Сохранение сайтов раннего интернета для междисциплинарных исследований на примере сайтов хостинга «Narod.ru» (2000–2013)

А. С. Козлова, И. В. Асланов, И. В. Бибилов, Е. В. Котельников

Европейский университет в Санкт-Петербурге

annakozlova@eu.spb.ru, iaslanov@eu.spb.ru, ibibilov@eu.spb.ru, kotelnikov.ev@gmail.com

Аннотация

Статья об исследовании сохранения и изучения сайтов хостинга «Narod.ru», активно функционировавшего в 2000–2013 гг. Авторы рассматривают сайты хостинга как исчезающие объекты цифрового наследия, сохранение и анализ которых может быть интересен экспертам разных предметных областей, в особенности культурологам и исследователям цифрового фольклора раннего интернета.

Данное исследование проводилось на случайно сформированной выборке в 10 тыс. сайтов из 600 тыс. сайтов хостинга. Для полученной выборки были сохранены скриншоты главных страниц и более 400 тыс. страниц сайтов в формате HTML. В дополнение сохранили более 2 млн отдельных файлов изображений документов, презентаций и таблиц, размещенных на сайтах выборки, и метаинформацию данных файлов. Эти данные собраны в проиндексированную базу с полнотекстовым поиском.

Авторы провели анализ собранных данных. Одним из направлений стало изучение языкового разнообразия сайтов хостинга, в ходе которого помимо русскоязычных были определены сайты на английском, украинском, сербском, болгарском, узбекском и других языках.

Важная часть работы — тематическое моделирование, которое позволило выделить тематические кластеры. «Narod.ru» содержал ресурсы различной направленности: сайты образовательных учреждений, о спорте, видеоиграх, животных. Тематическое моделирование определило наиболее популярные сайты на хостинге.

В статье приведены перспективы развития исследования и сделана оценка ограничений данных, в том числе связанных с этическими и правовыми аспектами, которые на данный момент могут являться препятствием для предоставления открытого доступа к данным.

Ключевые слова: ранний интернет, архивирование веб-сайтов, цифровое наследие, анализ данных, narod.ru, BERTopic, большие языковые модели

Библиографическая ссылка: Козлова А. С., Асланов И. В., Бибилов И. В., Котельников Е. В. Сохранение сайтов раннего интернета для междисциплинарных исследований на примере сайтов хостинга «Narod.ru» (2000–2013) // Информационное общество: образование, наука, культура и технологии будущего. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23 – 25 июня 2025 г. Сборник научных статей). — СПб: Университет ИТМО, 2025. С. 79-91. DOI: 10.17586/3033-5574-2025-9-79-91.

1. Введение

За 30 лет распространения общедоступного Интернета создавались и изменялись тысячи практик работы с вебом. Именно сейчас работа по сохранению и изучению цифрового фольклора — любительских практик по созданию контента раннего Интернета (в том числе веб-сайтов) [1] — особенно актуальна по нескольким причинам. С каждым годом таких данных становится все меньше. Четверть всех веб-страниц, существовавших с 2013 по 2023 гг., больше не доступны — спустя менее 10 лет после создания [2]. Можно предположить, что для более раннего периода процент исчезнувших сайтов более велик.

Мы можем потерять сайты в любой момент из-за изменения политики владельца хостинга. Так, например, произошло с сайтами «GeoCities», который компания «Yahoo!» закрыла в 2009 г. За несколько месяцев с объявления о закрытии «Archive Team» сумела спасти почти терабайт страниц «GeoCities», а спустя год они выложили данные в открытый доступ как «Geocities.archiveteam.torrent» [3]. Позднее на основе этих данных был создан проект «One Terabyte of Kilobyte Age» [4], который послужил источником исследовательских и художественных работ, включая выставку «Удаленный город» Ричарда Вийгена — проект, визуализирующий сообщества внутри «GeoCities» [5], и книгу «Digital Folklore Reader», ставшую важной отправной точкой для многих инициатив по сохранению цифрового наследия [1].

Как мы можем видеть, риск навсегда потерять материал для изучения цифрового фольклора, который также потенциально может содержать уникальные художественные, литературные, культурные и исторические артефакты, очень высок. Это значит, что сохранение и превращение сайтов в открытую базу данных, доступную для исследователей, — актуальная задача. Материалам раннего интернета может быть задан широкий круг исследовательских вопросов самых разных предметных областей — музеологии, цифровой антропологии, литературы, веб-дизайна и так далее. Собранные данные также могут послужить отличным источником для исследователей истории Интернета.

