

ЭЛЕКТРОННЫЕ РЕСУРСЫ ОТКРЫТОГО ДОСТУПА. ОТКРЫТЫЕ АРХИВЫ ИНФОРМАЦИИ

УДК 004.65 + 001.103 + 004.8:02

<https://doi.org/10.33186/1027-3689-2026-7-66-80>

Что такое датасет и зачем он нужен в библиотеке

М. В. Вахрушев¹, М. В. Гончаров², К. А. Колосов³

^{1,2}ГПНТБ России, Москва, Российская Федерация

³Ассоциация ЭБНИТ, Москва, Российская Федерация

¹vahrushev@gpntb.ru

²goncharov@gpntb.ru

³kolosov@gpntb.ru

Аннотация. В представленной статье рассматривается концепция датасетов, предназначенных для задач машинного обучения алгоритмов/моделей синтетических нейронных сетей, то есть моделей искусственного интеллекта. В Российской Федерации реализуется Стратегия развития информационного общества на 2017–2030 гг., цель которой – формирование в стране общества знаний. Современные библиотеки уже аккумулируют, сохраняют и распространяют существующие знания, а в рамках работы с датасетами выступают провайдерами концепции этой стратегии. Датасет можно определить как структурированный набор данных, в котором они организованы для решения тех или иных задач, к тому же структура датасета, в свою очередь, конструктивно зависит от типа, характера и формализованности первичных данных. Для обоснования этого утверждения и подготовлена настоящая статья. В тексте приводятся две базовые технологические модели формирования датасета (набора данных на основе текстового корпуса).

Публикация подготовлена в рамках Государственного задания ГПНТБ России № 075-00550-26-00 на 2026 г. по выполнению НИР по теме № НИОКТР 126011215424-5 (FNEG-2025-0004).

Ключевые слова: датасет, альтметрики, библиографическое описание

Для цитирования: Вахрушев М. В., Гончаров М. В., Колосов К. А. Что такое датасет и зачем он нужен в библиотеке // Научные и технические библиотеки. 2026. № 7. С. 66–80. <https://doi.org/10.33186/1027-3689-2026-7-66-80>

OPEN ACCESS DIGITAL RESOURCES. OPEN INFORMATION ARCHIVES

UDC 004.65 + 001.103 + 004.8:02
<https://doi.org/10.33186/1027-3689-2026-7-66-80>

What dataset is and why the libraries need it

Maxim V. Vakhrushev¹, Mikhail V. Goncharov² and Kirill A. Kolosov³

*^{1, 2}Russian National Public Library for Science and Technology,
Moscow, Russian Federation*

³ELNIT Association, Moscow, Russian Federation

¹vahrushev@gpntb.ru

²goncharov@gpntb.ru

³kolosov@gpntb.ru

Abstract. The authors examine the concept of datasets intended for machine learning of algorithms / synthetic neuronet models, i. e. artificial intelligence (AI) models. Russian Federation pursues the Strategy for Information Society for 2017–2030, with the goal to build knowledge society. Today the libraries have been accumulating, preserving and disseminating knowledge, and in the context of datasets, providing the strategy concept. The authors define dataset as a structured data array organized to solve particular problems: the dataset structure, in its turn, depends on the primary data type, character and formality. In the article, the authors substantiate this statement and discuss two basic technological models for building datasets (based on text corpus).

The paper is prepared under the Government Order to RNPLS&T No. 075-00550-26-00 for 2026, R&D theme No. 126011215424-5 (FNEG-2025-0004).

Keywords: dataset, altmetrics, bibliographic record

Cite: Vakhrushev M. V., Goncharov M. V., Kolosov K. A. What dataset is and why the libraries need it // Scientific and technical libraries. 2026. No. 7, pp. 66–80. <https://doi.org/10.33186/1027-3689-2026-7-66-80>

Введение

В современной научно-исследовательской и коммерческой практике, основанной на экономически оправданных бизнес-моделях, различают две главные концепции создания датасетов.

