оригинальная статья

eLibrary EDN: LBIBQV

Разработка веб-приложения для автоматизации создания корпуса метаданных публикаций в социальных сетях (на материале русскоязычных и англоязычных тревел-блогов о Республике Беларусь)

Красовская Юлия Юрьевна
Белорусский государственный университет, Беларусь, Минск
eLibrary Author SPIN: 3504-5564
hisradaar@gmail.com

Аннотация: В настоящее время одним из ключевых направлений в области анализа социальных медиа является извлечение и обработка метаданных для систематизации сведений о публикациях, авторах и вовлеченности аудитории. Целью исследования является разработка и тестирование веб-приложения OmniTrack, предназначенного для автоматизированного сбора корпуса метаданных русскоязычных и англоязычных публикаций о Республике Беларусь в тревел-блогах и их параметризации. Интеграция методов веб-скрейпинга, многоуровневой архитектуры и модульного подхода обеспечивает масштабируемость, воспроизводимость и расширяемость системы при изменении внешних интерфейсов платформ. Серверная часть реализована на языке Python с использованием фреймворка Flask; для взаимодействия с пользователем создан веб-интерфейс на HTML, CSS и JavaScript. Алгоритмы извлечения данных разработаны как независимые модули: для TikTok применена эмуляция браузера через undetected_chromedriver для обхода динамической отрисовки; для YouTube – библиотека yt_dlp для прямого получения JSON-метаданных; для Instagram[1] – инструмент instaloader, обеспечивающий высокоуровневый доступ к объектной модели публикации. Собранные метаданные приведены к унифицированной схеме с сохранением в формате Excel при помощи библиотеки openpyxl, что обеспечивает удобство последующей статистической обработки. Приложение прошло юзабилити-тестирование: 42 участника обработали более 400 публикаций, оценив простоту установки, скорость работы и интуитивность интерфейса; средняя оценка удобства составила 4,9 балла из 5; выявлены и устранены критические ошибки, включая несовместимость backend-модуля pywebview и некорректную обработку сокращенных ссылок TikTok. Предложенное авторское веб-приложение OmniTrack обеспечивает создание репрезентативного корпуса метаданных, необходимого для последующего анализа дискурсивных, жанровых и коммуникативных особенностей русскоязычных и англоязычных тревел-блогов о Республике Беларусь.

Ключевые слова: веб-приложение, метаданные, социальные сети, веб-скрейпинг, автоматизация сбора данных

Цитирование: Красовская Ю. Ю. Разработка веб-приложения для автоматизации создания корпуса метаданных публикаций в социальных сетях (на материале русскоязычных и англоязычных тревел-блогов о Республике Беларусь). *Виртуальная коммуникация и социальные сети.* 2026. Т. 5. № 2. С. 176–184. https://doi.org/10.21603/vcsn-2026-5-2-176-184

Поступила в редакцию 14.12.2025. Принята после рецензирования 24.03.2026. Принята в печать 30.03.2026.

[1] Компания *Meta Platforms*, владеющая социальными сетями *Facebook* и *Instagram* и онлайн-мессенджером *WhatsApp*, признана экстремистской организацией, ее деятельность запрещена на территории РФ.

original article

Web Application for Automated Metadata Corpus of Social Media Publications: Russian-Language and English-Language Travel Blogs about the Republic of Belarus

Yulija Yu. Krasowskaja
Belarusian State University, Belarus, Minsk
eLibrary Author SPIN: 3504-5564
hisradaar@gmail.com

Abstract: Metadata extraction and processing are crucial for social media analysis as they help to systematize information about publications, authors, and audience engagement. The article introduces the web application OmniTrack, designed for automated collection and parameterization of metadata from Russian-language and English-language travel blogs. The application integrates web-scraping methods with a multi-layer architecture and a modular approach, which provides scalability, reproducibility, and extensibility even with changing external platform interfaces. The backend is implemented in Python (Flask); the frontend utilizes HTML, CSS, and JavaScript for an interactive user experience. Data-extraction algorithms are independent modules: undetected_chromedriver for TikTok's dynamic rendering via browser emulation, yt-dlp for direct JSON-formatted metadata retrieval from YouTube, and Instaloader for high-level access to Instagram's[2] object model. Collected metadata are normalized to a unified schema in Excel format using the Openpyxl library, which facilitates subsequent statistical analysis. The application underwent usability testing: 42 participants processed 400 posts, evaluating installation simplicity, processing speed, and interface intuitiveness. The mean ease-of-use score was as high as 4.9 out of 5. Some critical issues were identified and resolved, including incompatibility of the pywebview backend module and incorrect handling of shortened TikTok links. The OmniTrack web application provides a robust framework for constructing a representative metadata corpus, supporting further linguistic research into the discursive, genre, and communicative features of Russian-language and English-language travel blogs.
Keywords: web application, metadata, social-media, web scraping, data automation

