PE Population and Economics

Население и экономика 5(4): 21–29
DOI 10.3897/popecon.5.e72783

СТАТЬЯ О БАЗЕ ДАННЫХ

# База данных публикаций электронных средств массовой информации о материнском (семейном) капитале в России в 2006–2019 гг.

Ирина Е. Калабихина[1], Герман А. Клименко[2], Евгений П. Банин[2], Екатерина К. Воробьева[2], Анна Д. Ламеева[2]

1 *МГУ имени М. В. Ломоносова, Москва, 119991, Россия*

2 *Московский государственный технический университет имени Н.Э. Баумана, Москва, 105005, Россия*

3 *Научно-исследовательский центр «Курчатовский институт», Москва, 123182, Россия*

**Получено** 1 December 2021 ♦ **Принято в печать** 1 December 2021 ♦ **Опубликовано** 7 December 2021

**Цитирование:** IE Kalabikhina, HA Klimenko, EP Banin, EK Vorobyeva, AD Lameeva (2021) Database of digital media publications on maternal (family) capital in Russia in 2006–2019. Population and Economics 5(4): 21–29. https://doi.org/10.3897/popecon.5.e78723

## Аннотация

В базе представлены данные публикаций электронных русскоязычных СМИ Российской Федерации по тематике материнского капитала в период с 10 мая 2006 г. по 30 июня 2019 г. База включает выгрузку общих данных по публикациям по тематике материнского капитала в формате .csv (кодировка UTF-8). Полные тексты публикаций представлены в формате .xml.

База составлена на основе специализированного запроса для агрегатора публикаций русскоязычных электронных средств массовой информации public.ru. Всего база данных состоит из 457 888 публикаций 7 665 издательств из 1 251 населенного пункта в России на территории 85 регионов. База данных включает в себя информацию о дате и типе публикации, издательстве и месте публикации (муниципалитет), тексты о материнском капитале и содержащееся в них количество уникальных положительных, отрицательных и нейтральных слов и фраз согласно словарю RuSentiLex2017, а также полноценные тексты публикаций.

## Ключевые слова

база данных, электронные СМИ, материнский (семейный) капитал, центральные и муниципальные СМИ, Россия, анализ тональности

**Коды JEL:** J10, J13, Z18.

**Название базы данных (на английском языке):** Database of digital media publications on maternal (family) capital in Russia in 2006–2019

## Формат данных и доступ

База данных состоит из полнотекстовых публикаций электронных средств массовой информации по тематике материнского капитала. Материалы на русском языке опубликованы в федеральных, региональных и локальных электронных средствах массовой информации. Период публикаций: с 10 мая 2006 г. по 30 июня 2019 г. Незначительное число публикаций за 2004–2005 гг. в базе (менее 100 штук) связано с упоминанием в них семейного капитала.

База данных состоит из 457 888 публикаций 7 665 издательств из 1 251 населенного пункта в России на территории 85 регионов. Формат представления .csv, .xml (полные тексты).

Доступ к данным: https://doi.org/10.5281/zenodo.5740417 [Kalabikhina et al., 2021].

Файл «Matkap_SMI_17_11_2021.csv» содержит обработанную информацию из расширенной полнотекстовой выборки по годам (содержится в архивном файле «XML.rar»).

## Методология сбора данных

Использован агрегатор публикаций русскоязычных электронных средств массовой информации public.ru. Выборка публикаций преимущественно ограничивалась временным периодом с 10 мая 2006 г. (Президент РФ В.В. Путин в послании Федеральному Собранию впервые анонсировал программу материнского капитала как один из механизмов по стимулированию рождаемости и выхода из демографического кризиса, [Федеральный закон…, 2006]) по 30 июня 2019 г. (до этой даты программа позволяла сделать полноценную выгрузку публикаций СМИ без потерь в период загрузки публикаций с 01.08.2021 по 15.08.2021).

Ключевые слова, использованные для отбора статей по тематике материнского капитала: «маткапитал», «материнский капитал», «семейный капитал», «отцовский капитал». В публикации должны были как минимум два раза повторяться фразы из запроса, при этом расстояние между фразами должно было быть не более четырех предложений. Тем самым исключались публикации, в которых тема материнского капитала упоминается случайно, косвенно.

