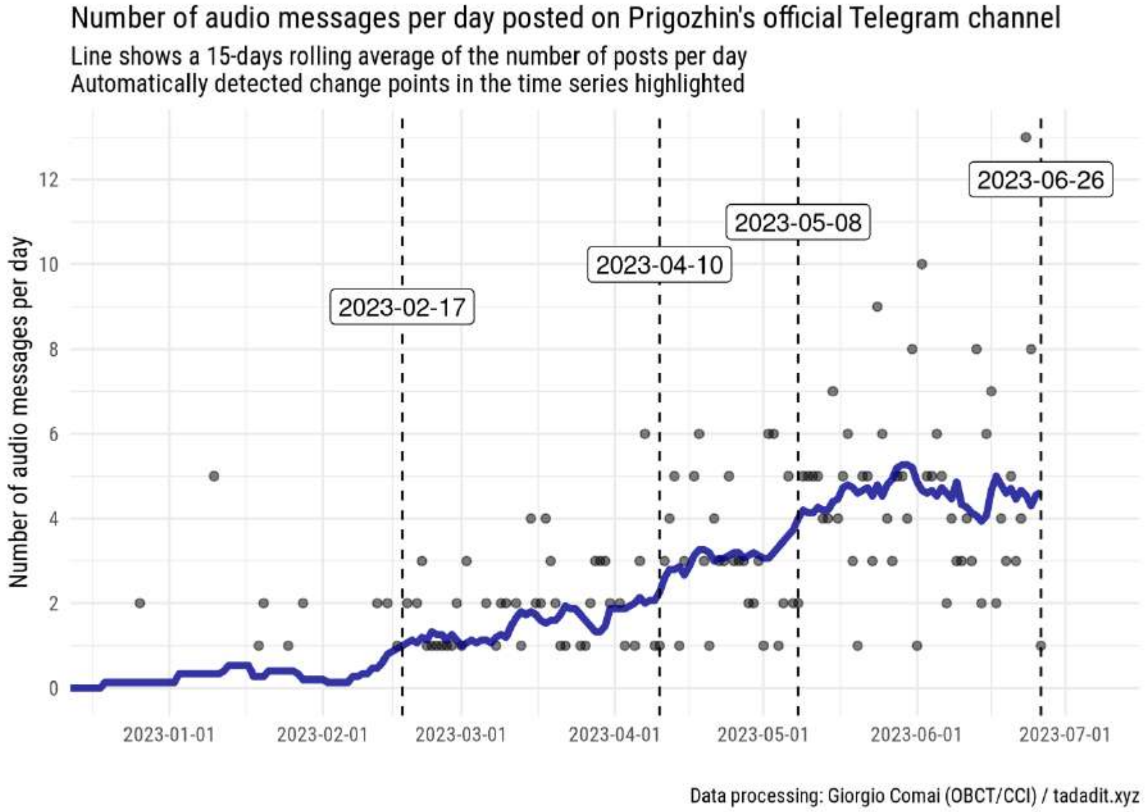
Let's plot this again, but zooming in on the graph setting the boundaries at the earliest audio message and the longest audio message to see things more clearly. In particular we notice that:

1. audio messages start being posted with some regularity around the second half of February 2023
2. they stop being posted all of a sudden on a specific day. You can easily spot Prigozhin's lengthy audio message on the top-right of the graph: Prigozhin's last message explaining and justifying the end of his mutiny, before effectively going into "radio silence".



Changes in the frequency of posting

Prigozhin's audio messages basically identify "Prigozhin's spring", from the "battle of Bakhmut" to his "march of justice". Indeed, just looking at the plain number of audio messages posted each day and using a dedicated function to automatically identify change points in time series, we obtain a visual depiction as well as some key dates marking the escalation.



Number of audio messages per day posted on Prigozhin's official Telegram channel

Line shows a 15-days rolling average of the number of posts per day
Automatically detected change points in the time series highlighted

Data processing: Giorgio Comai (OBCT/CCI) / tadadit.xyz

Without even looking at the contents of these messages, we can tentatively look at these dates and mark some phases in this escalation:

1. (before the beginning of this graph) since the day Prigozhin admitted he's Wagner's owner and leader on 26 September 2022 to 26 December 2022, when his first audio message was posted
2. between then and 19 February 2023, when audio messages were only occasional, but his criticism of the Ministry of Defence (MoD) became more open
3. Since late February, audio messages became more frequent; on 6 March, Prigozhin denounced explicitly the behaviour of the MoD as betrayal, but did not mention by name either Shoigu or Gerasimov
4. Starting with April, it became routine for Prigozhin to post three audio messages per day or more
5. On 5 May, in a video of himself surrounded by dead bodies, Prigozhin threatened to pull out of Bakhmut. In the weeks that followed, in particular after 9 May (the highly celebrated Victory Day in Russia) his denunciations became more vocal and his audio messages more frequent, averaging over 4 messages per day; it is only at this point that Prigozhin started to openly attack Shoigu and Gerasimov by name

6. The high frequency of posts broadly remained in place until the mutiny on 26 June 2023, then they stopped abruptly.

For reference, see the Moscow Times timeline of Prigozhin's standoff with the Ministry of Defence.

The contents: Prigozhin's audio messages, transcribed

I include below in tabular format a transcription of all audio messages posted by Prigozhin, first in English, then in Russian. Each line includes a timestamp and a link to the original message. The table can be used for quick searches.

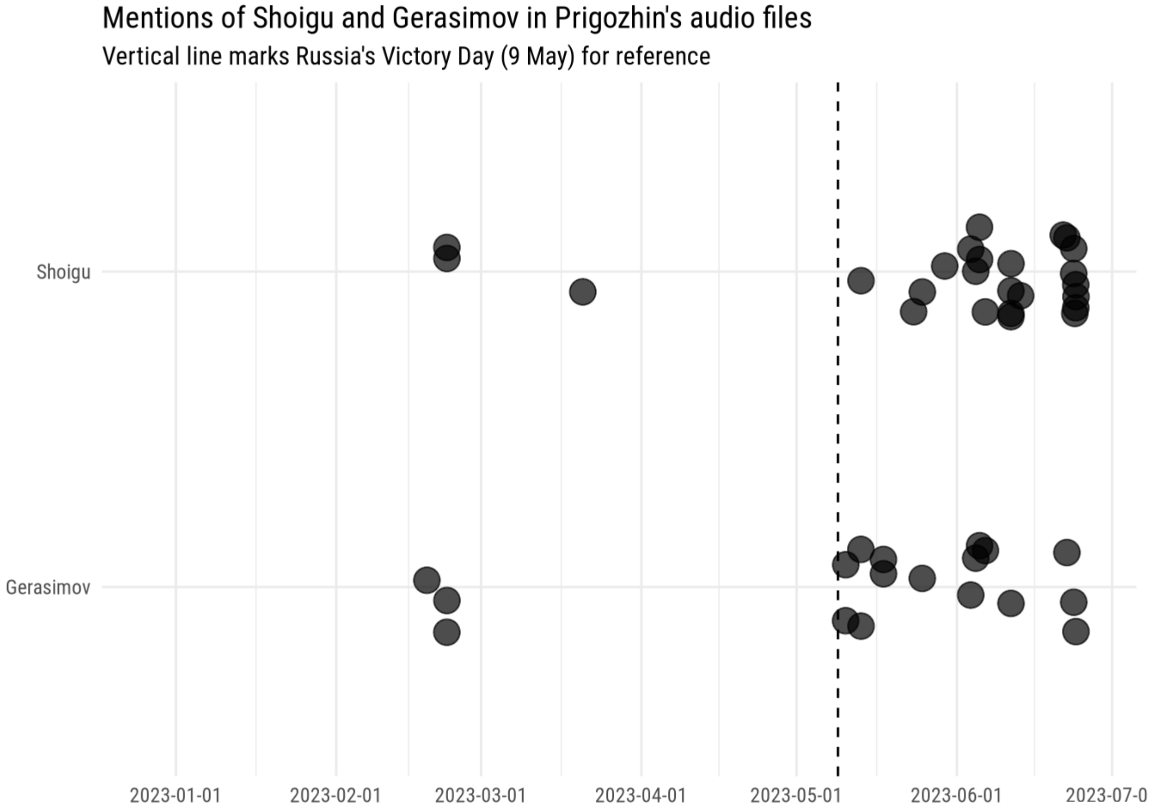
However, for convenience, I am sharing all of the transcribed audio files also as a downloadable dataset, and in a dedicated page where they can be conveniently browsed.

In both cases, be mindful of the errors in both transcription and translation.

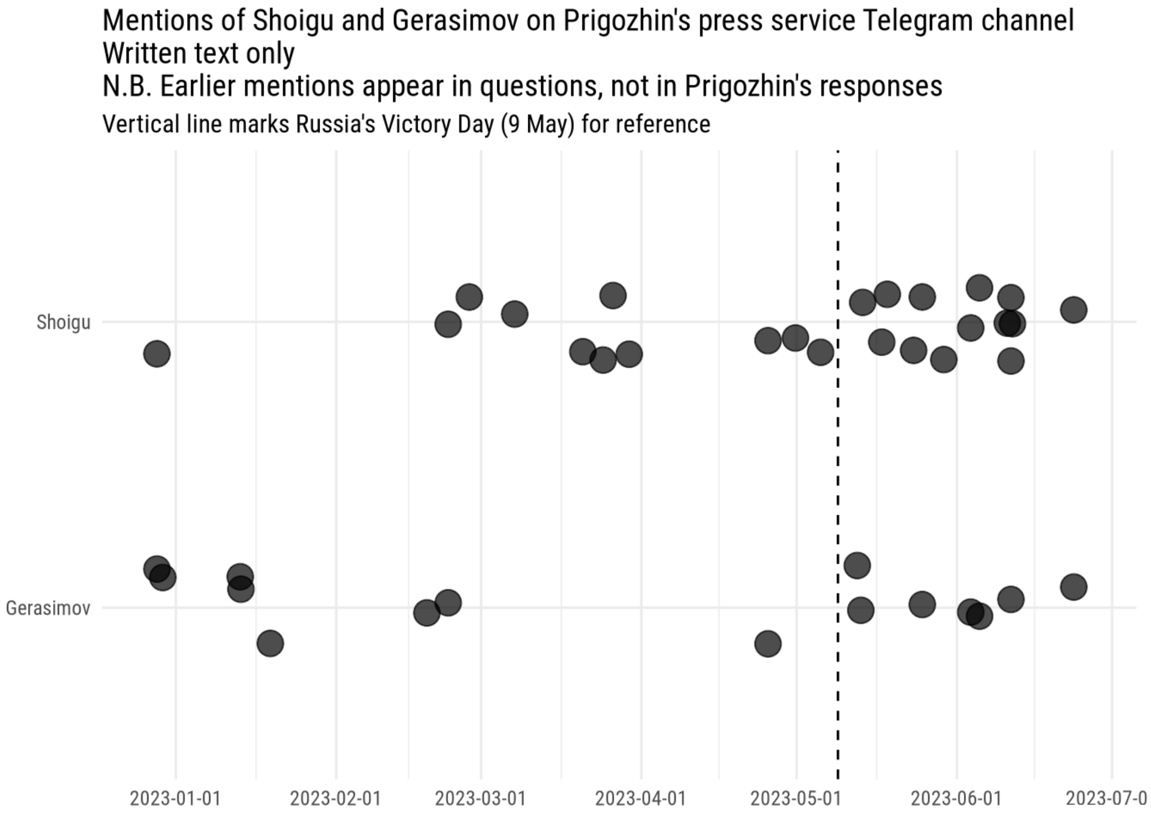
- full dataset with all of Prigozhin's audio messages transcribed is available for download
- all audio files transcribed in Russian and available in a single page
- all audio files transcribed and translated in English and available in a single page (automatic translation, more error-prone)

> ℹ Note
>
> **N.B.** This is a static preview. See the online version for exploring the data.

For reference, here is a graph showing mentions of Minister of Defence Shoigu and and Chief of Staff of the armed forces Gerasimov. It clearly shows how Prigozhin started making their name only after 9 May (Victory Day in Russia). As shown in the graphs above, at the same time, the average number of daily audio messages has also increased. Explicitly mentioning Shoigu and Gerasimov was a clear sign of escalation from Prigozhin's side. The fact that he was not rebuffed or criticised by the Kremlin at the time was probably instrumental in opening the way for the further escalation that led to Prigozhin's mutiny on 26 June.