Citation: Krasowskaja Yu. Yu. Web Application for Automated Metadata Corpus of Social Media Publications: Russian-Language and English-Language Travel Blogs about the Republic of Belarus. *Virtual Communication and Social Networks*, 2026, 5(2): 176–184. (In Russ.) https://doi.org/10.21603/vcsn-2026-5-2-176-184

Received 14 Dec 2025. Accepted after review 24 Mar 2026. Accepted for publication 30 Mar 2026.

Введение

Современное развитие цифровых технологий сопровождается стремительным ростом объема информации, распространяемой через социальные сети, которые стали не только пространством для коммуникации и самовыражения пользователей, но и значимыми источниками данных, представляющих интерес для исследователей в области лингвистики, маркетинга, аналитики и разработки программных решений [Chani et al. 2023: 1369; Zachlod et al. 2022: 1070]. Одно из ключевых направлений в области анализа социальных медиа – извлечение и обработка метаданных, позволяющих систематизировать сведения о публикациях, авторах и вовлеченности аудитории [Ohme et al. 2024; Pretorius 2024: 7]. Достижения в области научного приборостроения и вычислительных технологий обеспечивают возможность формирования беспрецедентных по объему и разнообразию массивов данных [Berman et al. 2018: 69]. Эффективность аналитики социальных сетей напрямую зависит от корректности и полноты сбора метаданных, что обусловливает необходимость разработки специализированных инструментов для их параметризации и сохранения в унифицированных форматах [Holom et al. 2020: 377].

Отдельный пласт вопросов связан с правовыми и этическими рамками применения веб-скрейпинга: ограничения, накладываемые платформами

[2] *Meta Platforms*, the parent company of *Facebook*, *Instagram* and *WhatsApp* Messenger, is banned in the Russian Federation as an extremist organization.

на использование API, вынуждают исследователей обращаться к неофициальным методам сбора данных [Brown et al. 2024]. Веб-скрейпинг рассматривается как один из ключевых инструментов вычислительных социальных наук [Жучкова, Ротмистров 2021]. Его привлекательность объясняется возможностью оперативного получения больших объемов данных с минимальными затратами времени и ресурсов. Среди типов данных, наиболее часто извлекаемых с помощью веб-скрейпинга, выделяются текстовые и числовые данные, изображения, мультимедиа, а также сетевые данные, фиксирующие социальные связи и взаимодействия. Современные системы разграничивают классы метаданных (дескриптивные, структурные, административные, технические и эксплуатационные), каждый из которых требует специализированных схем хранения и проверки качества [Park et al. 2010: 175; Yang et al. 2025].

Современные исследования в области управления и контроля качества метаданных демонстрируют необходимость комплексного подхода к их созданию, валидации и интероперабельности, при этом основными проблемами остаются семантическая неоднозначность, различия в стандартах и отсутствие единых руководств по созданию метаданных [Park, Tosaka 2010: 707]. Предлагаемые решения включают разработку единых международных протоколов, обучение специалистов и внедрение инструментов искусственного интеллекта для автоматизации валидации и исправления ошибок в метаданных [Huang et al. 2025; Subramaniam et al. 2021: 5]. Современная практика управления метаданными движется в сторону стандартизации, автоматизации и интеграции человеко-машинных методов для повышения их достоверности, согласованности и аналитической ценности [Skluzacek et al. 2022]. Подчеркивается, что внедрение систем управления метаданными значительно повышает зрелость управления данными, качество данных и прослеживаемость метаданных [Ahire 2025: 88; Díaz de la Paz et al. 2024: 107]. Отсутствие зрелого управления метаданными препятствует успешной реализации больших данных, приводя к дублированию, несогласованности и низкой надежности данных [Yulfitri et al. 2025]. Отдельно необходимо отметить проблему смещения выборки (*sampling bias*), возникающей вследствие особенностей веб-контента [Foerderer 2023].