В базе данных удалены дубликаты статей. К дубликатам относились публикации, которые включали в себя полноценное повторение самого текста публикации, издательства и муниципалитета (локации) издательства. Повторы (перепечатка) статей в других издательствах или в других регионах не исключались.

После лемматизации текста (а также после сокращения текста до нижнего регистра, удаления лишних пробелов, цифр и знаков препинания) были также подсчитаны уникальные положительные, отрицательные и нейтральные слова и фразы (переменные) — согласно словарю RuSentiLex2017 [Loukachevitch, Levchik, 2016]. Повторения тональных слов не засчитывались.

## Структура базы данных и описание переменных

Пояснения к переменным в базе данных даны в табл. 1.

База полных текстов (в архивном файле «XML.rar») содержит дополнительную информацию: заголовки публикаций, фамилию и имя автора(ов), субъект РФ, где находится редакция издательства (в наборе данных присутствуют публикации из 85 регионов РФ), и некоторую дополнительную информацию.

**Таблица 1.** Переменные в базе данных публикаций электронных средств массовой информации о материнском (семейном) капитале в России (файл «Matkap_SMI_17_11_2021.csv»)

| Заголовок столбца | Описание и комментарии |
|---|---|
| id | Идентификационный номер публикации (первые номера присваиваются более поздним публикациям) |
| pubData | Дата публикации в формате «YYYY-MM-DD» |
| text | Текст из «описания» в xml-файлах о материнском (семейном) капитале после лемматизации (удаление знаков препинания, строчных букв, пробелов, цифр) |
| source | Издательство, в котором была опубликована статья |
| place | Муниципалитет РФ, где находится редакция издательства. Всего в набор данных вошли публикации из 1 251 муниципалитета |
| type | Типы публикаций: бюллетень, интернет-ресурс, газета, журнал, интернет-издание, информационное агентство, пресс-релиз, радиопрограмма, телепрограмма |
| period | Частота публикаций: ежедневное, ежемесячное, еженедельное, два раза в неделю, раз в квартал, раз в две недели, другое |
| positive | Число уникальных позитивных слов и выражений из словаря RuSentiLex2017 |
| negative | Число уникальных негативных слов и выражений из словаря RuSentiLex2017 |
| neutral | Число уникальных нейтральных слов и выражений из словаря RuSentiLex2017 |

## Распределение публикаций по годам, типам и издательствам

Распределение публикаций в основной базе по годам за весь рассматриваемый период после удаления дубликатов по типам, регионам и издательствам приведено на рис. 1 и в табл. 2–4.

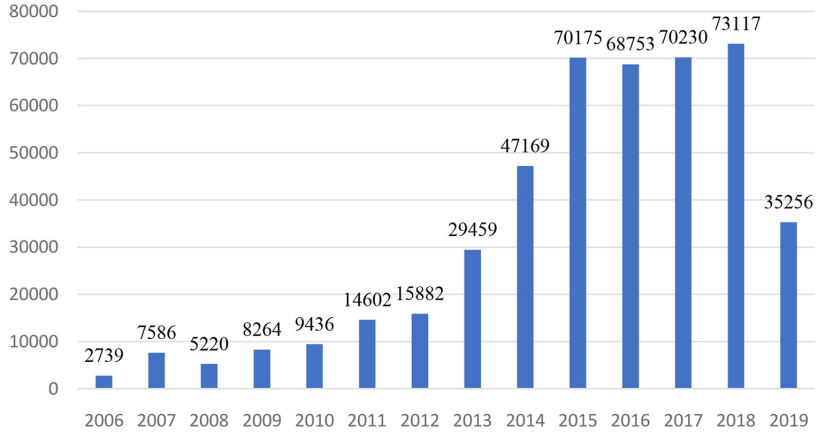


**Рис. 1.** Количество публикаций, попавших в выборку, по годам. *Примечание:* Количество публикаций в 2006 году рассчитывалось в период с 10.05.2006 по 31.12.2006; количество публикаций в 2019 году рассчитывалось в период с 01.01.2019 по 30.06.2019.

Наибольшее количество публикаций наблюдалось в период с 2015 по 2018 гг. Стремительный рост количества публикаций в период с 2012 по 2015 гг. связан с увеличением числа электронных СМИ в Российской Федерации, причиной которого стало распространение высокоскоростного доступа к интернету.