It is worth adding that these messages report Prigozhin's own words. Text included in the Telegram posts themselves often included both answer and reply, and quite often only the question asked. As it appears from the following graph pointing at mentions of Gerasimov and Shoigu in the text messages of Telegram posts, it easy to notice more earlier mentions. However, they all appear in the questions included in the post: even if journalists were actively asking about Shoigu and Gerasimov, Prigozhin himself never mentioned their name in his responses.



A separate post with further analysis of the contents may follow.

Conclusions

Audio messages, in the case of Prigozhin, and video messages in the case of many other influential voices commenting the war for local audiences in Russia have become an important part of the public conversation about the war, including topics and perspectives that do not appear on traditional media and may be scarcely mentioned in written text.

This post (full code available on this website's repository) demonstrated how audio messages - including audio messages in Russian, with often less than ideal audio quality and frequent use of slang - can be turned into written text using freely available tools. The speech-to-text process can be run fully offline without relying on third parties and is fully reproducible. It takes however a considerable amount of processing time even on relatively powerful copmuters: if many hours of audio need to be processed, then this is probably not a viable solution at this stage and other options (mostly, relying on commercial vendors) should probably be preferred.

The quality of the transcribed text even using *Whisper*'s `large` model is not perfect, but is of surprising accuracy. The text thus generated can easily be parsed to look for specific information or for quantitative analysis, as long as the some issues are kept in consideration, in particular varying spelling for names of people and places (e.g. Shoigu can also be Shaigu). When looking for specific patterns this is mostly rather easy to work around after some tentative exploration.

The quality of the text transcribed and translated in a single step (a feature offered by *Whisper*) is not as good, and includes some more errors and inaccuracies. Indeed, you may prefer going through Prigozhin's transcribed messages relying on the Russian version and running it through tools such as Google Translate rather than the automatically transcribed and translated version resulting from *Whisper*. Yet, the quality is really not that bad, and text can mostly be read and processed, even if being mindful of the fact that there are issues with the text.

Indeed, I recommend reading the transcription of Prigozhin's audio messages during the mutiny and the weeks that preceded it to see just how unhinged his criticism of Shoigu and Gerasimov had become; if one is allowed to criticise so very publicly the leadership of the army without being rebuffed, he may be (figuratively) excused from thinking that the country's leadership is really on his side.

As multi-media and multi-format contents become more central to public discourse in Russia and elsewhere, and online spaces where spoken words may be more prominent than written words become more important for public discourse, researchers should expand their analytical toolbox to better account for such sources, even if just in order to define subsets of materials to be analysed qualitatively or explored through complementary approaches.

Russophobia in Russian official statements and media

Originally published on tadadit.xyz on 26 September 2023

https://tadadit.xyz/posts/2023-09-russophobia/

Context

References to 'anti-Russian sentiments' or 'Russophobia' - have a long history that dates back to the 19th century (Feklyunina 2012; Darczewska e Żochowski 2015). However, in recent years references to the alleged spread of 'Russophobia' in the West have apparently become more common and more politically consequential (Darczewska e Żochowski 2015). A quantitative analysis of references to 'Russophobia' in statements by Russia's Ministry of foreign affairs has confirmed that the expression was barely used before 2012, but featured much more often in official statements in particular since 2014 (Robinson 2019).

In this post, I will provide a brief overview of the frequency of references to 'Russophobia' or 'anti-Russian' sentiments in official statements and press releases issued by the Kremlin, the Russian Ministry of Foreign Affairs, as well as in the textual version of news segments published by Russia's first channel (*Pervy Kanal*).

In order to let the reader gauge the tone of such references in context, I will also include tables with the five words preceding and following the reference to 'Russophobia' as well as a link to the specific occasion where the reference was found. I have previously discussed the usefulness of this approach in a dedicated article (G. Comai 2017).

The graphs included in this post are based on the absolute number of matches of relevant keywords, not their relative frequency as a share of total word count, even if the number of publications is not constant throughout the period under analysis. Additional graphs with the average number of publications per day for each source are included for reference. Analytically, relative word-count would not lead to significantly different conclusions.

Preliminary quality checks have not raised major inconsistencies or problems in the data collection process. Some issues possibly due to the way contents are archived by the sources (e.g. occasional duplicate posts) cannot however be completely excluded until more thorough checks are conducted.

I have published the textual dataset based on the Kremlin's website on Discuss Data. You can download the full dataset in different formats from there, and find a detailed note on how it was created. A more updated (even if not fully formalised) version of the datasets of the Kremlin (English and Russian version) and the Russian MFA (English and Russian version) are available for download.

Finally, here are links to a fully interactive version of the datasets mentioned in this post, allowing researchers to test the frequency of alternative expressions that may be used to express similar meaning. I have not yet finalised building the interface, but it should be sufficient for basic data exploration.

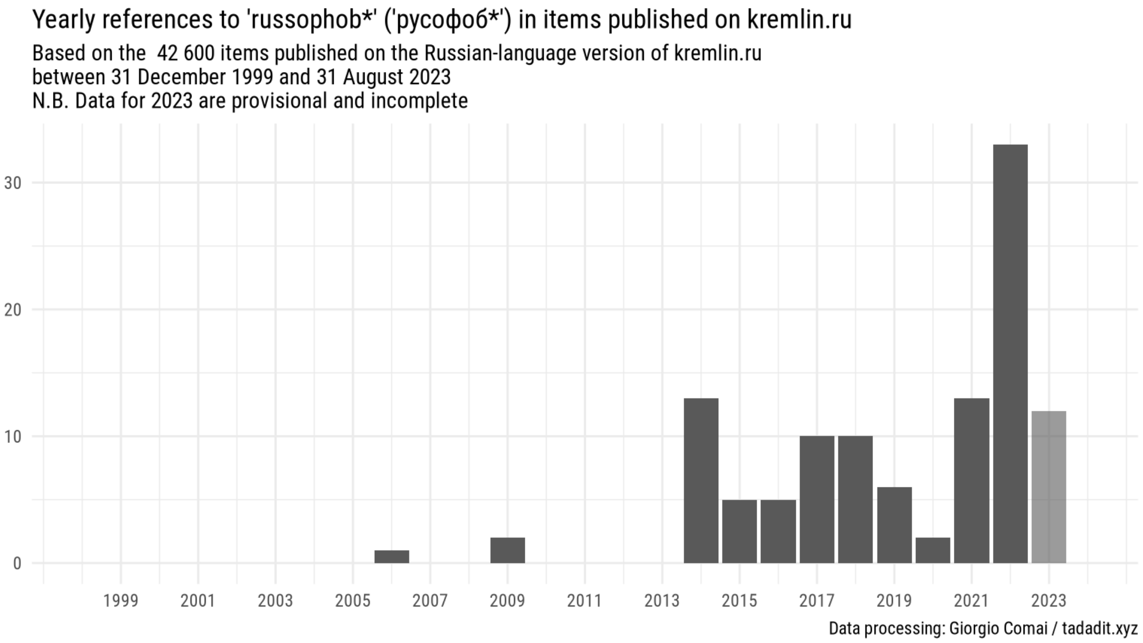
- Kremlin, Russian version
- Kremlin, English version
- Russian MFA, Russian version
- Russian MFA, English version

Preliminary observations

There are only occasional references to "anti-Russian" sentiments and "Russophobia" in official statements and in news segments aired by *Pervy Kanal* before 2014. Such references however have become much more common starting with 2014, after Russia's annexation of Crimea, and even more in 2022, with Russia's invasion of Ukraine. The trend is particularly noticeable in statements by the Russian Ministry of Foreign Affairs: over one thousand mentions of either "russophob*" or "anti-Russian" have been recorded in a single year.

The Kremlin

References to "Russophobia" were basically not to be found before 2014. Their sudden appearance corresponds to Russia's annexation of Crimea, but it is with Russia's invasion of Ukraine that the number of mentions increases even more substantially.

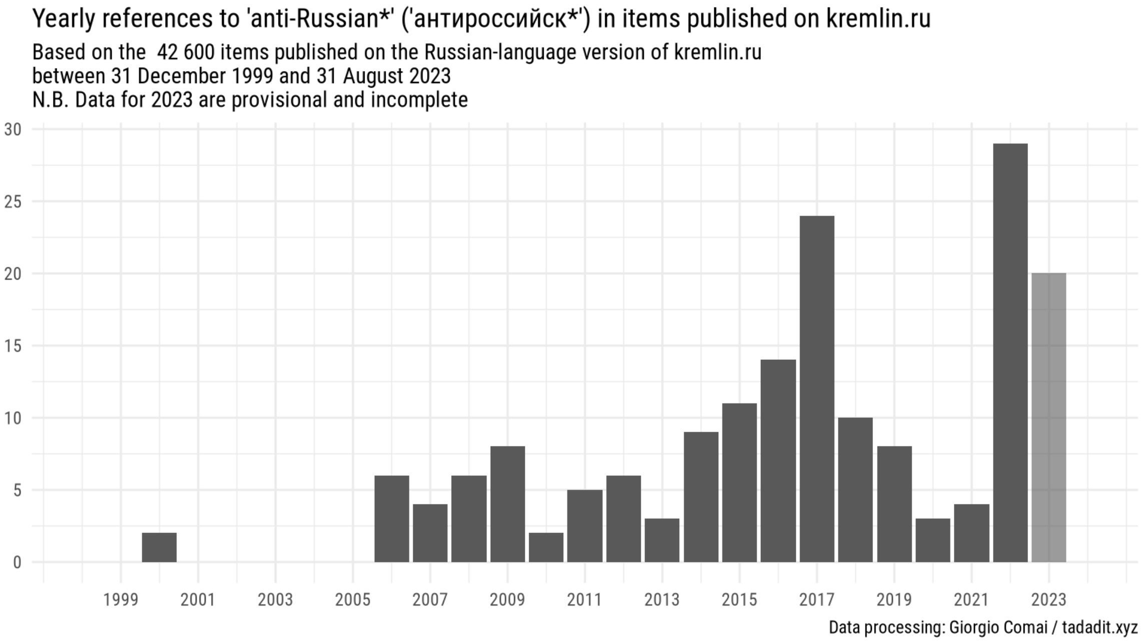


Yearly references to 'russophob*' ('русофоб*') in items published on kremlin.ru

Based on the 42 600 items published on the Russian-language version of kremlin.ru
between 31 December 1999 and 31 August 2023
N.B. Data for 2023 are provisional and incomplete

Data processing: Giorgio Comai / tadadit.xyz

References to "anti-Russian", however, started to appear earlier, in 2009. The reader should keep in mind that between 2008 and 2012 Dmitri Medvedev, not Vladimir Putin, was president of the Russian Federation, which likely had an impact on word choices.

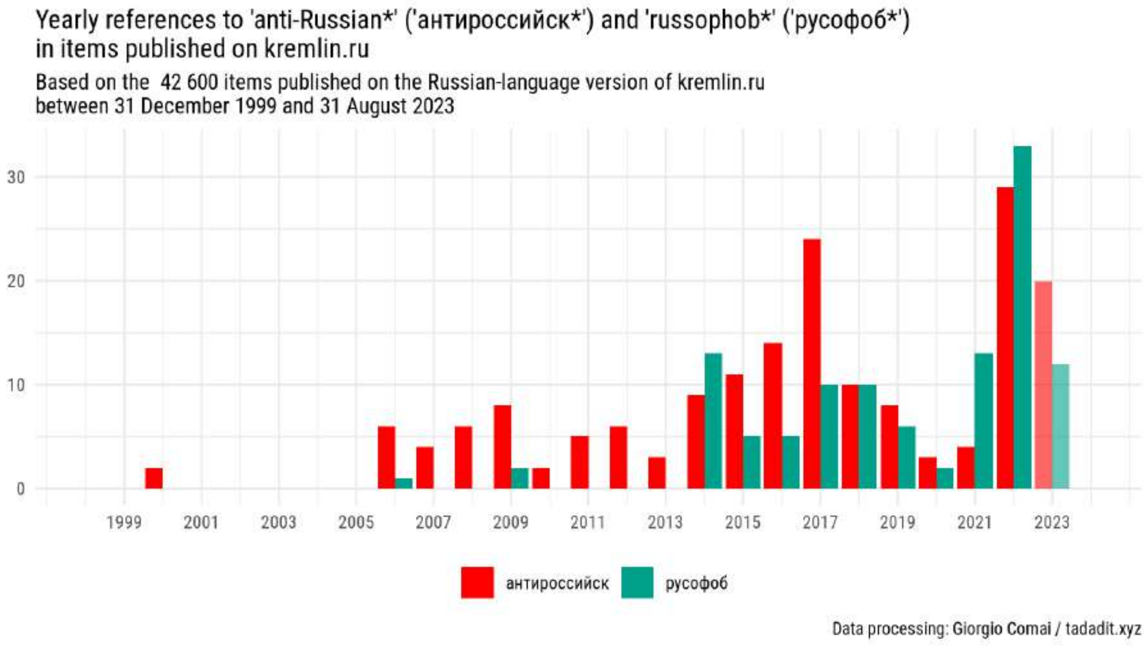


Yearly references to 'anti-Russian*' ('антироссийск*') in items published on kremlin.ru

Based on the 42 600 items published on the Russian-language version of kremlin.ru
between 31 December 1999 and 31 August 2023
N.B. Data for 2023 are provisional and incomplete

Data processing: Giorgio Comai / tadadit.xyz

The following graph shows both "anti-russian" and "Russophobia" with the same scale.