Целью исследования является разработка и тестирование веб-приложения OmniTrack, предназначенного для автоматизированного сбора корпуса метаданных русскоязычных и англоязычных публикаций о Республике Беларусь в тревел-блогах и их параметризации. Метаданные публикаций на самых популярных платформах в Республике Беларусь Instagram, TikTok и YouTube содержат ключевые параметры, отражающие динамику пользовательской активности и особенности медиапотребления, что делает их ценным ресурсом для лингвистических и прикладных исследований. Разработка собственного веб-приложения OmniTrack обусловлена необходимостью точного соответствия полноте и структурной единообразности собираемых метаданных, стабильности воспроизводимости результатов и гибкости масштабирования под новые платформы, а также предоставления данных в удобном унифицированном формате и устранения зависимости от лицензионного программного обеспечения. Подчеркивается, что правильная параметризация метаданных и структура сбора влияют на качество аналитики [Edara, Pasumansky 2021: 3091; Moreno-Ortiz, García-Gámez 2023: 249]. Принципы FAIR (*Findable, Accessible, Interoperable, Reusable*) в контексте нашего исследования применяются как методологическая основа для унификации и стандартизации метаданных, собираемых на различных социальных платформах [Wilkinson et al. 2016]. Сравнительный анализ с существующими решениями не проводился, поскольку разработанное приложение ориентировано на собственные исследовательские задачи и набор метаданных, не имеющий прямых функциональных аналогов.

Методы и материалы

Авторское веб-приложение OmniTrack основано на объединении средств веб-скрейпинга, алгоритмов структурирования информации и механизмов экспорта данных в табличные форматы в рамках единого программного комплекса OmniTrack, а также применении интегрированной архитектуры, обеспечивающей взаимодействие серверной части, реализованной на языке Python, и пользовательского веб-интерфейса, и предполагает следующие этапы:

1) определение требований к архитектуре веб-приложения, обеспечивающего параметризацию и сохранение метаданных в стандартизированном формате;
2) разработка алгоритмов сбора данных с целевых платформ и их интеграция в программный комплекс;
3) реализация функционала экспорта метаданных в табличный формат для последующей обработки;

 https://doi.org/10.21603/vcsn-2026-5-2-176-184

4) создание веб-интерфейса и интеграция всех модулей в единую систему;
5) юзабилити-тестирование веб-интерфейса приложения;
6) апробация разработанного решения на реальных данных.

Для реализации приложения использовались средства объектно-ориентированного программирования на языке Python, технологии организации табличных структур данных, а также инструменты веб-разработки для построения графического интерфейса на основе HTML. Тестирование и апробация разработанного комплекса проводились с применением эмпирических методов на реальных данных публикаций из Instagram, TikTok и YouTube.

Результаты

Архитектура веб-приложения

Алгоритмы извлечения метаданных реализованы как четыре независимых модуля: три модуля, ориентированные на специфику целевой платформы, и объединяющий модуль, приводящий все данные в табличный формат. Все модули объединены общими принципами:

- единая точка входа через URL (унифицированный указатель ресурса);
- устойчивое извлечение ключевых показателей вовлеченности и атрибутов публикации;
- приведение данных к унифицированной схеме, пригодной для последующего табличного экспорта.

Блок-схема разработанного алгоритма представлена на рисунке 1.

Разработка алгоритмов сбора данных с целевых платформ

В каждом модуле заложены механизмы повышения воспроизводимости и устойчивости к изменчивости интерфейсов. В начале по введенному URL происходит определение платформы и маршрутизация: выбирается один из разработанных нами соответствующих модулей: *scrape_tiktok, scrape_youtube, scrape_instagram*. Функция convert_k_to_number унифицирует записи вида 100K, 1.5M в целочисленный формат. Предусмотрены безопасные ветви для некорректных значений, пустых строк и уже числовых типов. Для аналитики сохраняется машинно-обрабатываемый объект datetime (UTC – всемирное координированное время), а для отображения – локализованная строка, где дата предоставляется в формате ДД.ММ.ГГГГ. Это разделяет представление и вычисления, исключая потери точности при экспорте в Excel. Архитектура разработанного приложения ориентирована на модульность и расширяемость, что позволяет добавлять новые платформы или параметры без перестройки системы, а также обеспечивает устойчивость при изменениях в источниках данных.