**Таблица 2.** Распределение публикаций, попавших в выборку, по типу

| Тип публикации | Количество публикаций | % от общего количества |
|---|---|---|
| Интернет-ресурс | 323 256 | 70,60 |
| Газета | 63 096 | 13,78 |
| Информационное агентство | 31 563 | 6,89 |
| Интернет-издание | 31 396 | 6,86 |
| Пресс-релиз | 3 519 | 0,77 |
| Телепрограмма | 2 466 | 0,54 |
| Журнал | 2 245 | 0,49 |
| Бюллетень | 323 | 0,07 |
| Радиопрограмма | 24 | 0,01 |

Основная доля публикаций по тематике материнского капитала приходится на публикации интернет-ресурсов. При этом суммарная доля публикаций типа «пресс-релиз», «телепрограмма», «журнал», «бюллетень» и «радиопрограмма» составляет менее 2% от всех публикаций.

**Таблица 3.** Регионы с наибольшим количеством публикаций СМИ по тематике материнского капитала (топ-15)

| Субъект РФ | Количество публикаций |
|---|---|
| Москва | 176 627 |
| Санкт-Петербург | 14 082 |
| Свердловская область | 11 843 |
| Чувашская Республика | 8 935 |
| ХМАО-Югра | 8 673 |
| Республика Татарстан | 8 663 |
| Ростовская область | 7 628 |
| Краснодарский край | 6 930 |
| Приморский край | 6 430 |
| Пензенская область | 5 957 |
| Челябинская область | 5 870 |
| Московская область | 5 414 |
| Пермский край | 5 325 |
| Алтайский край | 5 235 |
| Красноярский край | 5 227 |

По абсолютному числу публикаций с большим отрывом лидирует Москва, затем идут Санкт-Петербург и Свердловская область.

**Таблица 4.** Издательства и СМИ с наибольшим количеством публикаций по тематике материнского капитала (топ-20)

| Издательство | Количество публикаций | Издательство | Количество публикаций |
|---|---|---|---|
| Mngz.ru | 5 278 | Ttfinance.ru | 1 214 |
| Gorodskoyportal.ru/moskva/ | 4 381 | Regions.ru | 1 212 |
| ИА Регнум | 3 095 | Publishernews.ru | 1 198 |
| Gorodskoyportal.ru/ekaterinburg/ | 2 570 | Chelyabinsk.bezformata.ru | 1 187 |
| Sockart.ru | 1 970 | Podmoskovye.bezformata.ru | 1 187 |
| yodda.ru | 1 780 | Governors.ru | 1 130 |
| Российская газета | 1 724 | Ekaterinburg.bezformata.ru | 1 126 |
| Gov.cap.ru | 1 700 | Media-office.ru | 1 126 |
| РИА Новости | 1 674 | Pskov.bezformata.ru | 1 098 |
| Cherkesk.bezformata.ru | 1 648 | Barnaul.bezformata.ru | 1 077 |

Представленные в приведенном перечне издательства относятся как к федеральному, так и к региональному и местному уровням.

## Изменения в распределении публикаций по типам, регионам и издательствам

Изменения в распределениях публикаций по типам, регионам и издательствам в течение рассматриваемого периода приведены на рис. 2 и 3, в табл. 5 и 6.