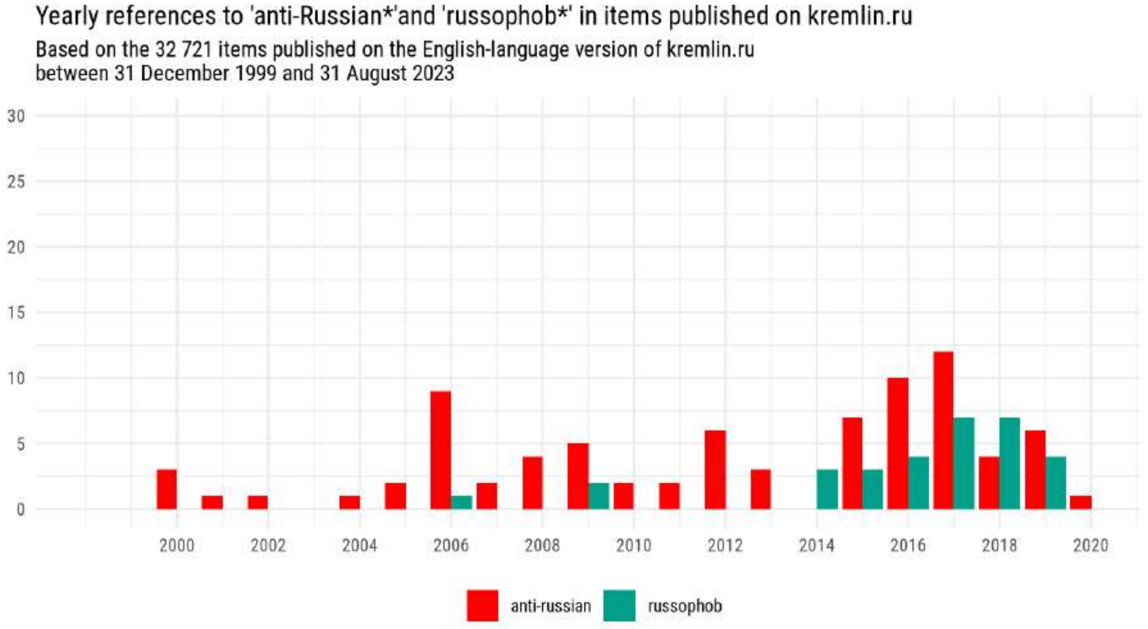
Yearly references to 'anti-Russian*' ('антироссийск*') and 'russophob*' ('русофоб*') in items published on kremlin.ru

Based on the 42 600 items published on the Russian-language version of kremlin.ru between 31 December 1999 and 31 August 2023

Data processing: Giorgio Comai / tadadit.xyz

Kremlin.ru, English version

I include for reference a version of the same graph, based on the English language version of Kremlin.ru, which has fewer contents available (not all items are translated in English).

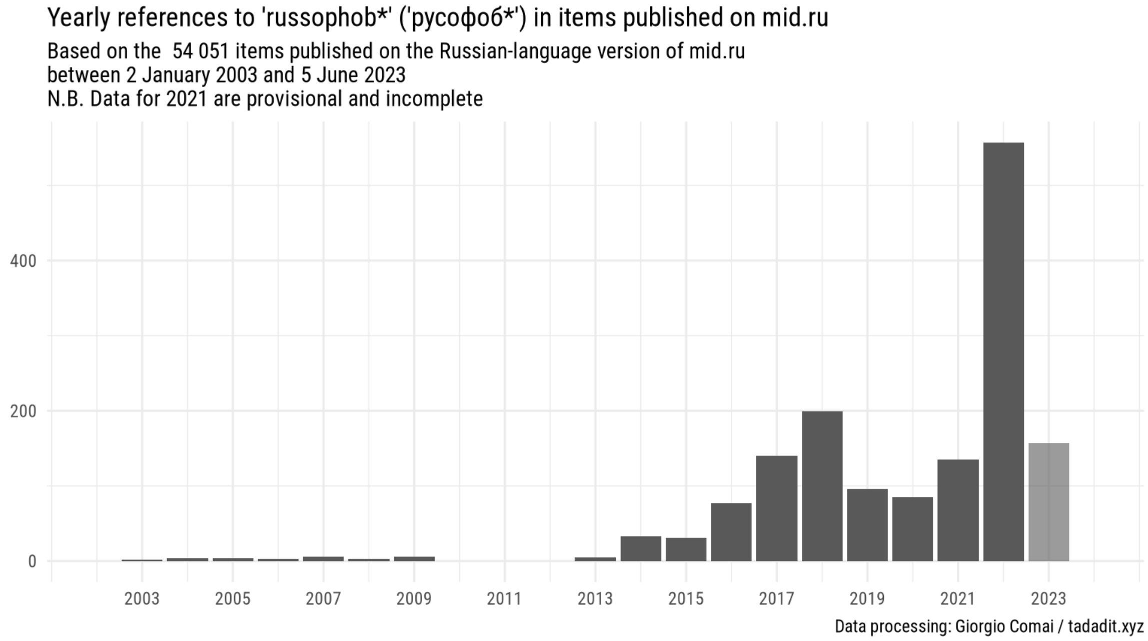


Yearly references to 'anti-Russian*'and 'russophob*' in items published on kremlin.ru

Based on the 32 721 items published on the English-language version of kremlin.ru between 31 December 1999 and 31 August 2023

The Russian Ministry of foreign affairs

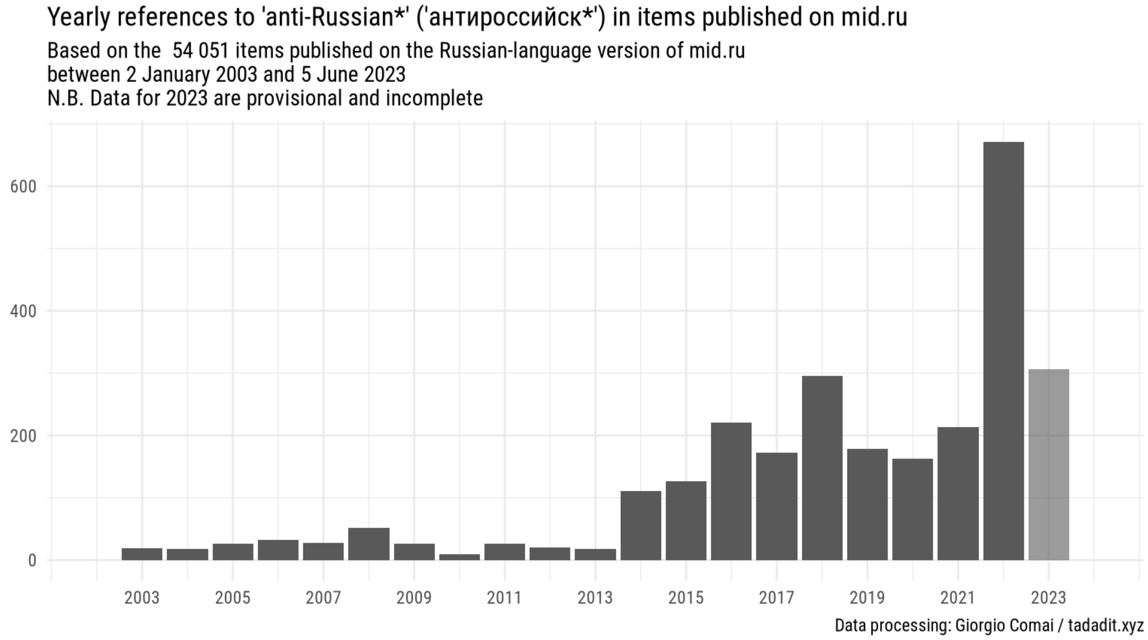
This section presents the same graphs as above, first for the Russian language version of mid.ru, then for the English language version. It emerges how the number of references to both “Russophobia” and “anti-Russian” increased very substantially in absolute terms, with over one thousands mentions of these expressions recorded in a single year.



Yearly references to 'russophob*' ('русофоб*') in items published on mid.ru

Based on the 54 051 items published on the Russian-language version of mid.ru
between 2 January 2003 and 5 June 2023
N.B. Data for 2021 are provisional and incomplete

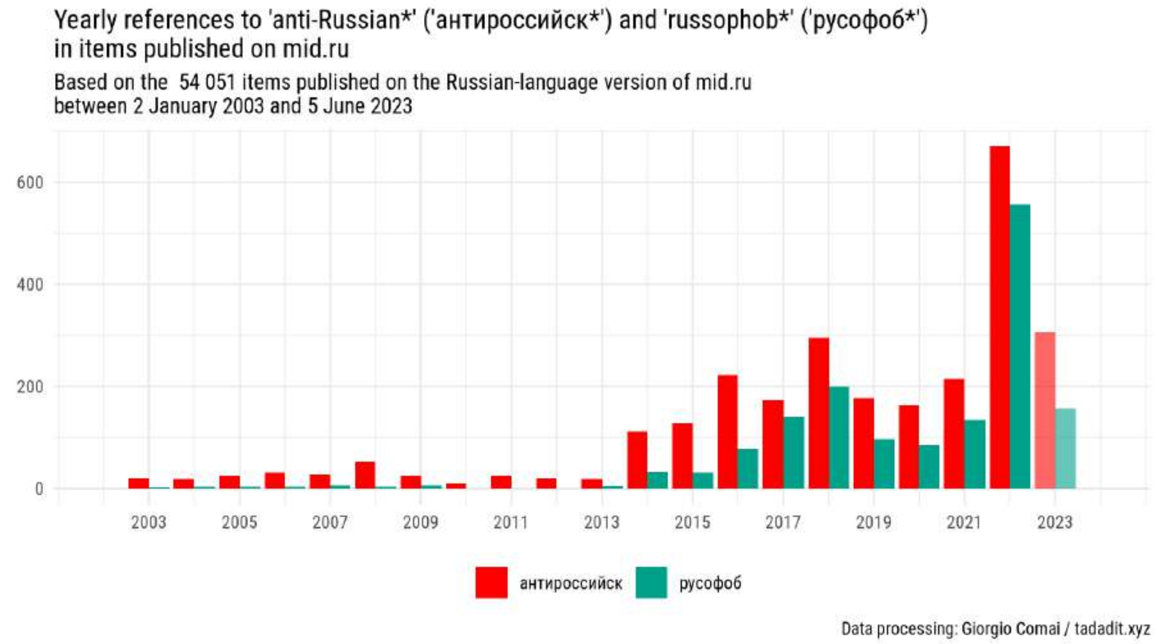
Data processing: Giorgio Comai / tadadit.xyz



Yearly references to 'anti-Russian*' ('антироссийск*') in items published on mid.ru

Based on the 54 051 items published on the Russian-language version of mid.ru
between 2 January 2003 and 5 June 2023
N.B. Data for 2023 are provisional and incomplete