История развития и значимость «GeoCities» и «Narod.ru» похожа для англоязычных и русскоязычных сегментов Интернета соответственно. Однако, для сайтов хостинга «Narod.ru», на сохранении и изучении которых сосредоточено представленное исследование, до этого момента не было предпринято систематических усилий по сбережению. «Narod.ru» — хостинговый сервис, основанный компанией «Яндекс» в 2000 г. и переданный в 2013 г. в собственность компании «uCoz» [6]. «Narod.ru» предоставлял пользователям возможность бесплатно создавать собственные веб-сайты, как самостоятельно, так и с помощью встроенного конструктора. «Narod.ru» пользовался популярностью среди широкого круга пользователей, на платформе создавались как персональные сайты, так и веб-ресурсы сообществ, бизнес-компаний и государственных организаций, включая школы и детские сады. Хостинговый сервис содержал обширную коллекцию разнообразных документов, включая фотографии, изображения, официальные бумаги и электронные таблицы, которые требуют сохранения и изучения.

Будет справедливо отметить, что проработанность вопроса сохранения цифрового наследия неравномерна для разных регионов, стран и типов наследия. Этот вопрос часто зависит от законодательства, финансовой поддержки институций, готовых взять на себя вопрос сбережения, наличия исследовательских и волонтерских объединений и других обстоятельств. В рамках этого исследования мы затронем несколько важных событий, связанных с проблемами сохранения.

В первую очередь, надо отметить ЮНЕСКО, деятельность которой играет ключевую роль в сохранении цифрового культурного наследия, начиная с принятия в 2003 г. «Хартии о сохранении цифрового наследия» [7]. В Хартии подчеркивается угроза утраты наследия и необходимость конкретных действий по сохранению и обеспечению доступности сохраняемых цифровых объектов, а также перечисляется широкий спектр форматов цифрового наследия, одними из которых являются как веб-сайты и содержащиеся в них

документы, тексты, движущиеся и неподвижные изображения и многое другое. Сайты хостинга «Narod.ru» — полноправная часть цифрового наследия и также находятся под угрозой исчезновения, нуждаются в сохранении.

Одним из самых масштабных проектов, занимающихся сохранением сайтов, по праву можно считать «Wayback Machine» команды «Internet Archive» [8]. Проект занимается сохранением миллиардов страниц в Интернете. Тем не менее, вопрос о степени охвата «Wayback Machine» сайтов, размещённых на хостинге «Narod.ru», требует отдельного рассмотрения. Проверить наличие копии сайта в «Internet Archive» можно при помощи API. Всего из 637 358 сайтов «Narod.ru», предоставленных компанией Яндекс авторам исследования, в «Internet Archive» не сохранен 183 061 сайт, что составляет 28,7 % от всех сайтов. Необходимо учитывать, что в данном случае речь идет лишь о снапшотах главных страниц сайтов — в ходе проверки наличия сохраненных копий для всех сайтов отфильтрованного списка в API подавались лишь адреса главных страниц, поэтому реальное количество сохраненных страниц и файлов для всего списка сайтов будет значительно меньше.

Следует предположить, что русскоязычные инициативы могли бы более эффективно решить вопрос о сохранении артефактов раннего российского сегмента Интернета. В России одной из ключевых в вопросе сохранения и архивации данных Интернета является инициатива АНО «Информационная культура», учрежденная И. В. Бегтиным. Один из проектов АНО «Информационная культура» — Национальный цифровой архив [9]: ресурс по «поиску и сохранению веб-сайтов и иных цифровых материалов, имеющих высокую общественную ценность и находящихся под угрозой уничтожения». Национальный цифровой архив также содержит копии сайтов «Narod.ru», но на данный момент в него включены всего 10 уникальных доменов «Narod.ru».

Таким образом, реализуемый нами проект приобретает особую значимость, поскольку направлен на архивирование и систематическое изучение исчезающего и слабо изученного сегмента раннего Интернета. На сегодняшний день ни одна из крупных международных или российских инициатив, среди которых нами были рассмотрены лишь наиболее значимые, не осуществляет архивирование сайтов хостинга «Narod.ru» в объемах, достаточных не только для качественного, но и для полноценного количественного анализа сохраненных данных.

2. Формирование базы данных

2.1 Сбор данных

Ключевым направлением реализации проекта стало формирование архива веб-сайтов, размещенных на хостинге «Narod.ru», для обеспечения исследователям возможности проводить анализ сохраненных данных. В качестве исходных данных был использован список с доменами третьего уровня сайтов хостинга (например, в исходном списке указано «site» для адреса https://site.narod.ru/), предоставленный компанией «Яндекс»: полный список адресов включал 637 тыс. доменных имён. В связи с большим объемом сайтов, подлежащих сохранению, для создания прототипа проекта была сформирована репрезентативная случайная выборка из 10 тыс. доменных имён. Такой объем выборки обеспечивает допустимую погрешность в пределах 1 % при доверительном интервале 95 %.

В рамках проекта предполагалось решение следующих задач: во-первых, было необходимо сохранить содержимое сайтов из выборки; во-вторых, обработать собранные данные для извлечения из них метаинформации; в-третьих, создать базу данных и веб-интерфейс, предоставляющий удобный доступ к сформированному датасету, включая возможности полнотекстового поиска по страницам и сохраненным документам, а также первичный анализ полученных данных.