Модуль 1. Модуль *scrape_tiktok* адресует две особенности платформы: динамическую отрисовку интерфейса и частичную персонализацию. Для повышения устойчивости используется эмуляция браузера (Chromium, headless) через заданный параметр *undetected_chromedriver*, имитирующий работу реального браузера в скрытом режиме и позволяющий корректно загружать страницы с элементами динамической отрисовки, и автоматическая установка драйвера (*webdriver_manager*). Введенный нами параметр *--lang=en-US* принудительно задает англоязычную локаль интерфейса, что обеспечивает стабильный формат сокращений (K – для тысяч и M – для миллионов) для последующей числовой нормализации. После загрузки страницы по URL предусмотрена искусственная задержка (около 5 секунд) для завершения отрисовки элементов, зависящих от JavaScript. Идентификатор автора извлекается из структуры ссылки (tiktok.com/@**username**), что надежно для стандартных форматов распространения контента. Для параметризации показателей вовлеченности алгоритм читает значения счетчиков через XPath-селекторы, таргетированные на устойчивые атрибуты data-e2e (*like-count, comment-count, share-count* для считывания показателей счетчиков лайков, комментариев и репостов соответственно). Значения упорядочиваются по позиции и приводятся к целым числам. Ввиду динамичности DOM (объектная модель документа – программный интерфейс, позволяющий программе получить доступ к содержимому HTML) дата публикации определяется через анализ исходного кода страницы и поиск параметра createTime, содержащего Unix-метку времени, которая затем переводится в стандартный формат даты.

Модуль 2. Для YouTube применен подход, минимизирующий зависимость от DOM – использование библиотеки yt_dlp, способной получать структурированный объект метаданных через внутренние механизмы платформы в виде JSON-структуры. Заданные параметры (*quiet, skip_download, forcejson, no_warnings*) ориентированы на бесшумный и быстрый режим. При запросе к YouTube вызывается функция extract_info с download=False, которая без скачивания видео