Формально первую концепцию можно отнести к устоявшейся практике протоколирования и подготовки отчетов по результатам проведения исследований, например, создание набора данных для проверки гипотез, анализа закономерностей, имитации реальных систем [1]. Сюда же можно отнести базы данных и электронные каталоги библиотек, которые формировались и совершенствовались годами, в силу нацеленности библиотечного сообщества на сохранение и популяризацию культурного и научно-исследовательского наследия человечества.

Ко второй группе датасетов можно отнести, во-первых, экономические банки данных (базы данных коммерческих структур, экономические статистические базы данных и т. д.), во-вторых, специально подготовленные наборы данных для решения узких коммерческих задач, например, задач бизнес-аналитики для целей расширения и развития бизнес-моделей, создания конкурентных преимуществ, повышения эффективности продаж продуктов и услуг и так далее [2, 3].

Отдельно можно рассматривать датасеты на стыке двух выше упомянутых концепций, например, уже подготовленные наборы датасетов [4]. В сложившейся мировой практике работы по формированию датасетов рассматривают как некую форму инженерии, которая поддерживает науку (датасеты могут использоваться для исследовательских целей), и современная наука сильно зависит и от инженерии,

и больших данных. Создание датасетов требует пристального внимания к соответствующей области знаний [5].

И в том и другом случае уже сложилась практика формирования датасетов из первичных (не маркированных / не размеченных) данных посредством приведения их к стандартизированному формату, пригодному к применению в задачах машинного обучения. В силу систематической подготовки научно-исследовательского сообщества – и не в последнюю очередь библиотечного – первичные данные, которыми они оперируют, уже являются размеченными. С учетом записей электронных каталогов библиотек и баз данных научных организаций можно сформировать множество датасетов, готовых к применению в машинном обучении. При этом получаемые датасеты соответствуют нормам использования с точки зрения соблюдения положений авторского права. В таблице приведен пример размеченного датасета на основе первичных данных из лог-файла веб-интерфейса ГПНТБ России [6].

Датасет, фрагмент которого приведен в таблице, может быть использован при расчетах альтиметрических показателей востребованности электронных ресурсов [7]. В примере наглядно видны результаты разметки первичных данных, которые в данном случае строятся на принятых правилах разметки базы данных. В представленном датасете демонстрируется время обращения к информационному ресурсу из каталога НТБ ГПНТБ России. Также указывается адрес обращавшегося и количество обращений к конкретному ресурсу. Первая колонка относится к стандартам нумерации самого датасета, `id` – нумерация из базы данных, `org_id` – поле-ключ (идентификатор) организации, `date_time` – дата обращения к записи, `addr` – адрес идентификатор обращавшегося, `bdname` – имя базы данных НТБ, `req` – имя ресурса из базы, к которому обращались, `count` – количество обращений к конкретному ресурсу. Датасет оформлен в формате `parquet` [8], это стандартизированный формат, разработанный для быстрой обработки данных. Например, из приведенного скриншота видно, что наиболее востребованным по количеству обращений является ресурс с библиографическим описанием:

«Крылова Н. П., Левашов Е. Н. Перспективы развития вузовских библиотек в научной информационной среде // Научные и технические библиотеки. 2021. № 8. С. 83–100. <https://doi.org/10.33186/1027-3689-2021-8-83-100>», число обращений к которому составляет 63 980.