Для реализации этапа сбора данных использовался веб-краулер, разработанный на языке программирования Python с применением библиотек Selenium и BeautifulSoup. Выбор инструментов был обусловлен их высокой степенью настраиваемости и низкими требованиями к ресурсам серверной инфраструктуры. Собранная информация была сохранена в базу данных PostgreSQL. HTML-документы и сопутствующие файлы, загруженные с сайтов, были отдельно сохранены на сервере.

В результате работы веб-краулера удалось обработать 9 956 из 10 000 сайтов, что соответствует уровню покрытия 99,6 %. В общей сложности было обработано 532 118 страниц, из которых 401 736 (75,5 %) были успешно сохранены в базу данных; остальные содержали различного рода ошибки. Кроме того, было сохранено более 2,2 миллиона файлов общим объемом 84,3 Гбайт, включая 2 130 401 изображение, 14 543 PDF-документа, 18 196 текстовых файлов, 1 934 электронные таблицы и 2 513 презентаций.

2.2. Разработка структуры данных и веб-интерфейса

Для хранения метаинформации о сайтах хостинга «Narod.ru» была использована система управления базами данных PostgreSQL[1].

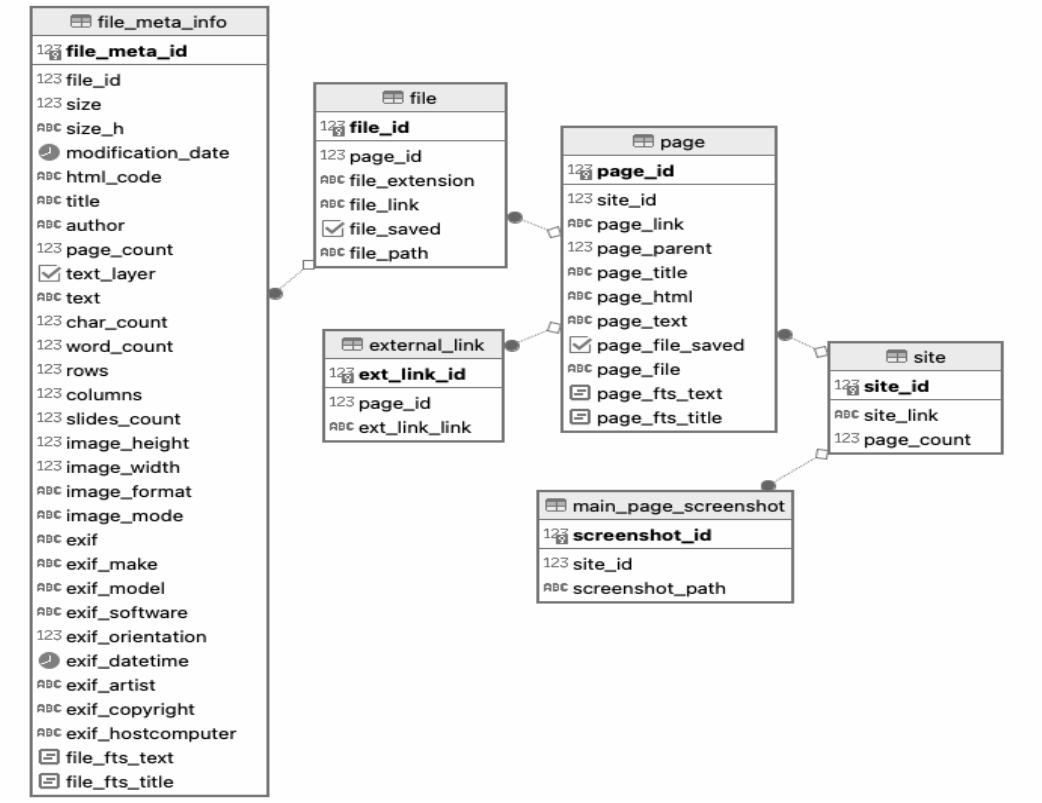

Рис. 1. Схема базы данных проекта по сохранению цифрового культурного наследия сайтов хостинга «Narod.ru»

[1]PostgreSQL // PostgreSQL Global Development Group. URL: https://www.postgresql.org/ (дата обращения: 30.03.2025).

Созданная база данных содержит следующую информацию: метаинформация веб-сайтов, дочерних страниц каждого сайта, скриншотов главных страниц (необходимы для последующей визуализации оригинального вида сохраняемых сайтов), а также файлов, размещенных на соответствующих сайтах.

К сохраненным файлам сайтов хостинга применялся парсер, извлекающий метаинформацию, которая была сохранена в отдельную таблицу базы данных. Наличие метаинформации предоставило возможность исследовать различные характеристики файлов: EXIF-данные изображений или тексты PDF-файлов и документов. Схема базы данных, разработанной в рамках проекта, представлена на рис. 1.