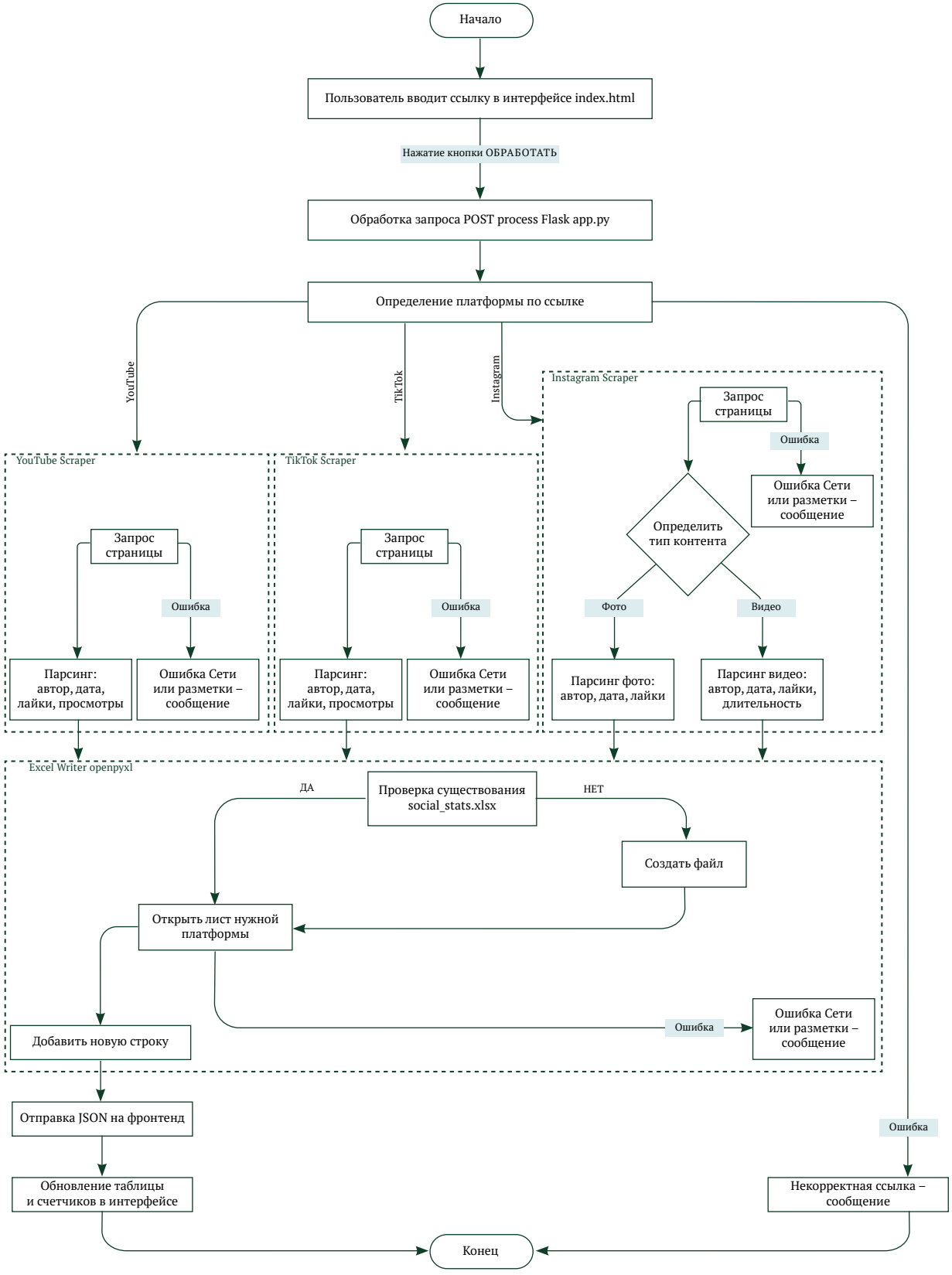

Рис. 1. Блок-схема алгоритма сбора и параметризации метаданных веб-приложения OmniTrack
Fig. 1. OmniTrack web application: metadata collection and parameterization algorithm

 https://doi.org/10.21603/vcsn-2026-5-2-176-184

возвращает словарь с ключевыми полями: извлекаются значения *uploader, title, view_count, like_count, comment_count*.

Модуль 3. Для Instagram использован специализированный инструмент instaloader, предоставляющий высокоуровневый доступ к объектной модели публикации. Это снижает зависимость от нестабильного DOM и упрощает логическое извлечение полей. Режимы download_pictures=False и download_videos=False отключают медиапотоки для прекращения загрузки объекта, а задаваемая нами функция quiet=True обеспечивает детерминированную работу в пакетной обработке. Модуль работает через уникальный идентификатор публикации (shortcode, https://www.instagram.com/p/**1234567890**/), извлекаемый из URL, после чего формируется объект Post через функцию Post.from_shortcode. Возвращаются параметры *owner_profile.full_name, owner_username*, а также *content_type* на основе булева признака *is_video* для отличия публикаций в видеоформате от изображений. Из объекта Post извлекаются значения *likes, comments* и *date_utc*.

Экспорт метаданных в табличный формат

Следующим ключевым компонентом разработанного приложения является модуль сохранения собранных метаданных в табличный формат (табл.). Реализация интегрирующего модуля основана на использовании библиотеки openpyxl, обеспечивающей работу с файлами формата .xlsx. На этапе обработки данные нормализуются и приводятся к единой структуре: автор публикации, дата размещения, количество просмотров (для видеоконтента), число лайков, комментариев, репостов, а также ссылка на исходный пост. Эти параметры фиксируются в виде строк таблицы, где каждый атрибут соответствует отдельному столбцу. Организация хранения предполагает использование разделения по платформам. Даты сохраняются в двух видах: как объект datetime для последующей машинной обработки и как отформатированная строка для визуального отображения. Функционал сохранения также предусматривает обработку ошибок: проверку существования файла, корректное добавление новых данных без перезаписи уже существующих, а также контроль целостности формата. Это позволяет использовать Excel-файл как накопительное хранилище, пригодное для долговременного анализа и экспорта в иные аналитические системы.