Пример размеченного датасета

№	id	org_id	date_time	addr	bdname	req	count
0	1	6	June	gpntb233.spsLnsr.ru	ntb	/subscribe/ntb/ntb96/5/file10.html	27
1	2	6	October	gpntb61.spsLnsr.ru	ntb	/subscribe/ntb/2023/11/NTB11_2023_4.pdf	202
2	3	8	March	ssti.ru	ntb	/subscribe/ntb/2021/8/NTB8_2021_4.pdf	639
3	4	8	March	ssti.ru	ntb	/subscribe/index.php?journal=ntb&year=2009&num=6&art=3	80
4	5	8	March	ssti.ru	ntb	/subscribe/ntb/1994/11/Astapovich_EG_Klassifikatsiya_biblioteknykh_tekhnologicheskikh_protessov.pdf	153
5	6	8	June	ssti.ru	ntb	/subscribe/index.php?journal=ntb&year=2007&num=10&art=7	377
6	7	8	June	ssti.ru	ntb	/subscribe/index.php?journal=ntb&year=2007&num=10&art=7	36
7	8	8	June	ssti.ru	ntb	/subscribe/index.php?journal=ntb&year=2007&num=10&art=7	36
8	9	8	September	ssti.ru	ntb	/subscribe/ntb/2023/3/NTB3_2023_5.pdf	36
9	10	30	March	h180.net32.bnsturu	ntb	/subscribe/ntb/ntb96/5/file10.html	171
10	11	30	June	h180.net32.bnsturu	ntb	/subscribe/index.php?journal=ntb&year=2006&num=4&art=4	27
11	12	30	June	h180.net32.bnsturu	ntb	/subscribe/index.php?journal=ntb&year=2008&num=4&art=7	15
12	13	30	August	h180.net32.bnsturu	ntb	/subscribe/ntb/1994/4/Stolyarov_YuN_Grezy_o_natsionalnoy_bibliotekte.pdf	48
13	14	30	September	w243-252.sturu	ntb	/subscribe/ntb/1991/2/Avraeva_YuB_Slovo_o_metodiste.pdf	44
14	15	30	September	w243-252.sturu	ntb	/subscribe/ntb/1991/2/Avraeva_YuB_Slovo_o_metodiste.pdf	134
15	16	30	September	w243-252.sturu	ntb	/subscribe/ntb/1991/2/Avraeva_YuB_Slovo_o_metodiste.pdf	134

Окончание таблицы

№	id	org_id	date_time	addr	bdname	req	count
16	17	30	September	w243-252.stu.ru	ntb	/subscribe/ntb/1991/2/Avraeva_YuB_Slovo_o_metodiste.pdf	134
17	18	30	September	w243-252.stu.ru	ntb	/subscribe/ntb/1991/2/Avraeva_YuB_Slovo_o_metodiste.pdf	134
18	19	30	September	w243-252.stu.ru	ntb	/subscribe/ntb/1995/5/Suslova_IM_Professionalizatsiya_met odicheskoy biblioteknoy deyatel'nosti.pdf	150
19	20	30	September	w243-252.stu.ru	ntb	/subscribe/ntb/1995/5/Suslova_IM_Professionalizatsiya_met odicheskoy biblioteknoy deyatel'nosti.pdf	150
20	21	30	September	w243-252.stu.ru	ntb	/subscribe/ntb/1995/5/Suslova_IM_Professionalizatsiya_met odicheskoy biblioteknoy deyatel'nosti.pdf	150
21	22	30	September	h180.net32.bmstu.ru	ntb	/subscribe/ntb/ntb97/7/FIG7_1.GIF	6
22	23	30	September	h180.net32.bmstu.ru	ntb	/subscribe/ntb/ntb97/7/FIG7_2.GIF	12

В приведенном примере представлены, как указывалось ранее, данные из лог-фалов веб-интерфейса НТБ ГПНТБ России в виде ссылочных данных, времени и количества обращений, что в свою очередь удобно для расчетов альтметрических показателей востребованности ресурсов. Но также датасеты превосходно подходят для структурирования обобщенных знаний в виде авторефератов, аннотаций и так далее, присущих библиографическим описаниям, применяемым в библиотеках. Так сформированные датасеты на основе больших текстовых корпусов из отечественных баз данных и библиотечных электронных каталогов могут быть использованы для обучения больших языковых моделей (LLM) [9, 10].