Для повышения производительности и обеспечения возможности проведения полнотекстового поиска в базе данных были дополнительно созданы индексы по заголовкам и текстовому содержимому страниц и файлов. Атрибуты базы данных, содержащие URL-адреса сайтов, отдельных страниц и файлов, были проиндексированы с использованием модуля «pg_trgm», что позволило реализовать механизм поиска по степени сходства строк, повышающий удобство работы с базой данных.

Для обеспечения взаимодействия пользователей с сохраненными данными был разработан веб-интерфейс на основе фреймворка Django. Интерфейс предоставляет возможность выполнения поисковых запросов по собранному датасету, включая фильтрацию данных по различным параметрам. Помимо этого, была реализована возможность полнотекстового поиска по содержимому страниц и документов. В перспективе будет расширение функциональности веб-интерфейса за счёт внедрения средств каталогизации сайтов и усовершенствования механизмов фильтрации, направленных на исключение ресурсов, содержащих нежелательный, вредоносный или запрещенный законодательством контент.

3. Анализ собранных данных

3.1. Изучение языкового разнообразия

Полученный датасет позволяет осуществлять исследования в области культурологии и лингвистики. В частности, в рамках данного проекта был проведен анализ языкового разнообразия сайтов, размещённых на хостинге «Narod.ru». Хостингом пользовались не только в России, но и в странах СНГ, других зарубежных государствах. Например, в выборке представлены такие ресурсы, как сайт Государственной филармонии Департамента культуры г. Астаны[2] и сайт с аудиокнигами, изданными в Беларуси[3].

Определение языка сайтов на основе метаинформации оказалось невозможным ввиду низкой степени её заполненности и наличия недостоверных данных. Поэтому анализ языкового распределения сайтов хостинга был реализован следующим образом: первоначально текстовое содержимое страниц обрабатывалось при помощи библиотеки Polyglot [10], выбор которой обусловлен поддержкой широкого спектра языков, включая языки стран постсоветского пространства, а также языки народов России (например, татарский и чувашский). В случаях, когда язык не мог быть идентифицирован библиотекой

[2]Астана қаласы Мәдениет департаментінің «Мемлекеттік филармония» МКҚК // Сайт государственной филармонии департамента культуры города Астаны. URL: https://filarmoniya-kz.narod.ru/ (дата обращения: 30.03.2025).

[3]Гукавыя кнігі // Сайт аудиокниг, изданных в Республике Беларусь. URL: https://gukkniga.narod.ru/ (дата обращения: 30.03.2025).

Polyglot, применялась библиотека Langdetect[4]. Если и после этого язык не определялся, информация извлекалась непосредственно из доступных метаданных страницы. Результаты анализа распределения языков сайтов представлены на рис. 2.

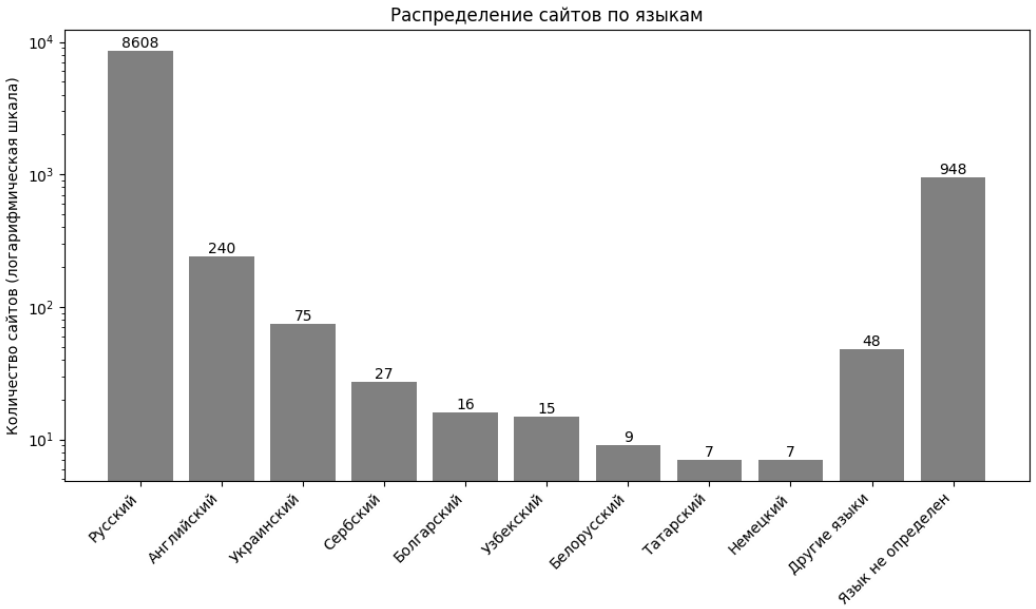

Рис. 2. Языковое разнообразие сайтов хостинга «Narod.ru»

Абсолютное большинство исследованных ресурсов было создано на русском языке. Язык 948 сайтов определить не удалось. Среди других языков, не изображенных на графике, присутствуют азербайджанский и румынский (по четыре сайта); армянский, киргизский, казахский, башкирский и эстонский (по три сайта); еще 18 языков, представленные одним или двумя сайтами.