Data processing: Giorgio Comai / tadadit.xyz

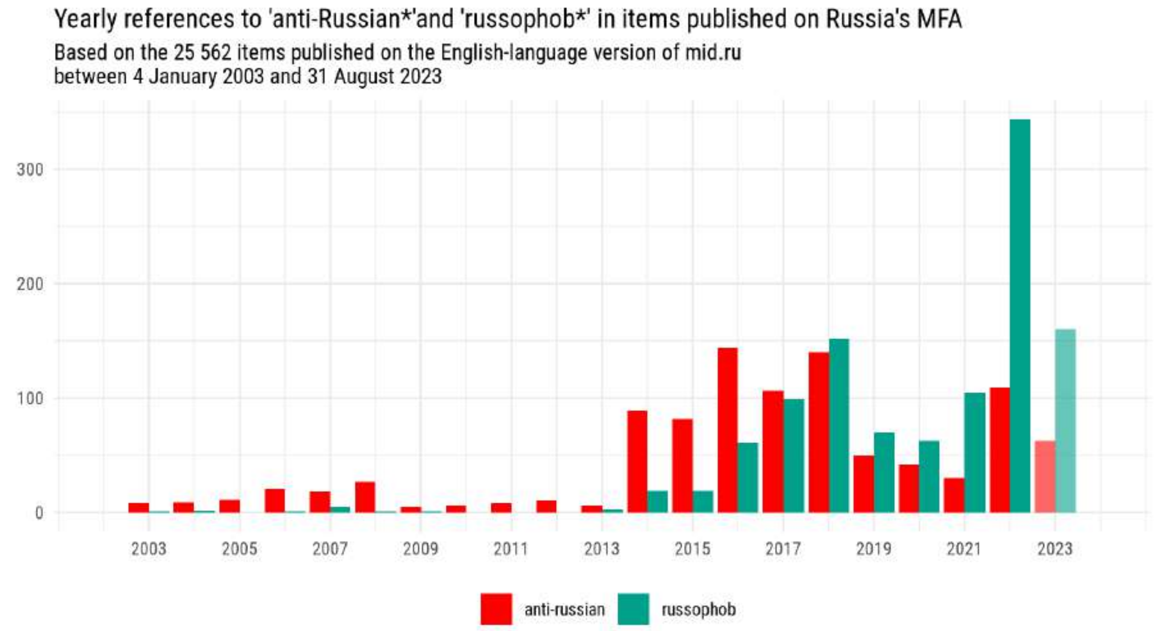


Yearly references to 'anti-Russian*' ('антироссийск*') and 'russophob*' ('русофоб*') in items published on mid.ru

Based on the 54 051 items published on the Russian-language version of mid.ru between 2 January 2003 and 5 June 2023

Data processing: Giorgio Comai / tadadit.xyz

Russian MFA, English version

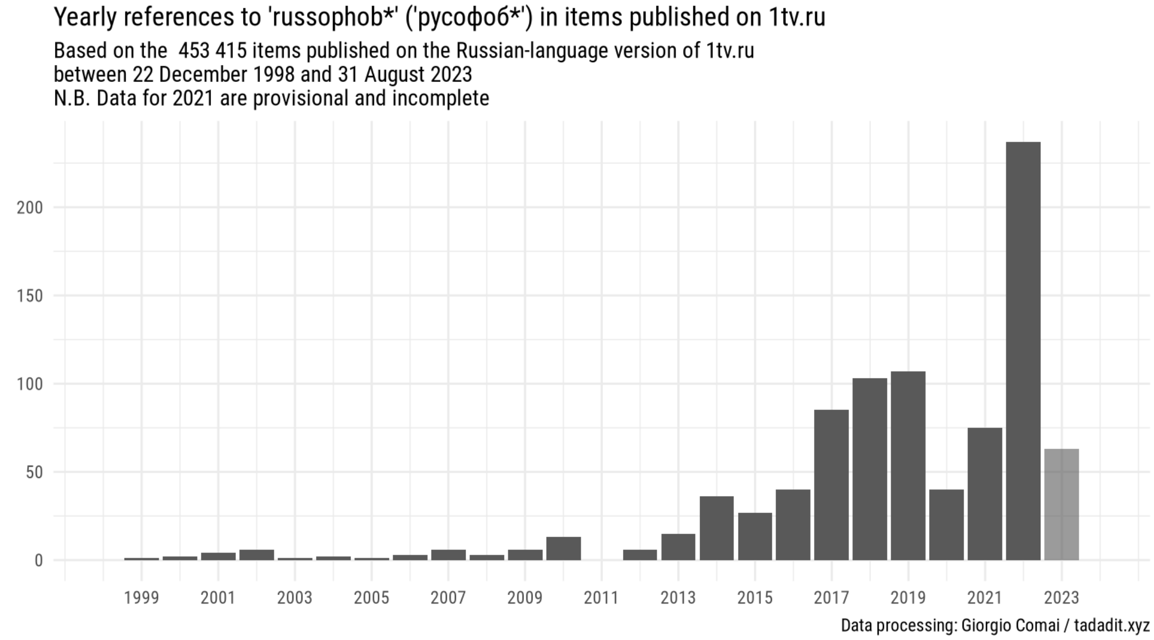


Yearly references to 'anti-Russian*'and 'russophob*' in items published on Russia's MFA

Based on the 25 562 items published on the English-language version of mid.ru between 4 January 2003 and 31 August 2023

Russia’s First Channel - Pervy Kanal

This dataset is based on all news items published on the websites of Russia’s first channel: Pervy Kanal. It is a much larger dataset, including 453 415 items published between 22 December 1998 and 31 August 2023. It mostly does not include full transcripts of talk shows, but rather transcripts of news segments. Starting in 2022, only a summary of news segments has been included; if full transcripts were included as in previous years, the number of recorded mentions would likely be much higher.



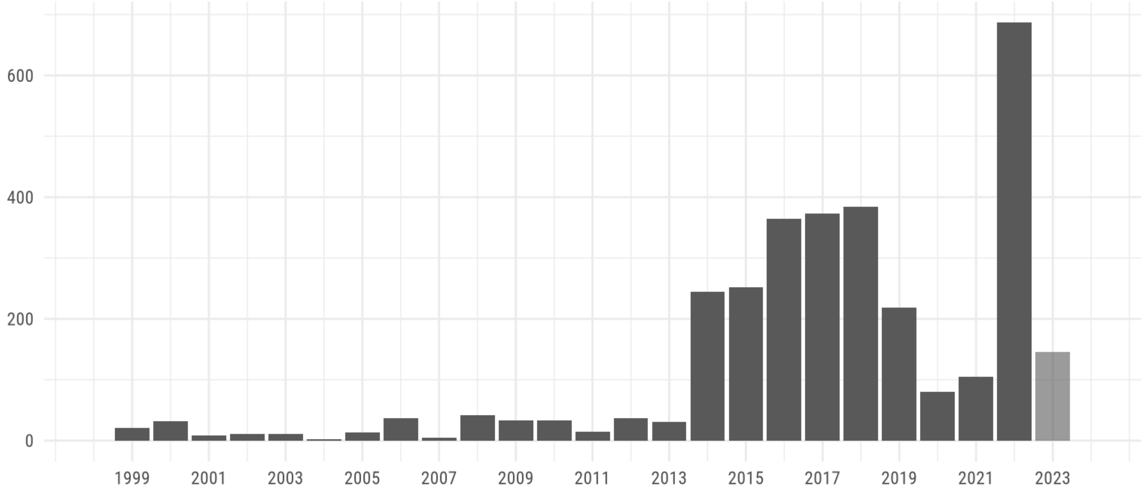
Yearly references to 'russophob*' ('русофоб*') in items published on 1tv.ru

Based on the 453 415 items published on the Russian-language version of 1tv.ru
between 22 December 1998 and 31 August 2023
N.B. Data for 2021 are provisional and incomplete

Data processing: Giorgio Comai / tadadit.xyz

Yearly references to 'anti-Russian*' ('антироссийск*') in items published on 1tv.ru

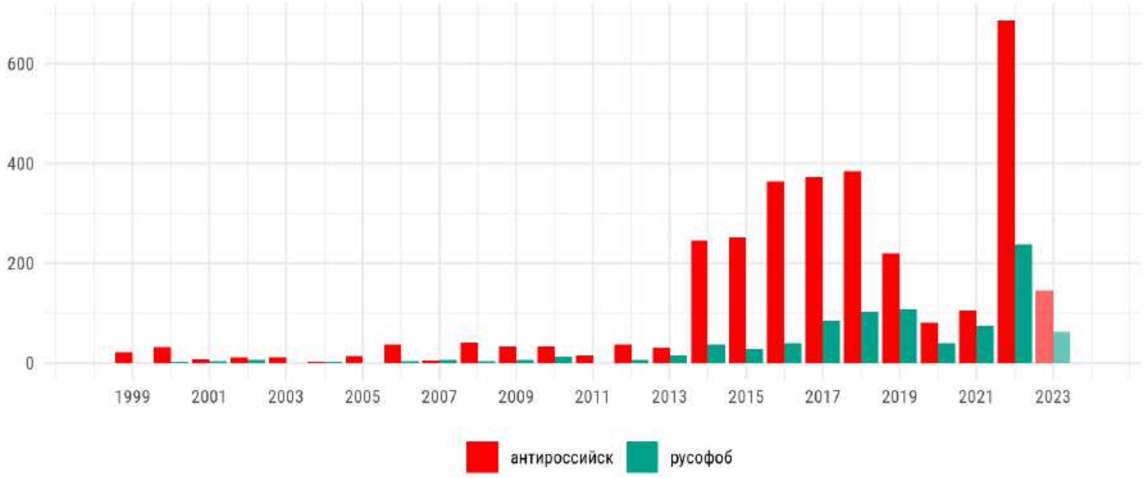
Based on the 453 415 items published on the Russian-language version of 1tv.ru
between 22 December 1998 and 31 August 2023
N.B. Data for 2021 are provisional and incomplete



Data processing: Giorgio Comai / tadadit.xyz

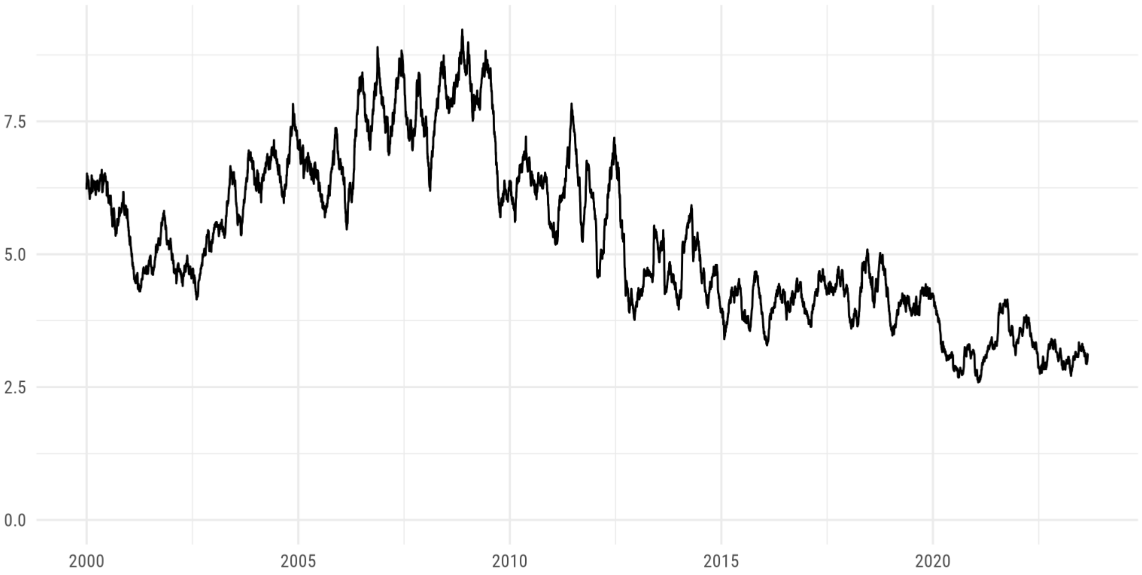
Yearly references to 'anti-Russian*' ('антироссийск*') and 'russophob*' ('русофоб*')
in items published on 1tv.ru

Based on the 453 415 items published on the Russian-language version of 1tv.ru
between 22 December 1998 and 31 August 2023