Практические аспекты формирования и жизненный цикл датасета

Технически методы формирования датасетов для машинного обучения включают сбор данных, аннотирование, фильтрацию и, в некоторых случаях, синтез. Эти методы помогают создать структурированный массив данных, в котором у каждого объекта есть конкретные свойства: признаки, связи между объектами [4].

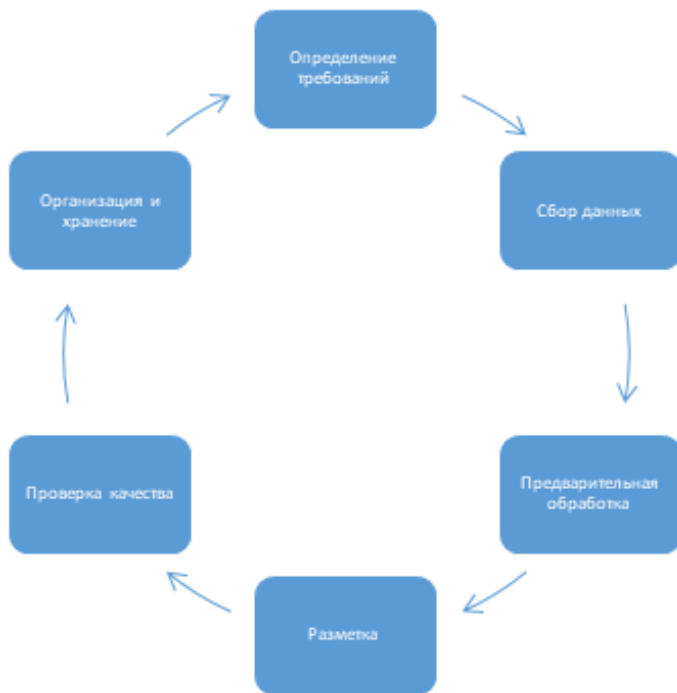
По своей сути датасет есть образ реляционной базы данных. Так называемый адаптер данных (DataAdapter) может выступать в качестве связующего моста между датасетом и источником данных.

Будучи неотъемлемой частью машинного обучения, наборы данных обладают всеми классическими свойствами ИТ-инфраструктуры: они представляют собой общие базы данных с четко определенной архитектурой и стандартизированными интерфейсами, создавая ценность за счет обеспечения быстрой итерации и новых форм процессов, из которых и вытекает их конечная ценность [5].

В современных условиях сложилась практика формирования датасетов путем проведения нескольких этапов:

- 1) определение требований,
- 2) сбор данных,
- 3) предварительная обработка,
- 4) разметка,
- 5) проверка качества,
- 6) организация и хранение.

Эту практику можно охарактеризовать как жизненный цикл датасета. Пример такого жизненного цикла приведен на рисунке. Справедливости ради стоит отметить, что в представленном жизненном цикле не рассматривается период использования датасета по назначению.



Жизненный цикл формирования датасета

Формирование датасета требует от специалиста знаний, умений и навыков обработки, разметки, организации и хранения первичных данных.

Существует еще один источник данных для формирования датасета. У любого ученого за годы профессиональной деятельности набирается целый корпус (называйте каталогом) материалов, исследований и других источников, не только результатов его собственной творческой деятельности. Часто в такой корпус материалов входят текстовые файлы стандартного TXT и PDF-форматов.

Предобработка текстового корпуса для последующего обучения большой языковой модели подразумевает приведение текста к маши-

ночитаемому формату. Другими словами, процедура формирования датасета на основе текстового корпуса состоит из нескольких предварительных этапов. В рамках исследования проводилась процедура предобработки на основе текстового корпуса из каталога НТБ ГПНТБ России. В приведенном корпусе текст представлен в виде текстовых файлов в формате TXT. Процедура предобработки включает в себя выделение токенов из общего корпуса текста, создание словаря, формирование так называемого корпуса train, на основе которого в дальнейшем проводится обучение алгоритма текстовой модели.