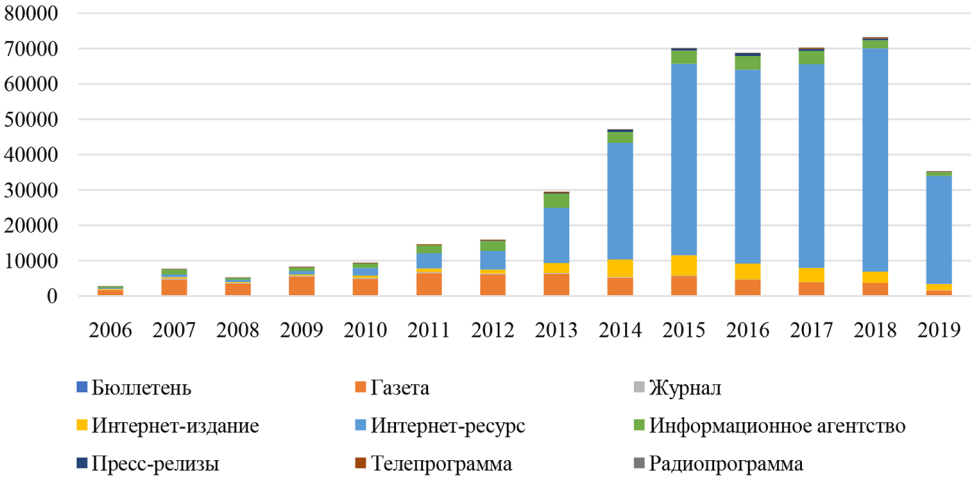


**Рис. 2.** Изменение в распределении публикаций по типам по абсолютному числу публикаций за каждый год. *Примечание:* Количество публикаций в 2006 году рассчитывалось в период с 10.05.2006 по 31.12.2006; количество публикаций в 2019 году рассчитывалось в период с 01.01.2019 по 30.06.2019.

Основной рост числа публикаций СМИ по тематике материнского капитала в 2012–2015 гг. происходил за счет увеличения числа публикаций интернет-ресурсов.

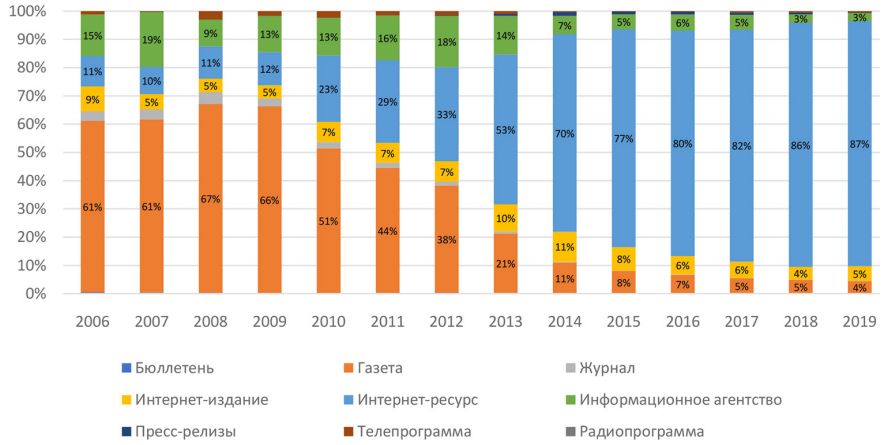


**Рис. 3.** Изменение в распределении публикаций по типам по относительному числу публикаций за каждый год

Прирост публикаций типа «интернет-ресурс» в основном происходил за счет уменьшения доли газет в общем наборе, доля которых за рассматриваемый период снизилась с 60,7% в 2006 г. (при этом максимум наблюдался в 2008 г. — 66,8%) до 4,4% в 2019 г., уступая по количеству публикаций интернет-изданиям. Также за рассматриваемый период снизилась доля публикаций информационных агентств: с 19,5% в 2007 г. до 2,9% в 2019 г.

Москва каждый год демонстрирует наибольшие показатели числа публикаций по теме материнского капитала (табл. 5). Доля других регионов часто демонстрирует сильную волатильность (к примеру, в Ханты-Мансийском АО, где доля публикаций в общем количестве публикаций снизилась с 4,78 % в 2017 г. до 0,77% в 2019 г.), однако доля каждого региона не превышает 5% ежегодно.

Многие издательства демонстрируют нулевые показатели за год (табл. 6). Это связано с отсутствием данного издательства в интернет-среде в данный год. Основной рост числа издательств в интернет-среде наблюдается в период с 2012 по 2015 гг.

## Возможное применение базы данных

Данные подходят для анализа публикационной активности отдельных издательств (переменная «Publishing House/Information Agency») или издательств территориально-административных единиц (переменная «Location (municipality) of the Publishing House» для муниципалитетов и «Region of the Publishing House» — для субъектов РФ) на территории РФ по тематике материнского капитала. Публикационная активность регионов позволяет проанализировать реакцию СМИ на события в сфере материнского капитала, происходящие как на федеральном, так и региональном или местном уровне: к примеру, реакцию СМИ на ключевые даты, связанные с рассмотрением, обсуждением, принятием и публикацией законопроекта на федеральном уровне. Измерение публикационной активности позволяет оценить настроения в регионе относительно демографического события, происходящего на любом уровне.