Создание веб-интерфейса и интеграция всех модулей в единую систему

Для взаимодействия пользователя с системой реализован веб-интерфейс, обеспечивающий удобный доступ к функционалу без необходимости работы с кодом. Интерфейс разработан на основе HTML, CSS и JavaScript, что позволило создать легкий и совместимый с большинством браузеров пользовательский слой. Интерфейс веб-приложения OmniTrack представлен на рисунке 2.

Ключевой элемент интерфейса – поле ввода ссылки. После нажатия кнопки «Обработать» инициируется запрос к серверной части приложения. Взаимодействие между фронтендом и бэкендом осуществляется через метод fetch() с передачей данных в формате JSON. Сервер возвращает извлеченные метаданные, которые ображаются в таблице на странице. Таблица имеет фиксированную структуру: строки соответствуют обработанным публикациям, а столбцы – параметрам. Особенностью реализации является динамическое обновление таблицы: при вводе новой ссылки данные автоматически добавляются в интерфейс, одновременно фиксируясь в Excel-файле. Таким образом, пользователь видит результат работы алгоритма в реальном времени.

Разработанная нами архитектура интерфейса обеспечивает баланс между функциональностью и простотой использования, позволяя лингвистам

Табл. Пример обработанных метаданных публикаций, сформированных веб-приложением OmniTrack
Tab. Processed publication metadata generated in OmniTrack

Канал	Название видео	Просмотры	Лайки	Комментарии	Дата	Ссылка
Syifa Adriana	TRAVELLING ALONE IN BELARUS – What is Minsk really like? [Ep. 1]	369721	10220	1600	4 декабря 2021 г.	https://www.youtube.com/watch?v=U8FWBa0x40o&ab_channel=SyifaAdriana

Рис. 2. Пользовательский веб-интерфейс приложения OmniTrack
Fig. 2. OmniTrack web application: frontend

сосредоточиться на работе с данными без необходимости углубляться в технические аспекты.

Завершающим этапом разработки веб-приложения для сбора и параметризации метаданных публикаций в русскоязычных и англоязычных тревел-блогах о Республике Беларусь является процесс интеграции всех модулей и формирование единого исполняемого файла. Данный этап имеет принципиальное значение для обеспечения удобства эксплуатации лингвистами, т. к. позволяет использовать приложение без необходимости установки интерпретатора Python и дополнительных библиотек. Для программной интеграции используется серверная основа на Flask, которая обеспечивает маршрутизацию запросов и координацию между фронтендом и бэкендом. Для преобразования исходного кода в исполняемый файл применяется инструмент PyInstaller, который автоматически анализирует зависимости проекта, включая внешние библиотеки, и формирует исполняемый файл, совместимый с операционной системой пользователя.

Юзабилити-тестирование веб-приложения OmniTrack

Для оценки качества пользовательского взаимодействия и выявления потенциальных проблем интерфейса проведено юзабилити-тестирование разработанного веб-приложения OmniTrack для получения обратной связи от представителей целевой аудитории (студентов-лингвистов) относительно удобства, простоты и интуитивности использования программного продукта.

В тестировании приняли участие 42 пользователя, преимущественно студенты 1–5 курсов факультета социокультурных коммуникаций Белорусского государственного университета (специальность «Современные иностранные языки»). Каждому участнику был направлен архив с собранным исполняемым файлом приложения OmniTrack (.exe), а также краткая инструкция по установке и запуску. Архив был адаптирован для запуска на персональных компьютерах с операционной системой Windows без необходимости установки дополнительных зависимостей, что позволило проверить корректность работы в условиях, максимально приближенных к реальным пользовательским сценариям.

Тестирование проводилось в индивидуальном порядке, т. е. студенты запускали приложение на собственных компьютерах и выполняли серию заданий:

1) ввод ссылок на публикации в TikTok, YouTube и Instagram;
2) проверка корректности отображения метаданных (автор, дата публикации, количество лайков, просмотров, комментариев, репостов);
3) проверка работы функции сохранения в Excel и последующего открытия файла;
4) оценка визуального оформления интерфейса (читаемость таблицы, удобство кнопок и поля ввода).