Представленное типовое решение предобработки текстового корпуса было реализовано на локальной машине с применением среды разработки WSL + Jupyter-lab, что демонстрирует наглядность реализации подобных решений доступными техническими средствами, без привлечения дорогостоящего оборудования. Ради объективности стоит отметить, что представленная реализация типового решения применялась в качестве механизма самообучения и практического тестирования эффективного алгоритма предобработки текстового корпуса для целей машинного обучения. К слову, представленный алгоритм был выбран эмпирическим путем из множества подобных алгоритмов с открытым кодом.

Для основной массы текущих технических задач достаточно типового алгоритма формирования датасета, с целью последующего его применения при обучении больших языковых моделей. Для большинства больших языковых моделей необходимо привести текстовой корпус к стандартизированному формату представления датасета. Существующая практика предусматривает формирование датасета с тремя основными критериями разбиения/разделения (англ. split). 1. Train – корпус, на основе которого будет производится обучение LLM. 2. Test – тестовый корпус, предназначенный для проверки правильности обучения. 3. Validation – валидационный корпус, на основе которого проводится проверка корректности итоговых результатов. Обычно сформированный датасет содержит корпуса train, test, validation в пропорции 60–20–20% от общего объема датасета. Стоит отметить, что указанные пропорции не являются требованиями к исполнению и в большинстве случаев зависят от технической реализации и задач, для которых формируется датасет. На основе уже рассмотренного текстового корпуса из каталога НТБ ГПНТБ России тех-

нически сформирован датасет. Приблизительно, алгоритм формирования датасета состоит из четырех последовательных шагов:

1. *Шаг технического формирования датасета.*

```
import pandas as pd
import os

file_names=os.listdir('...../NTB_dataset/dataset/')
doc={}
for file in file_names:
    with open('..../NTB_dataset/dataset/' +file, 'r') as target_file:
        doc[file]=target_file.read()
        ntb=(pd.DataFrame.from_dict(doc, orient='index')
            .reset_index().rename(index=str, columns={'index': 'file_name', 0:
'text'})))

NTB_train=ntb.sample(frac=0.7, random_state=42)
NTB_tmp=ntb.drop(NTB_train.index)
NTB_test=NTB_tmp.sample(frac=0.33333, random_state=42)
NTB_valid=NTB_tmp.drop(NTB_test.index)
del NTB_tmp

NTB_train.shape, NTB_test.shape, NTB_valid.shape
```

((416, 2), (60, 2), (119, 2)) – демонстрация пропорций в итоговом датасете

2. *Шаг сохранения результата в хранилище.*

```
NTB_test.to_parquet('...T/NTB_dataset/NTB3_test.parquet', en-
gine='pyarrow')
NTB_train.to_parquet('.../NTB_dataset/NTB3_train.parquet', en-
gine='pyarrow')
NTB_valid.to_parquet('.../NTB_dataset/NTB3_valid.parquet', en-
gine='pyarrow')
```

3. *Шаг формирования результирующего датасета.*

```
from datasets import load_dataset
```

```

data_files={'train': '...../NTB_dataset/NTB3_train.parquet', 'test':
'...../NTB_dataset/NTB3_test.parquet', 'validation':
'...../NTB_dataset/NTB3_valid.parquet'}
dataset=load_dataset('parquet', data_files=data_files)
dataset

```

```

DatasetDict({
  train: Dataset({
    features: ['file_name', 'text', '__index_level_0__'],
    num_rows: 416
  })
  test: Dataset({
    features: ['file_name', 'text', '__index_level_0__'],
    num_rows: 60
  })
  validation: Dataset({
    features: ['file_name', 'text', '__index_level_0__'],
    num_rows: 119
  })
})

```

4. Шаг сохранения датасета.

```
dataset.save_to_disk('...../GIT/NTB_dataset/pre')
```

К слову, 2-й и 4-й шаги – это два разных по значению и содержанию этапа. В случае со 2-м шагом происходит сохранение промежуточного результата, для целей последующей его обработки. Это продемонстрировано на шаге 3. На 4-м шаге происходит сохранение датасета на локальном диске.