**Таблица 5.** Распределение публикаций по регионам (15 регионов с наибольшим общим количеством публикаций СМИ), 2006–2019 гг., % по столбцу

| Регион | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Москва | 45,02 | 30,62 | 31,21 | 26,33 | 27,65 | 24,85 | 26,63 | 40,12 | 43,49 | 40,78 | 42,33 | 40,12 | 36,54 | 39,28 |
| Санкт-Петербург | 4,82 | 4,67 | 2,32 | 2,69 | 3,25 | 3,16 | 3,58 | 3,99 | 3,31 | 2,25 | 1,90 | 3,29 | 3,64 | 3,74 |
| Свердловская область | 2,15 | 2,37 | 1,30 | 1,43 | 2,30 | 2,64 | 2,13 | 2,37 | 3,00 | 3,32 | 2,97 | 2,25 | 2,18 | 2,31 |
| Чувашская Республика | 0,40 | 1,11 | 0,82 | 0,83 | 0,35 | 0,56 | 0,68 | 1,42 | 3,18 | 2,72 | 2,09 | 2,20 | 1,74 | 1,19 |
| ХМАО-Югра | 0,99 | 0,80 | 1,07 | 0,93 | 1,02 | 0,98 | 0,93 | 0,94 | 1,80 | 0,58 | 2,40 | 4,78 | 1,72 | 0,77 |
| Республика Татарстан | 0,51 | 2,89 | 1,13 | 1,04 | 1,58 | 1,49 | 1,24 | 1,86 | 1,88 | 1,82 | 1,89 | 2,32 | 1,90 | 1,96 |
| Ростовская область | 1,61 | 2,56 | 2,61 | 2,88 | 2,79 | 3,24 | 3,39 | 1,84 | 1,46 | 1,47 | 1,38 | 1,46 | 1,43 | 1,30 |
| Краснодарский край | 1,06 | 1,60 | 2,80 | 2,86 | 2,62 | 3,05 | 2,86 | 2,21 | 2,18 | 1,43 | 1,08 | 1,12 | 0,90 | 1,08 |
| Приморский край | 1,83 | 2,68 | 1,95 | 1,67 | 1,59 | 3,17 | 3,60 | 1,64 | 0,99 | 0,99 | 1,28 | 1,23 | 1,14 | 1,50 |
| Пензенская область | 0,47 | 0,26 | 0,25 | 0,48 | 0,39 | 0,15 | 0,12 | 0,57 | 1,54 | 1,97 | 1,46 | 1,15 | 1,69 | 1,31 |
| Челябинская область | 0,51 | 1,40 | 1,86 | 3,34 | 2,16 | 1,34 | 1,53 | 1,60 | 1,17 | 1,28 | 0,98 | 1,13 | 1,11 | 1,50 |
| Московская область | 0,51 | 1,04 | 0,98 | 0,65 | 0,48 | 0,95 | 1,11 | 0,56 | 0,91 | 1,21 | 1,32 | 0,76 | 1,80 | 1,86 |
| Пермский край | 1,83 | 1,67 | 2,28 | 1,56 | 1,38 | 1,34 | 1,65 | 1,32 | 1,48 | 1,26 | 1,29 | 0,94 | 0,75 | 0,69 |
| Алтайский край | 1,06 | 1,53 | 0,69 | 1,16 | 1,48 | 1,76 | 1,34 | 1,09 | 0,88 | 0,81 | 1,00 | 1,37 | 1,27 | 1,32 |
| Красноярский край | 1,10 | 1,16 | 1,02 | 1,32 | 1,95 | 3,42 | 1,64 | 1,14 | 1,10 | 1,11 | 1,04 | 0,92 | 0,95 | 0,87 |