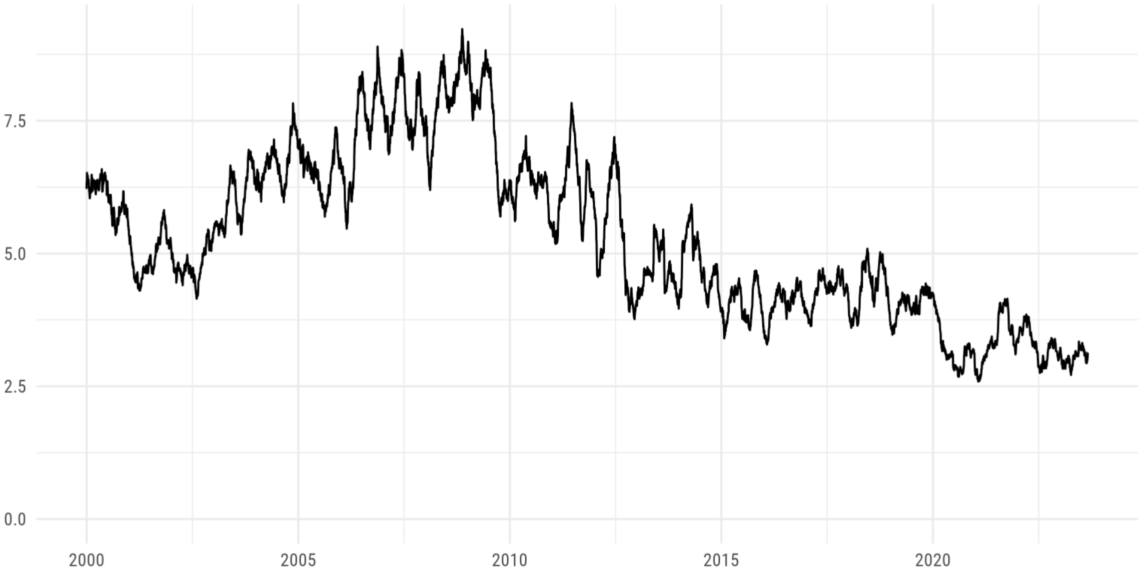
Data processing: Giorgio Comai / tadadit.xyz

Frequency of publications in each of the sources



Number of publications per day on the Russian language version of Kremlin.ru

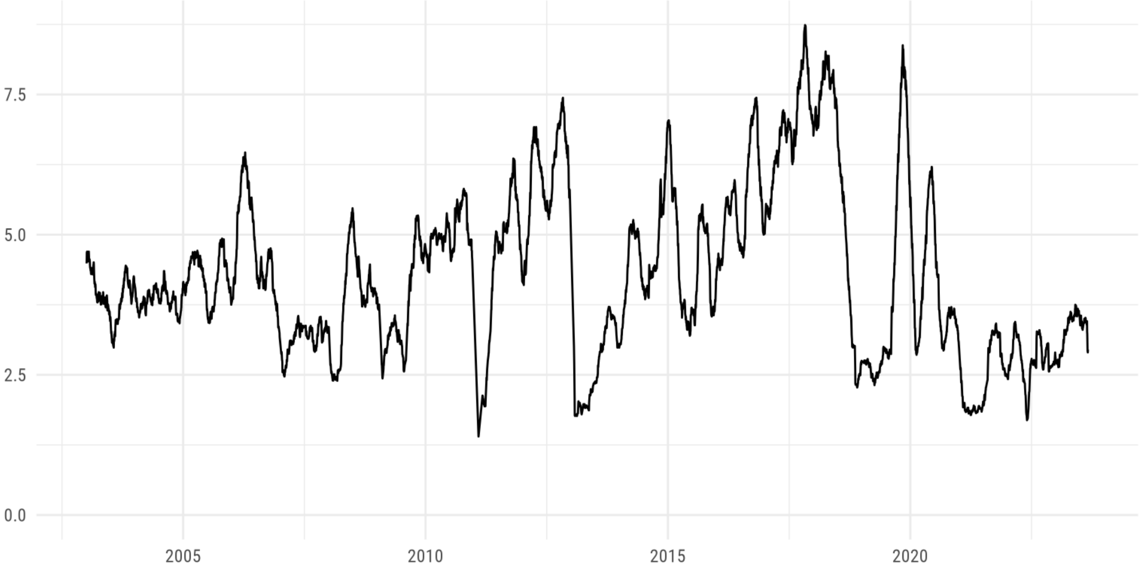
* Calculated on a rolling mean of 91 days



Number of publications per day on the Russian language version of Kremlin.ru

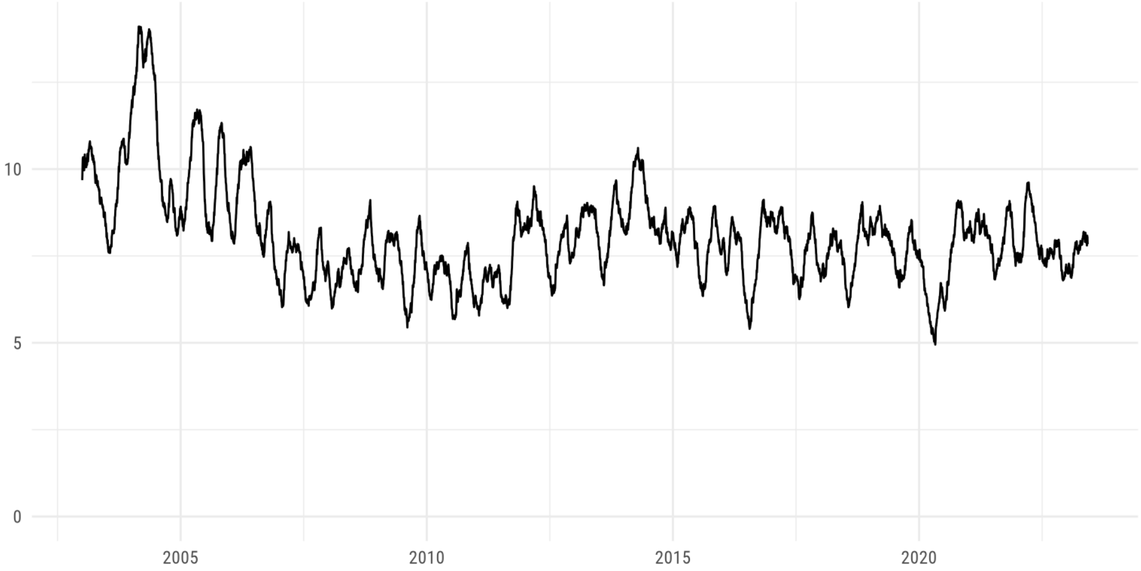
* Calculated on a rolling mean of 91 days

Number of publications per day on the English language version of mid.ru



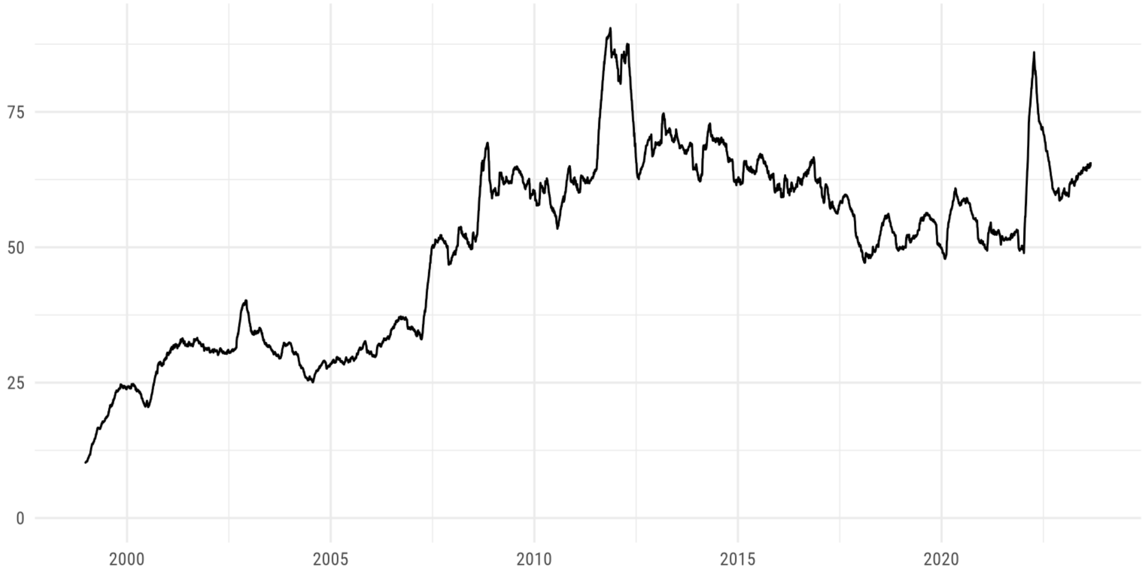
* Calculated on a rolling mean of 91 days

Number of publications per day on the Russian language version of mid.ru



* Calculated on a rolling mean of 91 days

Number of publications per day on 1tv.ru



* Calculated on a rolling mean of 91 days

Extracting textual contents from the Kremlin's website with `castarter`

This is a tutorial demonstrating how to extract textual contents with metadata from the Kremlin's website using `castarter`, a package for the R programming language.

As this is an introductory post, all steps and functions are explained in detail.

Step 1: Install `castarter`

This tutorial assumes some familiarity with the R programming language. At the most basic, you should have installed R and an IDE such as Rstudio, start a new project, and then you can just copy/paste and run commands in a scripts and things should work. Once your setup is in place, make sure you install `castarter`.

```r
# install.packages("remotes")
remotes::install_github("giocomai/castarter")
```

Step 2: Think of file and folder locations

Everyone organises stuff on their computer in their own way. But when getting ready to retrieve hundreds of thousands of pages, some consideration should be given to organising file in a consistent manner.

With `castarter`, by default, everything is stored in the current working directory, but a common pattern for `castarter` users will be to keep the file with their R script in a location, often a synced folder, and then to have all the html pages stored for text mining in a non-synced folder. This is because it is common to get hundreds of thousands of pages even for relatively small projects, and storing very large number of small files slows down sync clients for the likes of Dropbox, Google Drive, and Nextcloud. If you rely on git-based approaches, you will need to add the data folders to your `.gitignore`.

The following setup will store all files in a subfolder of your R installation folder; set it to whatever works for you in the `base_folder` parameter.

The following code sets a few option for the current session. You will typically want to have something like this at the beginning of each `castarter` script.

This sets:

- a `base_folder` - everything will happen starting from there
- a `project` - which will generate a folder within the `base_folder`. In this case, I'll set this to "Russian institutions", as I plan to store here a number of textual datasets from relevant websites, including the website of the Russian president, the Ministry of defence, and the Ministry of Foreign affairs.
- a `website` - you are free to call it as suits you best, but I have grown accustomed to use the bare domain, an underscore, and then the main language, so in this case it will be `kremlin.ru_en` (so that it's easy to differentiate from the Russian language version, which would be `kremlin.ru_ru`). This is however just a convention, and anything will work.

```r
library("castarter")

cas_set_options(
  base_folder = fs::path(fs::path_home_r(),
                         "R",
                         "castarter_tadadit"),
  project = "Russian institutions",
  website = "kremlin.ru_en"
)
```

By default, `castarter` stores information about the links and downloaded files in a local database, so that it will always be possible to determine when a given page was downloaded, or where a given link comes from. By default, this is stored in a `SQLite` database in the website folder.

Each time a file is downloaded or processed, by default relevant information is stored in the local database for reference.

Step 3: Get the urls of index pages

Conceptually, `castarter` is based on the idea common to many text mining projects that there are really two types of pages:

- **index pages**: they are, e.g, lists to posts or articles, the kind of pages you ofen see if you click on "See all news". Sitemap files can also be understood as index pages in this context. The key part is that these pages are mostly dynamic, and we mostly care about them because they include links to the pages we are actually interested in.
- **content pages**: these are the pages we mostly actually care about. They have unique urls, and their content is mostly expected to remain unchanged after publication.

In the case of the Kremlin's website, we can quickly figure out that index pages have these urls:

- http://en.kremlin.ru/events/president/news/page/1 (the latest posts published)
- http://en.kremlin.ru/events/president/news/page/2 (previous posts)
- http://en.kremlin.ru/events/president/news/page/3 (etc.)
- …

Incidentally, I'll also point out that there's a version of the website aimed to the visually impaired:

http://en.special.kremlin.ru/events/president/news

We will stick to the standard version in this tutorial to keep things interesting, but it is worth remembering that websites may have alternative versions, e.g. a mobile-focused version, that have less clutter and may be easier to parse. For example, from that version of the website it is easier to see how many index pages there are.