Из представленного листинга (см. шаг 3) видно итоговое пропорциональное соотношение внутри датасета. Готовый датасет можно посмотреть по адресу [11].

Заключение

Любые методы, технологии или подходы работы с датасетами помогают сделать первичные данные пригодными для анализа, библиометрических или альтметрических расчетов показателей, анализа и реферирования больших объемов текста и т. д.

В статье представлены основные приемы работы с датасетами. Приведенный в статье листинг примечателен в качестве примера, как данная техническая работа может быть полезна в библиотеке для совершенствования внутренней библиотечной информационно-поисковой системы. Сформулировано определение датасета, гласящее, что датасет – это структурированный набор данных, в котором данные организованы для решения тех или иных задач, к тому же структура датасета, в свою очередь, конструктивно зависима от типа, характера и формализованности первичных данных.

В рамках концепции библиотечно-информационной деятельности подразумевается, что специалисты должны уметь работать с первичными данными и, соответственно, уметь организовывать и хранить датасеты. По факту получается, что современная библиотека уже обладает достаточной и емкой компетенцией для подобного рода деятельности. Функциональная нагрузка библиотекарей и самой библиотеки, предполагающая работу с источниками, публикациями, изданиями – в общем, информационными ресурсами различного характера, их библиографическое описание, реферирование, аннотирование и так далее, позволяет утверждать: вся эта работа ведется с первичными данными. Другими словами, в фондах библиотек уже имеются готовые датасеты, различного масштаба и по различным тематикам. К тому же, в силу сложившейся практики, каждая библиотека готовит библиографические сборники, которые можно расценивать как объекты авторского права, и библиографические справочники, являясь набором первичных данных, уже технически и юридически подготовлены для организации и хранения датасета.

Список источников

1. **Что** такое датасет и как его использовать. URL: <https://thecode.media/chto-takoe-dataset-i-chto-s-nim-delayut/> (дата обращения: 06.11.2025).
2. **Что** такое датасеты и как их использовать для анализа данных. URL: <https://productstar.ru/blog/datasety-chto-eto-vidy-i-kak-ispolzovat-dlya-analiza-dannykh> (дата обращения: 06.11.2025).
3. **Zhang J. et al.** An open dataset of EurekAlert! press releases for science communication research // Scientific Data. 2025. Vol. 12. № 1. P. 1606. DOI 10.1038/s41597-025-05865-1.