В рамках дальнейших исследований языкового состава сайтов «Narod.ru» планируется расширение методологии анализа, в частности, реализация распознавания языка не только текстового контента, но и текста на изображениях. Отдельной проблемой, требующей внимания, является некорректное определение кодировок страниц. В настоящий момент значительная часть сайтов, идентифицированных как ресурсы на сербском или болгарском языке, а также страницы с неопределённым языком, отображаются некорректно ввиду несоответствия исходных кодировок текстов.

3.2 Тематическое моделирование

Одной из ключевых исследовательских задач, решаемых в рамках анализа сайтов хостинга «Narod.ru», является определение тематического состава сохраненных ресурсов. Сайты на хостинге характеризовались значительной тематической гетерогенностью: от личных страниц и ресурсов, посвященных популярным сериалам и мультфильмам, до сайтов государственных учреждений. Для выявления тематического разнообразия была выполнена процедура тематического моделирования с применением больших языковых моделей, что позволило выявить более двадцати отдельных тематических кластеров.

[4] langdetect // GitHub. URL: https://github.com/Mimino666/langdetect (дата обращения: 30.03.2025).

В качестве основного инструмента тематического моделирования был использован фреймворк BERTopic [11], выбранный благодаря высокой степени кастомизации и эффективности при обработке многоязычных текстовых данных. BERTopic предполагает поэтапное выполнение тематического моделирования: генерацию эмбеддингов текстов, уменьшение размерности, кластеризацию и репрезентацию сформированных тематических кластеров.

На начальном этапе подготовки данных для моделирования были отобраны главные страницы сайтов выборки, содержащие не менее 50 символов, что составило 9 466 сайтов. Тексты были очищены от стоп-слов и лемматизированы с использованием библиотек Pymorphy3[5], NLTK [12] и Fast-langdetect[6]. Предобработанный текст был необходим для последнего этапа моделирования — репрезентации тематических кластеров. Для получения эмбеддингов текстов страниц была использована многоязычная модель Multilingual-E5-large [13], основанная на архитектуре Transformer [14].

В качестве изначального алгоритма кластеризации в рамках анализа данных был выбран DBSCAN [15]. Однако, кластеризация сайтов данным методом не дала удовлетворительных результатов: DBSCAN определил 43 % сайтов выборки (более 4 тыс. сайтов) как выбросы. В связи с этим было принято решение использовать алгоритм k-средних [16], оптимальное число кластеров для которого определялось при помощи коэффициента силуэта [17]. Анализ данной метрики показал, что 34 кластера являются оптимальным числом для выборки сайтов проекта (см. рис. 3).

Рис. 3. Метрика коэффициента силуэта для метода k-средних

Среди полученных кластеров 4 были выделены как выбросы, включающие сайты с техническими ошибками и неправильными кодировками. Эти кластеры насчитывали 360 сайтов, что составляет 3,8 % от выборки.

На следующем этапе для улучшения репрезентации тематик была проведена процедура объединения схожих кластеров при помощи большой языковой модели компании OpenAI

[5] pymorphy3 // GitHub. URL: https://github.com/no-plagiarism/pymorphy3 (дата обращения: 30.03.2025).

[6] fast-langdetect // GitHub. URL: https://github.com/LlmKira/fast-langdetect (дата обращения: 30.03.2025).

o3-mini-2025-01-31[7]. Применение языковой модели позволило минимизировать субъективность человеческого фактора и дало возможность оценить применимость языковых моделей для выполнения таких задач. В модель был передан JSON-файл, содержащий ранжированный список из 50 ключевых слов для каждого кластера, исходя из меры c-TF-IDF [11]. Это позволило сократить итоговое количество кластеров с 30 до 25 путем объединения таких тем как личные страницы и сайты по продаже товаров. Дополнительно модель присвоила кластерам названия, соответствующие их тематике. Итоговая кластеризация представлена на рис. 4 и 5, показывающих количественное распределение сайтов и наиболее значимые слова каждого кластера соответственно.