At the time of writing, there are 1350 index pages for the news section of the website. This is how those urls look:

```r
cas_build_urls(url = "http://en.kremlin.ru/events/president/news/page/",
               start_page = 1,
               end_page = 1350)
```

```
# A tibble: 1,350 x 3
      id url                                                     index_group
   <dbl> <chr>                                                   <chr>
```

```
1      1 http://en.kremlin.ru/events/president/news/page/1  index
2      2 http://en.kremlin.ru/events/president/news/page/2  index
3      3 http://en.kremlin.ru/events/president/news/page/3  index
4      4 http://en.kremlin.ru/events/president/news/page/4  index
5      5 http://en.kremlin.ru/events/president/news/page/5  index
6      6 http://en.kremlin.ru/events/president/news/page/6  index
7      7 http://en.kremlin.ru/events/president/news/page/7  index
8      8 http://en.kremlin.ru/events/president/news/page/8  index
9      9 http://en.kremlin.ru/events/president/news/page/9  index
10    10 http://en.kremlin.ru/events/president/news/page/10 index
# i 1,340 more rows
```

On many websites, the main "news" section will include links to all relevant articles. In our case, we see that the Kremlin has also a few more sections:

- News
- Speeches and transcripts
- Presidential Executive Office
- State Council
- Security Council
- Commissions and Councils

As it turns out, "Speeches and transcripts" seem to be mostly included in the news feed, but e.g. items from the "Commissions and councils" are not. Less we miss some materials, we may want to get all of these. Or, depending on the research we're working on, we may be interested only in "Speeches and transcripts". Either way, to distinguish among different feeds and facilitate future steps, including possibly automatic updating of the procedure, `castarter` includes an `index_group` parameter.

So we can build urls separated by `index_group`, and store them in our local database.

```
cas_build_urls(url = "http://en.kremlin.ru/events/president/news/page/",
               start_page = 1,
               end_page = 1350,
               index_group = "news") |>
  cas_write_db_index()
```

And we can then proceed and add links from other categories if we so wish. We do not need to worry about running this script again: only new urls will be added by default.

```
cas_build_urls(url = "http://en.kremlin.ru/events/president/transcripts/page/",
               start_page = 1,
               end_page = 470,
               index_group = "transcripts") |>
  cas_write_db_index()

cas_build_urls(url = "http://en.kremlin.ru/events/administration/page/",
               start_page = 1,
               end_page = 57,
               index_group = "administration") |>
  cas_write_db_index()

cas_build_urls(url = "http://en.kremlin.ru/events/state-council/page/",
               start_page = 1,
               end_page = 16,
               index_group = "state-council") |>
  cas_write_db_index()

cas_build_urls(url = "http://en.kremlin.ru/events/security-council/page/",
               start_page = 1,
               end_page = 20,
               index_group = "security-council") |>
  cas_write_db_index()

cas_build_urls(url = "http://en.kremlin.ru/events/councils/page/",
               start_page = 1,
               end_page = 50,
               index_group = "councils") |>
  cas_write_db_index()
```

So here are all urls to index pages, divded by group:

```
cas_read_db_index() |>
  group_by(index_group) |>
  tally() |>
  collect()
```

```
# A tibble: 6 x 2
  index_group          n
  <chr>            <int>
1 administration      57
```

```
2 councils           50
3 news             1350
4 security-council   70
5 state-council      16
6 transcripts       470
```

Have you noticed?

All `castarter` functions start with a consistent prefix, `cas_`, to facilitate autocompletion as you write code, and are generally followed by a verb, describing the action.

Time to start downloading pages.

Step 4: Download the index pages

By default, there is a wait of one second between the download of each page, to reduce pressure on the server. The server can also respond with a standard error code, requesting longer wait times, and by default such requests will be honoured. People will have different opinions on how to go about such things, ranging from having lower download rates, to much higher rates with no wait time and concurrent download of pages. If you are not in a hurry and you do not have to download zillions of pages, you can probably leave the defaults. Some websites, however, may not like high number of requests coming from the same IP address, so you may have to wait and increase the wait time through the dedicated parameters, e.g. `cas_download(wait = 10)`).

We should be ready to start the download process, but as it turns out, there's one more thing you may need to take care of.

Dealing with limitations to access with systematic approaches

On most websites there is not a hard limit on the type and amount of contents you can retrieve, and if there is, it's usually written in the `robots.txt` file of each website. However, there's number of issues that can play out. Some relate to the geographic limitations or the use of VPNs: sometimes it's convenient to use a VPN to overcome geographic limitations on traffic, sometimes using a VPN will cause more issues.

In the case of the website of the Kremlin, it appears they have introduced an undeclared limitation on the “user_agent”, i.e., the type of software that can access their contents. Notice that in general there may be good reasons for these kind of things, and indeed, you may want to use the “user_agent” field to give hints to the website on the receiving end of your text mining adventures what it is that you want to do. If you feel you need to circumvent limitations, take a moment to consider if what you are about to do is fully appropriate. In this case, we are accessing an institutional website without putting undue pressure on the server, all contents we are retrieving are published under a creative commons license, and the website itself does not insist on any limitations on its `[robots.txt](http://kremlin.ru/robots.txt)` file. In this case, there is an easy solution: we can claim to be accessing the website as a generic modern browser, and the Kremlin’s servers will gladly let us through.

So… we’ll declare we are accessing the website as a modern browser, slightly increase the waiting time between download of each page to 3 seconds to prevent hitting rate limits, and without further ado we can start the download process of the index pages!

```
cas_download_index(
  user_agent = "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
  wait = 3)
```

In most cases, just running `cas_download_index()` without any other parameter should just work.

Time to take a break, and let the download process go. If for any reason you need to stop the process, you can safely close the session, and re-run the script at another time, and the download process will proceed from where it stopped. This is usually not an issue when you are downloading content pages, but, depending on how index pages are built, may lead to some issues if new contents are published in the meantime.

As you may see, information about the download process is stored locally in the database for reference:

```
cas_read_db_download(index = TRUE)
```

```
# A tibble: 2,013 x 5
      id batch datetime            status        size
   <dbl> <dbl> <dttm>               <int> <fs::bytes>
 1     1     8 2023-09-12 16:00:47    200         47K
 2     2     8 2023-09-12 16:00:54    200       46.4K
 3     3      8 2023-09-12 16:01:00    200        45.2K
 4     4      8 2023-09-12 16:01:07    200        44.3K
 5     5      8 2023-09-12 16:01:14    200        46.8K
 6     6      8 2023-09-12 16:01:20    200        44.7K
 7     7      8 2023-09-12 16:01:27    200          45K
 8     8      8 2023-09-12 16:01:34    200        47.7K
 9     9      8 2023-09-12 16:01:40    200          49K
10    10      8 2023-09-12 16:01:47    200        45.6K
# i 2,003 more rows
```

> Have you noticed?
>
> It is mostly not necessary to give `castarter` much information about the data themselves. Once project and website are set at the beginning of each section, then `castarter` knows where to look for finding data stored in the local database, including which urls to download, which of them have already been downloaded, and where relevant files are. This is why you don't need to tell `cas_download_index()` (and later, once urls to contents pages will be extracted, `cas_download()`) which pages should be downloaded: everything is stored in the local database and the package knows where to look for information and where to store new files without further instructions.

Step 5: Getting the links to the content pages

In order to extract links to the content pages, we should first take a look at how index pages are built, We can use the following command to see a random index page in our browser:

```
cas_browse(index = TRUE)
```

At this point, some familiarity with html comes in handy. At the most basic, you look at the page, press on "F12" on your keyboard or right-click and select "View page source", and see where links to contents are stored. They will mostly be wrapped into a container, such as a "div" or a "span" or a title such as "h2", and this will have a name.

After some trial and error, the goal is to get to a selector that consistently takes the kind of links we want.

More advanced users can pass directly the `custom_xpath` or `custom_css` parameter.

For less advanced users, I plan to add further convenience functions to `castarter` in order to smooth out the process, but until then, trial and error will have to do.

A pratical way to go about it to set `sample` to 1 (each time a new random index page will be picked), set `write_to_db` to FALSE (so that we do not store useless links in the database), and then run the command a few times until we see that we get a consistent and convincing number of links from each of these calls (obviously, more formalised methods for checking accuracy and consistency exist).

If you want to make sure you know exactly what page you are extracting links from in order to troubleshoot potential issues, you can just pick a random id, and then look at its outputs.

```r
test_file <- cas_get_path_to_files(index = TRUE, sample = 1)

test_id <- test_file$id

# cas_browse(index = TRUE, id = test_id)

cas_extract_links(id = test_id,
                  write_to_db = FALSE,
                  container = "div",
                  container_class = "hentry h-entry hentry_event")
```

In this case, I found that links to articles are always inside a "div" container of class "hentry h-entry hentry_event". However, I noticed that if I leave it at that, I'll capture also links to pages with photos or videos of the events, which I am not interested in, so I'll add a parameter to exclude all links that include "/photos" or "/videos/ in the url. I also see that links are extracted without domain name, so I make sure it is added consistently.

After checking some more at random…

```r
cas_extract_links(sample = 1,
                  write_to_db = FALSE,
                  container = "div",
                  container_class = "hentry h-entry hentry_event",
                  exclude_when = c("/photos", "/videos"),
                  domain = "http://en.kremlin.ru/")
```

…we are ready to process all the files we have collected and store the result in the local database.

```
cas_extract_links(write_to_db = TRUE,
                  reverse_order = TRUE,
                  container = "div",
                  container_class = "hentry h-entry hentry_event",
                  exclude_when = c("/photos", "/videos"),
                  domain = "http://en.kremlin.ru/")
```

This may take a couple of minutes, but as usual, no worries, if you interrupt the process, you can re-run the script and everything will proceed from where it stopped without issues.

We get links to close to 40 000 content pages.

```
cas_read_db_contents_id() |>
  pull(id) |>
  length()
```

```
[1] 40178
```

We can give a quick look at them to see if everything looks alright.

```
cas_read_db_contents_id() |>
  collect() |>
  View()
```

If it doesn't, we can simply remove the extracted links to the database with the following command, and retake it from `cas_extract_links()`.

```
cas_reset_db_contents_id()
```

If all looks fine, then it's finally time to get downloading.

Step 6: Download index pages

At a slow pace, this will take many hours, so you can probably leave your computer on overnight, and take it from there the following night. If you want to proceed with the tutorial, of course, you can just download a small subset of pages, and then use them in the following steps. You can still download the rest later, and if you re-run the script, the final dataset will include all contents as expected.

```
# we're again using the user_agent trick
cas_download(user_agent = "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537
             wait = 3)
```

Step 7: Extracting text and metadata

This is quite possibly the most technical part. I will proceed to explain the core idea, but will not get into details of how this works, which will be object of a separate post.

At the most basic, the idea is to create a named list of functions. Each of them will take the downloaded html page as an input, and store the selected part in a dedicated column. As you will see, this will extract textual contents of a page, trying to associate it with as much metadata as possible, including the location where a given press release was issued, as well as tagged individuals and themes.

```
extractors_l <- list(
  title = \(x) cas_extract_html(html_document = x,
                                container = "title") |>
    stringr::str_remove(pattern = " • President of Russia$"),

  date = \(x) cas_extract_html(
    html_document = x,
    container = "time",
    container_class = "read__published",
    attribute = "datetime",
    container_instance = 1),
  time = \(x) cas_extract_html(
    html_document = x,
    container = "div",
    container_class = "read__time",
    container_instance = 1),
  datetime = \(x) {
    stringr::str_c(cas_extract_html(
      html_document = x,
      container = "time",
      container_class = "read__published",
      attribute = "datetime",
      container_instance = 1),
      " ",
      cas_extract_html(
        html_document = x,
        container = "div",
        container_class = "read__time",
        container_instance = 1)) |>
      lubridate::ymd_hm()

  },
  location = \(x) cas_extract_html(
    html_document = x,
    container = "div",
    container_class = "read__place p-location",
    container_instance = 1),

  description = \(x) cas_extract_html(
    html_document = x,
    container = "meta",
    container_name = "description",
    attribute = "content"),

  keywords = \(x) cas_extract_html(
    html_document = x,
    container = "meta",
    container_name = "keywords",
    attribute = "content"),

  text = \(x) cas_extract_html(
    html_document = x,
    container = "div",
    container_class = "entry-content e-content read__internal_content",
    sub_element = "p") |>
    stringr::str_remove(pattern = " Published in sectio(n|ns): .*$"),

  tags = \(x) x %>%
    rvest::html_nodes(xpath = "//div[@class='read__tagscol']//a") %>%
    rvest::html_text2() %>%
    stringr::str_c(collapse = "; "),

  tags_links = \(x) x %>%
    rvest::html_nodes(xpath = "//div[@class='read__tagscol']//a") %>%
    xml2::xml_attr("href")|>
    stringr::str_c(collapse = "; "))
```

A typical workflow for finding the right combinations may look as follows: first test what works on a single page or a small set of pages, and then let the extractor process all pages.

```
current_file <- cas_get_path_to_files(sample = 1)
# current_file <- cas_get_path_to_files(id = 1)
# cas_browse(id = current_file$id)

test_df <- cas_extract(extractors = extractors_l,
                       id = current_file$id,
                       write_to_db = FALSE) %>%
  dplyr::collect()

test_df

test_df |> dplyr::pull(text)
```

```
cas_extract(extractors = extractors_l)
```

So here we are, with the full corpus of English-language posts published on the Kremlin's website.