4. **Для** чего аналитику данных датасет и где его взять (Практикум / Блог Яндекс). URL: <https://practicum.yandex.ru/blog/dataset-dlya-mashinnogo-obucheniya-i-analiza> (дата обращения: 11.08.2025).
5. **Hutchinson B. et al.** Towards accountability for machine learning datasets: Practices from software engineering and infrastructure // Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021. P. 560–575. DOI 10.1145/3442188.3445918.
6. **Датасет** из лог-файла веб-интерфейса НТБ ГПНТБ России. URL: <https://disk.yandex.ru/d/6hPLWBvOWOX7Lw> (дата обращения: 07.04.2026).
7. **Земсков А. И., Колосов К. А.** Библиометрия в библиотеках // Научные и технические библиотеки. 2016. № 11. С. 5–23.
8. **Описание** формата parquet. URL: <https://datafinder.ru/products/chto-takoe-parquet-obzor-osnovnyh-preimushchestv-i-sluchaev-ispolzovaniya?ysclid=mf558b144740088306> (дата обращения: 04.09.2025).
9. **Двойникова А. А., Карпов А. А.** Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных // Информационно-управляющие системы. 2020. № 4. С. 20–30. DOI 10.31799/1684-8853-2020-4-20-30.
10. **Коллекция** источников elibrary.ru. URL: <https://disk.yandex.ru/d/62LSxukdaXPioA> (дата обращения: 02.09.2025).
11. **Готовый** датасет из листинга. URL: <https://huggingface.co/datasets/vahrush/pre4> (дата обращения: 16.04.2026).
12. **Смирнова О. С., Шишков В. В.** Выбор топологии нейронных сетей и их применение для классификации коротких текстов // International Journal of Open Information Technologies. 2016. Т. 4. № 8. С. 50–54.
13. **Стратегия** развития информационного общества в Российской Федерации на 2017–2030 годы. URL: <https://www.garant.ru/products/ipo/prime/doc/71570570/?ysclid=mf57f2m1f3852664316> (дата обращения: 04.09.2025).
14. **Rogers A. et al.** RuSentiment: An enriched sentiment analysis dataset for social media in Russian // Proceedings of the 27th international conference on computational linguistics. 2018. P. 755–763.
15. **Полное** руководство по библиотеке Gensim для начинающих. URL: <https://pythonru.com/biblioteki/gensim?ysclid=mgylzf9l2n931144963> (дата обращения: 20.04.2026).
16. **Михальченкова Н. А.** Инфраструктура научно-технической информации России: систематизация видов научно-технической информации и научно-технической литературы в контексте федерального проекта по развитию научно-технических библиотек // Научные и технические библиотеки. 2026. № 6. С. 15–30. <https://doi.org/10.33186/1027-3689-2026-6-15-30>.
17. **Свидетельство** о государственной регистрации базы данных № 2026621985. База данных «Датасет НТБ ГПНТБ России. Вер. 0.2» : № 2026621498 : заявл. 09.04.2026 : опубл. 30.04.2026 / М. В. Вахрушев, М. В. Гончаров ; правообладатель Федеральное государственное бюджетное учреждение «Государственная публичная научно-техническая библиотека России» (RU). 1 с.

References

1. **Chto** takoe dataset i kak ego ispol'zovat'. URL: <https://thecode.media/chto-takoe-dataset-i-chto-s-nim-delayut/> (data obrashcheniia: 06.11.2025).
2. **Chto** takoe datasety` i kak ikh ispol'zovat` dlia analiza danny`kh. URL: <https://productstar.ru/blog/datasety-chto-eto-vidy-i-kak-ispolzovat-dlya-analiza-dannykh> (data obrashcheniia: 06.11.2025).
3. **Zhang J. et al.** An open dataset of EurekAlert! press releases for science communication research // Scientific Data. 2025. Vol. 12. № 1. P. 1606. DOI 10.1038/s41597-025-05865-1.
4. **Dlia** chego analitiku danny`kh dataset i gde ego vziat` (Praktikum / Blog Yandex). URL: <https://practicum.yandex.ru/blog/dataset-dlya-mashinnogo-obucheniya-i-analiza> (data obrashcheniia: 11.08.2025).
5. **Hutchinson B. et al.** Towards accountability for machine learning datasets: Practices from software engineering and infrastructure // Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021. P. 560–575. DOI 10.1145/3442188.3445918.
6. **Dataset** iz log-fai`la veb-interfei'sa NTB GPNTB Rossii. URL: <https://disk.yandex.ru/d/6hPLWBvOWOX7Lw> (data obrashcheniia: 07.04.2026).
7. **Zemskov A. I., Kolosov K. A.** Bibliometriia v bibliotekakh // Nauchny`e i tekhnicheskie biblioteki. 2016. № 11. S. 5–23.
8. **Opisanie** formata parquet. URL: <https://datafinder.ru/products/chto-takoe-parquet-obzor-osnovnyh-preimushchestv-i-sluchaev-ispolzovaniya?ysclid=mf558b144740088306> (data obrashcheniia: 04.09.2025).
9. **Dvoi`nikova A. A., Karpov A. A.** Analiticheskii` obzor podhodov k raspoznavaniu tonal'nosti russkoiazyc'nykh tekstovy`kh danny`kh // Informatcionno-upravliaiushchie sistemy`. 2020. № 4. S. 20–30. DOI 10.31799/1684-8853-2020-4-20-30.
10. **Kollektsiia** istochnikov elibrary.ru. URL: <https://disk.yandex.ru/d/62LSxukdaXPioA> (data obrashcheniia: 02.09.2025).
11. **Gotovy`i`** dataset iz listinga. URL: <https://huggingface.co/datasets/vahrush/pre4> (data obrashcheniia: 16.04.2026).
12. **Smirnova O. S., Shishkov V. V.** Vy`bor topologii nei`ronny`kh setei` i ikh primenenie dlia klassifikatsii korotkikh tekstov // International Journal of Open Information Technologies. 2016. T. 4. № 8. S. 50–54.
13. **Strategiia** razvitiia informatcionnogo obshchestva v Rossii`skei` Federatsii na 2017–2030 gody`. URL: <https://www.garant.ru/products/ipo/prime/doc/71570570/?ysclid=mf57f2m1f3852664316> (data obrashcheniia: 04.09.2025).
14. **Rogers A. et al.** RuSentiment: An enriched sentiment analysis dataset for social media in Russian // Proceedings of the 27th international conference on computational linguistics. 2018. P. 755–763.