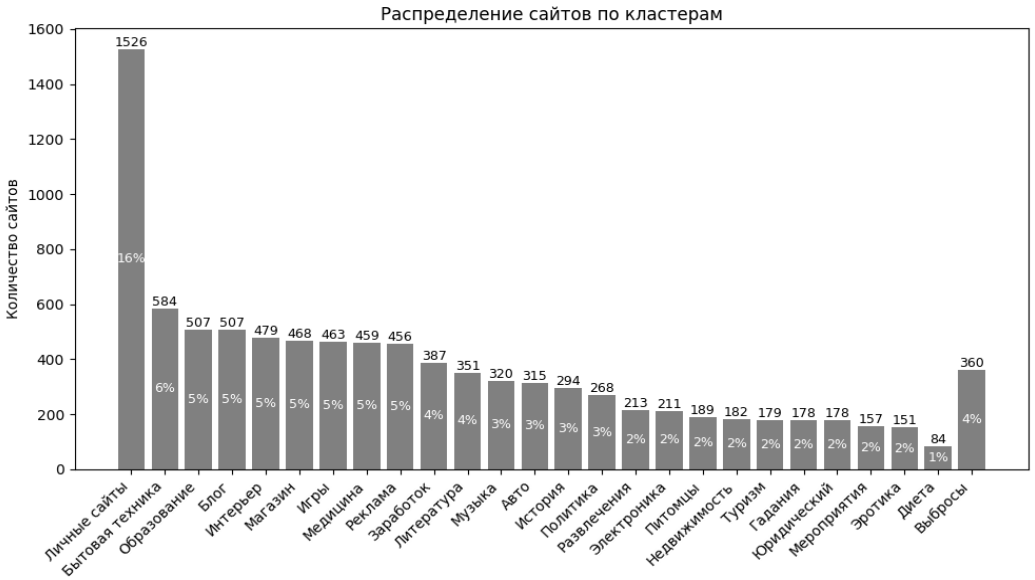

Рис. 4. Распределение сайтов хостинга «Narod.ru» по 25 тематическим кластерам

Полученные результаты демонстрируют, что применение BERTopic в сочетании с языковыми моделями является эффективным для задач эксплораторного анализа данных сайтов хостинг «Narod.ru». Коэффициент силуэта после объединения кластеров составил 0,38, что указывает на наличие умеренно выраженных кластеров в выборке, что также подтверждается визуализацией близости эмбеддингов в двумерном пространстве (рис. 6).

Языковая модель o3-mini продемонстрировала удовлетворительные результаты, корректно выделив тематические группы, такие как продажа товаров или сайты с большим числом страниц.

Вместе с тем проведенный анализ указывает на возможности дальнейшего улучшения процедуры тематического моделирования. Так, например, в случае объединения сайтов личных страниц, стоит отметить, что хотя в кластере и наблюдается большое количество персональных сайтов, можно встретить и сайты, не соответствующие данной тематике. Также возникают вопросы к точности присвоения названий некоторым кластерам: например, кластер «Диета» включает ресурсы, связанные с вредными привычками, такими как курение, а кластер «Реклама» в том числе состоит из сайтов компаний и страниц с торговлей различными товарами.

---

[7]OpenAI O3 Mini // OpenAI. URL: https://openai.com/index/openai-o3-mini/ (дата обращения: 30.03.2025).

Рис.5. Ключевые слова для тематических кластеров сайтов хостинга «Narod.ru»

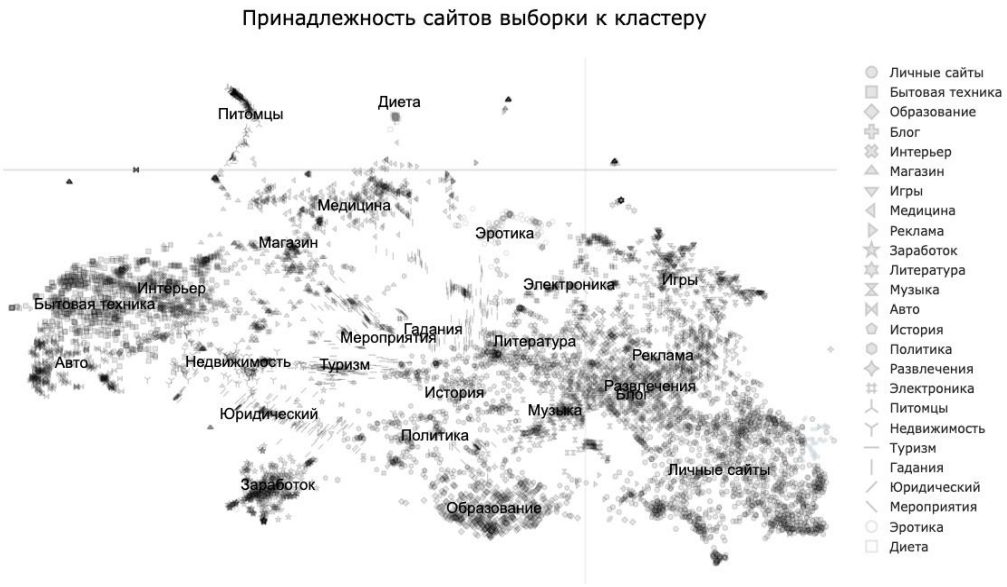

Рис. 6. Визуализация сайтов хостинга «Narod.ru» в двумерном пространстве

Таким образом, дальнейшие исследования должны быть направлены на уточнение тематического моделирования и оптимизацию подходов к интерпретации результатов с применением языковых моделей.