Again, if you realise that something went wrong with the extraction, you can reset the extraction process and extract again (convenience functions for custom dropping of specific pages without necessarily re-proccesing all of them will be introduced in a future version).

```
cas_reset_db_contents_data()
```

Step 8: Celebrate! You have a textual dataset

```
corpus_df <- cas_read_db_contents_data()  |>
  collect()
```

Have you noticed?

Here as elsewhere we are using `collect()` after retrieving data from the database. Since this dataset is not huge in size, we can import it fully into

> memory for ease of use, but we don't need to: we can read a subset of data from the local database without loading into memory, thus enabling the processing of larger-than RAM datasets. A dedicated tutorial will clarify all issues that may stem from this approach.

The resulting dataset is a wide table, with the following columns:

```r
corpus_df |>
  slice(1) |>
  colnames()
```

```
 [1] "id"          "url"         "title"       "date"        "time"
 [6] "datetime"    "location"    "description" "keywords"    "text"
[11] "tags"        "tags_links"
```

And here is an example page:

```r
corpus_df |>
  slice(1) |>
  dplyr::mutate(text = stringr::str_trunc(string = text, width = 760)) %>%
  tidyr::pivot_longer(cols = everything()) |>
  knitr::kable() %>%
  kableExtra::kable_styling(full_width = TRUE)
```

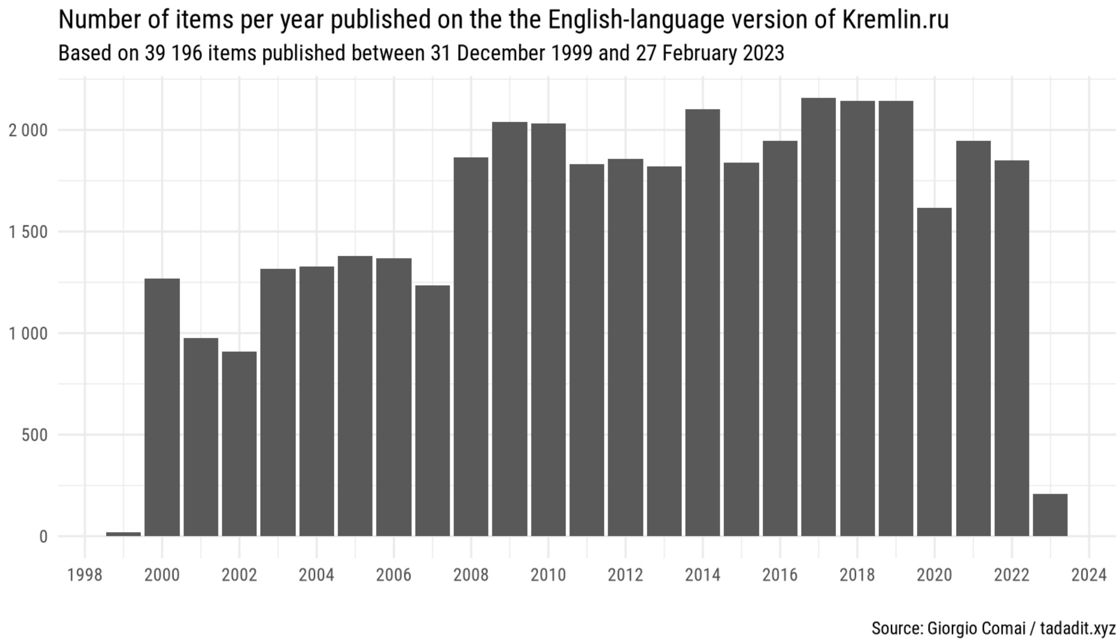
name	value
id	1
url	http://en.kremlin.ru/events/councils/7584
title	Meeting of Commission for Modernisation and Technological Development of Russia’s Economy
date	2010-04-29
time	16:30
datetime	2010-04-29 16:30:00
location	Obninsk
description	Dmitry Medvedev held meeting of Commission for Modernisation and Technological Development of Russia’s Economy on development of nuclear medicine and establishment of innovation centre in Skolkovo.
keywords	News ,President ,Commissions and Councils
text	Russia has a good scientific and technological basis for the production of radiopharmaceuticals, many qualified professionals, and a positive experience with the most advanced diagnostic techniques and treatment, primarily in relation to cancer, the President noted in his speech. However, Russia still lags behind the global average in broad use of these techniques. President Medvedev stressed the need to promptly consolidate all fields of nuclear medicine and further encourage its general development. Mr Medvedev also noted that this field has good export potential. According to the President, the innovation centre in Skolkovo should act as the impetus for establishing an innovative environment in Russia. Skolkovo will not just be a base for innov...
tags	Healthcare; Science and innovation
tags_links	/catalog/keywords/47/events; /catalog/keywords/39/events

As you may have noticed, we have kept the keyword and tag fields in a raw format, without processing them. To be most useful, they would need to be formally separated, and ideally matched with some unique identifiers, but for the time being we will keep things as they are for the sake of simplicity. We've been busy doing the extraction part, this kind of post-processing can be left for later.

Step 9: Basic information about the dataset and missing data

Time to have a quick look at the dataset to see if there's some evident data issue.

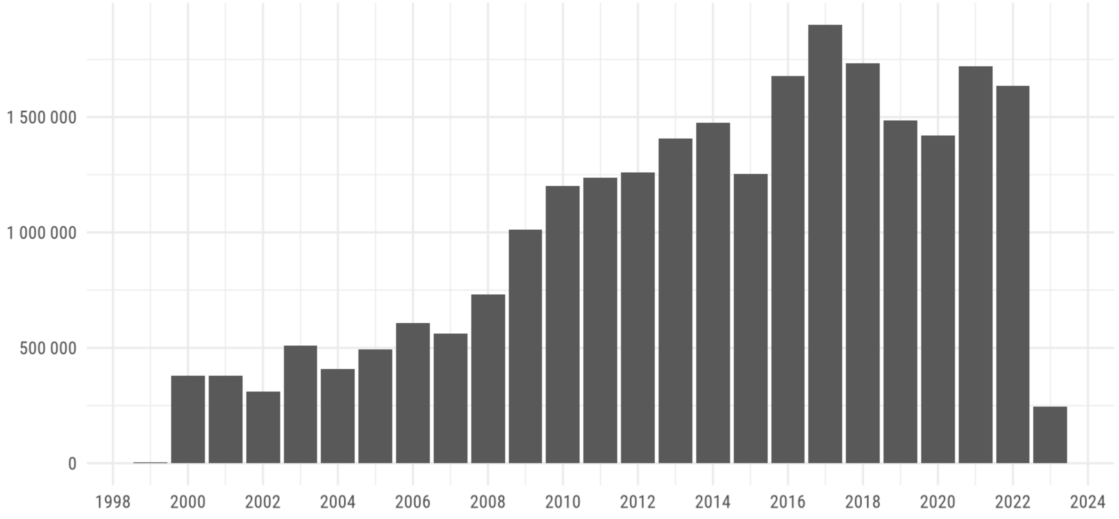
```r
corpus_df |>
  mutate(year = lubridate::year(date)) |>
  count(year) |>
  ggplot(mapping = aes(x = year, y = n)) +
  geom_col() +
  scale_y_continuous(name = "", labels = scales::number) +
  scale_x_continuous(name = "", breaks = scales::pretty_breaks(n = 10)) +
  labs(
    title = "Number of items per year published on the the English-language version
    subtitle = stringr::str_c(
      "Based on ",
      scales::number(nrow(corpus_df)),
      " items published between ",
      format.Date(x = min(corpus_df$date), "%d %B %Y"),
      " and ",
      format.Date(x = max(corpus_df$date), "%d %B %Y")),
    caption = "Source: Giorgio Comai / tadadit.xyz"
  )
```



We see a slight increase in the number of posts since 2008, a slump in 2020 (covid-related?), but not much. Looking at the word count, we also see that the posts have been getting longer on average, so there's more than three times as much text from recent years than on previous ones (which is relevant, as it impacts non-weighted word counts).

```r
words_per_day_df <- corpus_df |>
  cas_count_total_words() |>
  mutate(date = lubridate::as_date(date),
         pattern = "total words")

words_per_day_df |>
  cas_summarise(period = "year", auto_convert = TRUE) |>
  rename(year = date) |>
  ggplot(mapping = aes(x = year, y = n)) +
  geom_col() +
  scale_y_continuous(name = "", labels = scales::number) +
  scale_x_continuous(name = "", breaks = scales::pretty_breaks(n = 10)) +
  labs(
    title = "Number of words per year published on the the English-language version
    subtitle = stringr::str_c("Based on ",
                              scales::number(nrow(corpus_df)),
                              " items published between ",
                              format.Date(x = min(corpus_df$date), "%d %B %Y"),
                              " and ",
                          format.Date(x = max(corpus_df$date), "%d %B %Y")),
   caption = "Source: Giorgio Comai / tadadit.xyz")
```



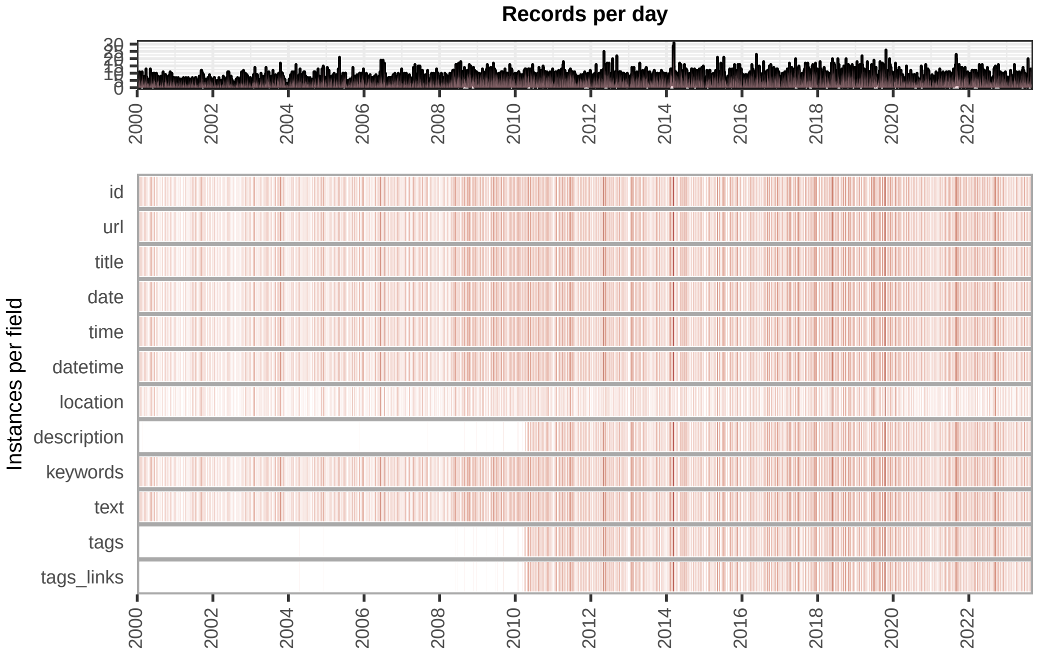
Number of words per year published on the the English-language version of Kremlin.ru

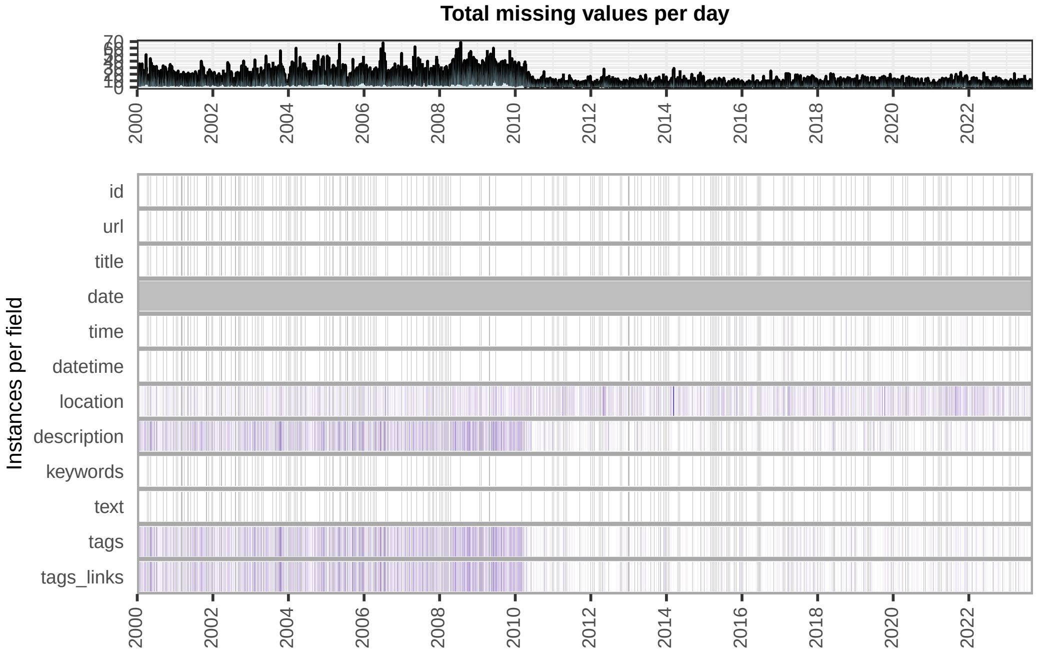
Based on 39 196 items published between 31 December 1999 and 27 February 2023

Source: Giorgio Comai / tadadit.xyz

But these are just basic impressions about what is going on… in generale, it’s a good idea to make some proper data quality checks. A fully formalised approach to this part of the process will be detailed in a separate post, but we can achieve a lot of information on the availability of data by looking at the reports generated with the `daiquiri` package Quan (2022).