15. **Polnoe** rukovodstvo po biblioteke Gensim dlia nachinaiushchikh.
URL: <https://pythonru.com/biblioteki/gensim?ysclid=mgyzlf9l2n931144963>
(data obrashcheniia: 20.04.2026).
16. **Mihal'chenkova N. A.** Infrastruktura nauchno-tekhnicheskoi` informatsii Rossii: sistematizatsiia vidov nauchno-tekhnicheskoi` informatsii i nauchno-tekhnicheskoi` literatury` v kontekste federal'nogo proekta po razvitiu nauchno-tekhnicheskikh bibliotek // Nauchny'e i tekhnicheskie biblioteki. 2026. № 6. S. 15–30. <https://doi.org/10.33186/1027-3689-2026-6-15-30>.
17. **Svidetel'stvo** o gosudarstvennoi` registratsii bazy` danny'kh № 2026621985. Baza danny'kh «Dataset NTB GPNTB Rossii. Ver. 0.2» : № 2026621498 : zaivl. 09.04.2026 : opubl. 30.04.2026 / M. V. Vakhrushev, M. V. Goncharov ; pravoobladatel' Federal'noe gosudarstvennoe biudzhethnoe uchrezhdenie «Gosudarstvennaia publichnaia nauchno-tekhnicheskaia biblioteka Rossii» (RU). 1 s.

Информация об авторах / Authors

Вахрушев Максим Васильевич – канд. пед. наук, старший научный сотрудник группы перспективных исследований и аналитического прогнозирования ГПНТБ России, Москва, Российская Федерация
vahrushev@gpntb.ru

Гончаров Михаил Владимирович – канд. техн. наук, доцент, ведущий научный сотрудник, руководитель группы перспективных исследований и аналитического прогнозирования ГПНТБ России, Москва, Российская Федерация
goncharov@gpntb.ru

Колосов Кирилл Анатольевич – канд. техн. наук, программист Ассоциации ЭБНИТ, Москва, Российская Федерация
kolosov@gpntb.ru

Maxim V. Vakhrushev – Cand. Sc. (Pedagogy), Senior Researcher, Perspective Research and Analytical Forecast Group, Russian National Public Library for Science and Technology, Moscow, Russian Federation
vahrushev@gpntb.ru

Mikhail V. Goncharov – Cand. Sc. (Engineering), Associate Professor, Leading Researcher, Head, Perspective Research and Analytical Forecast Group, Russian National Public Library for Science and Technology, Moscow, Russian Federation
goncharov@gpntb.ru

Kirill A. Kolosov – Cand. Sc. (Engineering), programmer, ELNIT Association, Moscow, Russian Federation
kolosov@gpntb.ru