4. Ограничения и перспективы

Основные цели представленного проекта — максимально полное сохранение сайтов, размещённых на хостинге «Narod.ru», проведение комплексного анализа собранных данных и обеспечение удобного доступа к материалам различных исследователей, включая антропологов, культурологов и историков. На текущем этапе работы удалось сохранить в созданной базе данных содержимое выборки, включающей 10 тыс. сайтов, а также значительный объём сопутствующих файлов (более 2 миллионов). Существенным компонентом исследования стала оценка масштаба потерь и ограничений данных, анализ их содержимого: распределение языков и тематическое моделирование, разработка прототипа веб-интерфейса для дальнейшего исследования данных.

Вместе с тем проект сталкивается с рядом ограничений, наиболее существенный из которых — невозможность предоставления собранного датасета в открытый доступ. Это обусловлено тем, что часть сохраненных ресурсов может содержать нежелательный или запрещенный законодательством Российской Федерации контент. Полноценная фильтрация такого контента представляет серьезный вызов для участников проекта. В настоящее время ведется разработка алгоритмов выявления сайтов, содержащих потенциально проблемные материалы, с использованием методов машинного обучения. Однако, на данном этапе, открытая публикация данных невозможна.

Несмотря на это, в ближайшей перспективе запланировано масштабирование процессов сохранения и обработки сайтов полного списка ресурсов хостинга «Narod.ru», совершенствование интерфейса, привлечение специалистов различных предметных областей для реализации исследовательских, художественных и образовательных проектов, которые могут способствовать сохранению и изучению наследия сайтов «Narod.ru» даже без публичного доступа к данным.

Авторы статьи благодарят Наталью Горбачеву, Александру Горваль, Наталью Дмитриеву, Михаила Котова, Татьяну Максимову за неоценимый вклад в развитие концепции, а также разработку кода по парсингу, сохранению и анализу данных, которые были созданы в рамках полугодового проекта[8] ставшего первым шагом для развития данного исследования.

Литература

[1] Digital Folklore: to computer users, with love and respect / eds. O. Lialina [et al.]. Stuttgart: Merz & Solitude, 2009. 286 p.
[2] Chapekis A. [et al.] When Online Content Disappears // Pew Research Center. 17 May. 2024. URL: https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/ (дата обращения: 29.03.2025).
[3] Lialina O. Ruins and Templates of Geocities // Still there. URL: https://contemporary-home-computing.org/still-there/geocities.html (дата обращения: 29.03.2025).
[4] Lialina O., Espenschied D. One terabyte of kilobyte age. // Tumblr. URL: https://oneterabyteofkilobyteage.tumblr.com/ (дата обращения: 29.03.2025).
[5] Vijgen R. The Deleted City: A Digital Archaeology // Parsons Journal for Information Mapping. URL: http://piim.newschool.edu/journal/issues/2013/02/pdfs/ParsonsJournalForInformationMapping_Vijgen_Richard.pdf (дата обращения: 29.03.2025).
[6] Переезд сайтов narod.ru на платформу uCoz // Официальный блог uCoz. URL: https://blog.ucoz.ru/blog/pereezd_sajtov_narod_ru_na_platformu_ucoz/2013-01-31-255 (дата обращения: 29.03.2025).
[7] Charter on the Preservation of Digital Heritage // UNESCO. 15 October 2003. URL: https://www.unesco.org/en/legal-affairs/charter-preservation-digital-heritage (дата обращения: 29.03.2025).
[8] Internet Archive: About IA // Internet Archive. URL: https://archive.org/about/ (дата обращения: 29.03.2025).
[9] Задачи. Russian national digital archive (ruarxive.org) // Национальный цифровой архив. URL: https://ruarxive.org/kb/volunteers/volunteers-tasks (дата обращения: 30.03.2025).
[10] Chen Y., Skiena S. Building Sentiment Lexicons for All Major Languages // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) ACL 2014 / eds. K. Toutanova, H. Wu. Baltimore, MD: Association for Computational Linguistics, 2014. P. 383-389.
[11] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv, 2022. URL: https://arxiv.org/abs/2203.05794 (дата обращения: 29.03.2025).
[12] Bird S., Klein E., Loper E. Natural language processing with Python. Sebastopol, CA: O'Reilly Media Inc., 2009. 479 p.
[13] Wang L. [et al.]. Multilingual E5 Text Embeddings: A Technical Report // arXiv, 2024. URL: https://arxiv.org/abs/2402.05672 (дата обращения: 29.03.2025).
[14] Vaswani A. [et al.]. Attention Is All You Need // arXiv, 2017. URL: https://arxiv.org/abs/1706.03762 (дата обращения 29.03.2025).
[15] Ester M., Kriegel H.-P.m, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining / eds. E. Simoudis, J. Han, U. Fayyad. Portland, OR: AAAI Press, 1996. P. 226-231.

[8] Достояние Naroda: дата-археология 2000-х // Narod и цифровое наследие URL: https://projects.pandan.eusp.org/narod (дата обращения: 19.05.2025).

[16]Lloyd S. P. Least squares quantity in PCM // Information Theory, IEEE Transactions on. 1982. Vol. 28., no. 2, P. 129-137.
[17]Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53-65.