As the following graphs show, tags and description are available only starting with 2010. Location is often missing, but it is occasionally available starting with the early years. Depending on the type of analysis we are interested in, these data may need to go through some further quality checks, but at its most basic, things look as expected.





Step 10: Archive and backup

All looks good, and we can move on to create other textual datasets or to analyse what we have collected. But to keep things in good order, it's a good idea to compress the tens of thousands of html files we have downloaded to free up some space.

This can be easily achieved with the following command.

```
cas_archive()
```

You will find yourself with some big compressed files that can be stored in a backup on your favourite service. The local database keeps track of which file is where.

Depending on the type of dataset you are working on, it may make sense to ensure that you are not the only person to hold a copy of the original source. The following functions check if the original pages exist on the Internet Archive Wayback Machine, and if they are not, they try, very slowly, to add them to the archive. All of the Kremlin's website is available there already, so no issues in this case, but for smaller websites I feel there is some use in ensuring their long term availability. More efficient approaches to achieve the same result will be considered.

```
cas_ia_check()

cas_ia_save()
```

Step 11: Keep the dataset updated

This is quite easy, as this is part of the whole point of having built `castarter`, but I will describe this in a separate post.

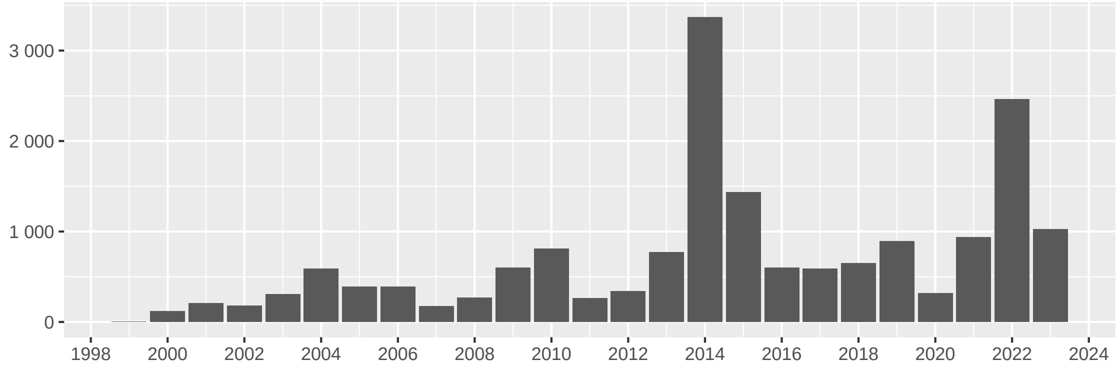
Step 12: At long last, look at the data

Here's just a couple of quick examples for reference. But here is where this tutorial ends, and the actual content analysis begins.

```
corpus_df |>
  cas_count(pattern = "Ukrain") |>
  cas_summarise(period = "year",
                auto_convert = TRUE) |>
  rename(year = date) |>
  ggplot(mapping = aes(x = year, y = n)) +
  geom_col() +
  scale_y_continuous(name = "", labels = scales::number) +
  scale_x_continuous(name = "", breaks = scales::pretty_breaks(n = 10)) +
  labs(
    title = stringr::str_c(
      "Number of mentions of ",
      sQuote("Ukraine"),
      " per year on Kremlin.ru"),
    subtitle = stringr::str_c(
      "Based on ",
      scales::number(nrow(corpus_df)),
      " items published in English between ",
      format.Date(x = min(corpus_df$date), "%d %B %Y"),
      " and ",
      format.Date(x = max(corpus_df$date), "%d %B %Y")),
    caption = "Source: Giorgio Comai / tadadit.xyz")
```

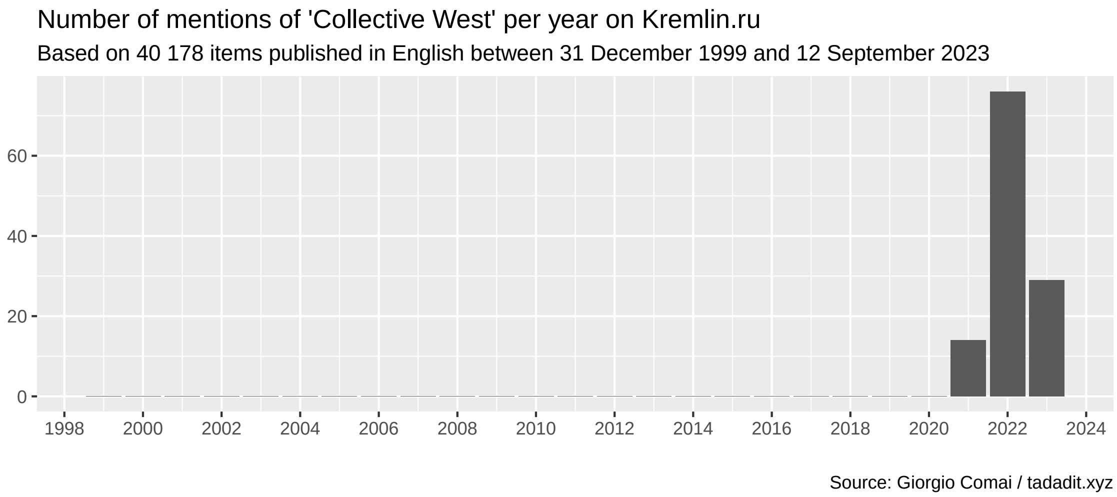
Number of mentions of 'Ukraine' per year on Kremlin.ru

Based on 40 178 items published in English between 31 December 1999 and 12 September 2023



Source: Giorgio Comai / tadadit.xyz

```
corpus_df |>
  cas_count(pattern = "Collective West") |>
  cas_summarise(period = "year",
                auto_convert = TRUE) |>
  rename(year = date) |>
  ggplot(mapping = aes(x = year, y = n)) +
  geom_col() +
  scale_y_continuous(name = "", labels = scales::number) +
  scale_x_continuous(name = "", breaks = scales::pretty_breaks(n = 10)) +
  labs(
    title = stringr::str_c(
      "Number of mentions of ",
      sQuote("Collective West"),
      " per year on Kremlin.ru"),
    subtitle = stringr::str_c(
      "Based on ",
      scales::number(nrow(corpus_df)),
      " items published in English between ",
      format.Date(x = min(corpus_df$date), "%d %B %Y"),
      " and ",
      format.Date(x = max(corpus_df$date), "%d %B %Y")),
    caption = "Source: Giorgio Comai / tadadit.xyz")
```



Number of mentions of 'Collective West' per year on Kremlin.ru

Based on 40 178 items published in English between 31 December 1999 and 12 September 2023

Source: Giorgio Comai / tadadit.xyz

For an example of basic content analysis on a dataset such as this one, see the post: “**Who said it first? ‘The collective West’ in Russia’s nationalist media and official statements**”

References

Chimiris, Ekaterina. 2022. «The Collective West Concept and Selected Western Actors (Germany, Norway, Estonia, NATO) in the Russian Media: Post-Crimea Dynamics». *Global Journal of Human-Social Science* 22 (F1): 1–12. https://doi.org/10.34257/GJHSSFVOL22IS1PG1.

Comai. 2021. «Kremlin_en - A Textual Dataset Based on the Contents Published on the English-Language Version of the Kremlin's Website – A Corpus in Tabular Format with All Posts Published on the Official Website of the President of the Russian Federation Between 31 December 1999 and 31 December 2020». *Https://Discuss-Data.net/Dataset/5eb1481e-Ae89-45bf-9c88-03574910730a/*, maggio. https://doi.org/10.48320/5EB1481E-AE89-45BF-9C88-03574910730A.

Comai, Giorgio. 2017. «Quantitative Analysis of Web Content in Support of Qualitative Research. Examples from the Study of Post-Soviet De Facto States». *Studies of Transition States and Societies* 9 (1).

Darczewska, Jolanta, e Piotr Żochowski. 2015. *Russophobia in the Kremlin's Strategy. A Weapon of Mass Destruction*. Wyd. 1. Punkt Widzenia, Numer 56, październik 2015. Warszawa: Ośrodek Studiów Wschodnich im. Marka Karpia.

Feklyunina, Valentina. 2012. «Constructing Russophobia». In, a cura di Ray Taras, 91–109. London; New York: Routledge.

Laruelle, Marlene. 2016. «Russia as an anti-liberal European civilisation». In, a cura di Pål Kolstø e Helge Blakkisrud, 275–97. Edinburgh: Edinburgh University Press. https://www.universitypressscholarship.com/view/10.3366/edinburgh/9781474410427.001.0001/upso-9781474410427-chapter-011.

∅stb∅, Jardar. 2017. «Securitizing "Spiritual-Moral Values" in Russia». *Post-Soviet Affairs* 33 (3): 200–216. https://doi.org/10.1080/1060586X.2016.1251023.

President of the Russian Federation. 2015. «Strategiya natsionalnoi bezopasnosti Rossiiskoi Federatsii». https://rg.ru/documents/2015/12/31/nac-bezopasnost-site-dok.html.

———. 2021. «Strategiya natsionalnoi bezopasnosti Rossiiskoi Federatsii». http://kremlin.ru/acts/bank/47046.

———. 2023. «Kontseptsiya vneshnei politiki Rossiiskoi Federatsii». https://rg.ru/documents/2023/03/31/prezident-ukaz229-site-dok.html.

Quan, T. Phuong. 2022. «Daiquiri: Data Quality Reporting for Temporal Datasets». *Journal of Open Source Software* 7 (80): 5034. https://doi.org/10.21105/joss.05034.

Quenoy, Irina du, e Dmitry Dubrovskiy. 2018. «Violence and the Defense of "Traditional Values" in the Russian Federation». In, a cura di Olga Oliker, 93–116. Washington, DC: Center for Strategic; International Studies. https://www.csis.org/analysis/religion-and-violence-russia.

Robinson, Neil. 2019. «Russophobia in official Russian political discourse». *De Europa* 2 (2): 61–77. https://ulir.ul.ie/handle/10344/8429.

Rodkiewicz, Witold, e Jadwiga Rogoża. 2015. «Potemkin Conservatism: an Ideological Tool of the Kremlin». Warsaw. https://www.osw.waw.pl/en/publikacje/point-view/2015-02-03/potemkin-conservatism-ideological-tool-kremlin.

Sharafutdinova, Gulnaz. 2014. «The Pussy Riot affair and Putin's démarche from sovereign democracy to sovereign morality». *Nationalities Papers* 42 (4): 615–21. https://doi.org/10.1080/00905992.2014.917075.

Tsygankov, Andrei. 2016. «Crafting the State-Civilization Vladimir Putin's Turn to Distinct Values». *Problems of Post-Communism* 63 (3): 146–58. https://doi.org/10.1080/10758216.2015.1113884.