**Таблица 6.** Изменение в распределении публикаций по издательствам (20 издательств с наибольшим общим количеством публикаций) по абсолютному числу публикаций за каждый год

| Издательства | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mngz.ru | 0 | 0 | 0 | 0 | 0 | 0 | 26 | 53 | 620 | 0 | 1 091 | 2 899 | 589 | 0 |
| Gorodskoyportal.ru/moskva/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 353 | 823 | 1 469 | 1 624 | 80 | 32 |
| ИА Регнум | 70 | 444 | 182 | 406 | 198 | 268 | 328 | 374 | 137 | 221 | 199 | 149 | 61 | 58 |
| Gorodskoyportal.ru/ekaterinburg/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 343 | 824 | 814 | 163 | 289 | 137 |
| Sockart.ru | 0 | 0 | 0 | 0 | 0 | 0 | 94 | 390 | 630 | 620 | 234 | 2 | 0 | 0 |
| yodda.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 968 | 715 | 97 | 0 |
| Российская газета | 87 | 140 | 109 | 124 | 120 | 142 | 121 | 144 | 129 | 152 | 142 | 126 | 121 | 67 |
| Gov.cap.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 164 | 561 | 582 | 165 | 162 | 66 | 0 |
| РИА Новости | 0 | 0 | 0 | 0 | 0 | 212 | 318 | 364 | 156 | 79 | 89 | 212 | 149 | 95 |
| Cherkesk.bezformata.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 101 | 346 | 411 | 539 | 251 |
| Ttfinance.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 25 | 49 | 403 | 482 | 209 | 41 |
| Regions.ru | 55 | 161 | 72 | 112 | 81 | 94 | 103 | 131 | 136 | 85 | 55 | 66 | 52 | 9 |
| Publishernews.ru | 0 | 0 | 0 | 2 | 1 | 3 | 7 | 85 | 183 | 175 | 251 | 195 | 251 | 45 |
| Chelyabinsk.bezformata.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 88 | 140 | 253 | 413 | 293 |
| Podmoskovye.bezformata.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 45 | 106 | 217 | 567 | 252 |
| Governors.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 330 | 307 | 317 | 176 | 0 | 0 | 0 |
| Ekaterinburg.bezformata.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 61 | 127 | 258 | 445 | 235 |
| Media-office.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 280 | 273 | 242 | 194 | 108 | 29 |
| Pskov.bezformata.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 79 | 212 | 264 | 404 | 139 |
| Barnaul.bezformata.ru | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 60 | 127 | 263 | 380 | 247 |

Анализ тональности выполнен для элементов всей базы данных, что позволяет изучать тональность различных групп публикаций.

В следующих версиях мы планируем расширить базу данных публикаций СМИ по другим запросам и тематикам, связанным с политикой в сфере рождаемости: пособия и декретные отпуска.

## Список литературы

1. Loukachevitch N., Levchik A. (2016) Creating a General Russian Sentiment Lexicon. In: N.Calzolari, K.Choukri et al. (eds.) Proceedings of the Tenth International Language Resources and Evaluation Conference (LREC-2016), Portorož (Slovenia), May 2016. URL: https://aclanthology.org/L16-1186.pdf
2. Kalabikhina I.E., Klimenko G.A., Banin E.P., Vorobieva E.K., Lameeva A.D. (2021). Database of digital media publications on maternal (family) capital in Russia in 2006-2019 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5740417

## Другие источники информации

1. Федеральный закон (2006) «О дополнительных мерах государственной поддержки семей, имеющих детей» (с изменениями и дополнениями) от 29 декабря 2006 г. № 256-ФЗ. URL: https://base.garant.ru/12151286/ (дата обращения 15.11.2021)

## Сведения об авторах

- Калабихина Ирина Евгеньевна — доктор экономических наук, профессор, заведующая кафедрой народонаселения экономического факультета, школа «Мозг, когнитивные системы, искусственный интеллект», МГУ имени М.В. Ломоносова, E-mail: kalabikhina@econ.msu.ru
- Клименко Герман Андреевич — аспирант, инженер научно-исследовательской лаборатории экономики народонаселения и демографии экономического факультета МГУ имени М.В. Ломоносова, E-mail: german89000@mail.ru
- Банин Евгений Петрович — кандидат технических наук, инженер-исследователь, Научно-исследовательский центр «Курчатовский институт», Московский государственный технический университет имени Н.Э. Баумана, E-mail: evg.banin@gmail.com
- Воробьева Екатерина Кирилловна — студентка магистратуры, экономический факультет МГУ имени М.В. Ломоносова, E-mail: ekaterina.vrb@mail.ru
- Ламеева Анна Дмитриевна — студентка магистратуры, экономический факультет МГУ имени М.В. Ломоносова, E-mail: lameevaanna@mail.ru