Preservation of Early Internet Sites for Interdisciplinary Research on the Example of the «Narod.ru» (2000–2013)

A. Kozlova, I. Aslanov, I. Bibilov, E. Kotelnikov

European University at St. Petersburg

This study examines the preservation and analysis of websites hosted on Narod.ru, a major Russian free web hosting platform active from 2000 to 2013. The authors consider Narod.ru sites as vanishing digital heritage artifacts of significant value for researchers in cultural studies, digital folklore, and internet history. The research involved a random sample of 10,000 sites from the total of approximately 600,000 Narod.ru websites. The authors preserved screenshots of homepage interfaces and archived over 400,000 HTML pages, along with more than 2 million individual files, including images, documents, presentations, and spreadsheets with their metadata. The materials were stored in an indexed database with full-text search capabilities. The study analyzed linguistic diversity, revealing content in multiple languages such as Russian, English, Ukrainian, Serbian, Bulgarian, and Uzbek. Topic modeling using the BERTopic framework identified over twenty distinct thematic groups, showcasing popular interests and internet culture during Narod.ru's operational period. The research acknowledges data limitations, including technical issues like content encoding problems, incomplete metadata, and potentially illegal content. Future research perspectives include expanding data preservation efforts, improving web interfaces for researchers, and employing advanced analysis tools. The study emphasizes the importance of interdisciplinary collaboration for preserving and interpreting early internet culture, with datasets supporting diverse scholarly, artistic, and educational projects.

Keywords: early Internet, website archiving, digital heritage, data analysis, narod.ru, BERTopic, large language models

Reference for citation: A. Kozlova, I. Aslanov, I. Bibilov, E. Kotelnikov Preservation of Early Internet Sites for Interdisciplinary Research on the Example of the «Narod.ru» (2000–2013) // Information Society: Education, Science, Culture and Technology of Future. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). – St. Petersburg: ITMO University, 2025. P. 79-91. DOI: 10.17586/3033-5574-2025-9-79-91.

Reference

[1] Digital Folklore: to computer users, with love and respect / eds. O. Lialina [et al.]. Stuttgart: Merz & Solitude, 2009. 286 p.
[2] Chapekis A. [et al.] When Online Content Disappears // Pew Research Center. 17 May. 2024. URL: https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/ (accessed date: 29.03.2025).
[3] Lialina O., Ruins and Templates of Geocities // Still there. URL: https://contemporary-home-computing.org/still-there/geocities.html (accessed date: 29.03.2025).
[4] Lialina O., Espenschied D. One terabyte of kilobyte age. // Tumblr. URL: https://oneterabyteofkilobyteage.tumblr.com/ (accessed date: 29.03.2025).
[5] Vijgen R. The Deleted City: A Digital Archaeology // Parsons Journal for Information Mapping. URL:

http://piim.newschool.edu/journal/issues/2013/02/pdfs/ParsonsJournalForInformationMapping_Vijgen_Richard.pdf (accessed date: 29.03.2025).
[6] Moving sites narod.ru on the uCoz platform // The official uCoz blog. URL: https://blog.ucoz.ru/blog/pereezd_sajtov_narod_ru_na_platformu_ucoz/2013-01-31-255 (accessed date: 29.03.2025).
[7] Charter on the Preservation of Digital Heritage. // UNESCO. 15 October 2003. URL: https://www.unesco.org/en/legal-affairs/charter-preservation-digital-heritage (accessed date: 29.03.2025).
[8] Internet Archive: About IA // Internet Archive. URL: https://archive.org/about/ (accessed date: 29.03.2025).
[9] Tasks. Russian national digital archive (ruarxive.org) // National Digital Archive. URL: https://ruarxive.org/kb/volunteers/volunteers-tasks (accessed date: 30.03.2025).
[10] Chen Y., Skiena S. Building Sentiment Lexicons for All Major Languages // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) ACL 2014 / eds. K. Toutanova, H. Wu. Baltimore, MD: Association for Computational Linguistics, 2014. P. 383-389.
[11] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv, 2022. URL: https://arxiv.org/abs/2203.05794 (accessed date: 29.03.2025).
[12] Bird S., Klein E., Loper E. Natural language processing with Python. Sebastopol, CA: O'Reilly Media Inc., 2009. 479 p.
[13] Wang L. [et al.]. Multilingual E5 Text Embeddings: A Technical Report // arXiv, 2024. URL: https://arxiv.org/abs/2402.05672 (accessed date: 29.03.2025).
[14] Vaswani A. [et al.]. Attention Is All You Need // arXiv, 2017. URL: https://arxiv.org/abs/1706.03762 (accessed date: 29.03.2025).
[15] Ester M., Kriegel H.-P.m, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining / eds. E. Simoudis, J. Han, U. Fayyad. Portland, OR: AAAI Press, 1996. P. 226–231.
[16] Lloyd S. P. Least squares quantity in PCM // Information Theory, IEEE Transactions on. 1982. Vol. 28., no. 2, P. 129-137.
[17] Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53-65.