Для полноты эксперимента каждому участнику предложено обработать 5–10 публикаций из различных социальных сетей, что в совокупности обеспечило более 400 протестированных ссылок. После завершения тестирования участники заполняли специально разработанную Google-форму для юзабилити-оценки. Получены следующие результаты:

- 95 % респондентов отметили, что установка и первый запуск не вызвали затруднений; средняя оценка интуитивности интерфейса составила 4,9 по пятибалльной шкале;
- 88 % пользователей отметили приемлемое время обработки одной ссылки (до 10 секунд для TikTok, до 5 секунд для YouTube и Instagram); общая оценка веб-приложения составила 4,9 по пятибалльной шкале.

Выявленные ошибки и их устранение

В процессе юзабилити-тестирования разработанного веб-приложения OmniTrack зафиксированы две основные проблемы.

Первая проблема связана с запуском исполняемого файла при использовании PyInstaller и PyWebView, т. к. у части пользователей при первом запуске скомпилированного исполняемого файла возникало аварийное завершение программы. Анализ показал, что библиотека pywebview, применяемая

https://doi.org/10.21603/vcsn-2026-5-2-176-184

для создания десктопного графического интерфейса, по умолчанию выбирает в Windows бэкенд WinForms, который опирается на пакет pythonnet. При упаковке с помощью PyInstaller динамическая библиотека Python.Runtime.dll загружалась некорректно, вследствие чего процесс инициализации .NET-окружения прерывался и приложение не запускалось. Было решено использовать графический бэкенд EdgeChromium, не требующий pythonnet. После внесенных изменений приложение успешно запускается на всех протестированных конфигурациях Windows. Дополнительное требование – наличие установленного браузера Microsoft Edge (Chromium), что является стандартом в современных версиях системы и не вызвало проблем среди пользователей.

Вторая ошибка возникла среди пользователей, которые сообщали о невозможности получения метаданных при вводе сокращенных ссылок формата *https://**vm**.tiktok.com/…*, которые автоматически формируются при копировании адреса публикации из мобильного приложения TikTok. Скрейпер TikTok изначально обрабатывал только полные URL с явным указанием имени пользователя. При поступлении короткой ссылки модуль не мог корректно извлечь идентификатор поста. Нами было решено добавить в модуль TikTok-скрейпера предварительный этап разрешения коротких ссылок: перед парсингом выполняется HTTP-запрос с автоматическим следованием редиректу для получения полного URL, который затем обрабатывается остальным алгоритмом. Внесенная доработка не требует дополнительных действий со стороны пользователя и обеспечивает одинаковую работу приложения с любым типом ссылок. Повторное тестирование показало стопроцентную успешность извлечения данных независимо от формата TikTok-ссылки.

Заключение

Разработанное веб-приложение OmniTrack представляет собой комплексный инструмент сбора и параметризации метаданных, обладающий практической ценностью для лингвистов и специалистов в области анализа цифровых коммуникаций, а также имеющий потенциал дальнейшего совершенствования и интеграции в более масштабные аналитические системы.

Собранный при помощи веб-приложения OmniTrack корпус метаданных необходим для дальнейшего лингвистического анализа русскоязычных и англоязычных публикаций о Республике Беларусь в тревел-блогах, а именно выявления динамики тематических предпочтений, сезонных колебаний интереса к туристическим направлениям, а также сопоставления речевых стратегий и паттернов вовлеченности аудитории в разных языковых сообществах.

Конфликт интересов: Автор заявил об отсутствии потенциальных конфликтов интересов в отношении исследования, авторства и / или публикации данной статьи.

Conflict of interests: The author declared no potential conflict of interests regarding the research, authorship, and / or publication of this article.

Литература / References

Жучкова С. В., Ротмистров А. Н. Автоматическое извлечение текстовых и числовых веб-данных для целей социальных наук. *Социология: методология, методы, математическое моделирование*. 2021. № 50-51. С. 141–183. [Zhuchkova S. V., Rotmistrov A. N. Automatic extraction of text and numeric web data for social science purposes. *Sociology: Methodology, Methods, Mathematical Modeling (AM)*, 2021, (50-51): 141–183. (In Russ.)] https://elibrary.ru/xytjoy

Ahire V. Y. Assessing the effectiveness of metadata management systems in enhancing data governance: A primary study of IT and data-driven organizations. *Management Journal for Advanced Research*, 2025, 5(3): 85–90. https://doi.org/10.5281/zenodo.16792143

Berman F., Rutenbar R., Hailpern B., Christensen H., Davidson S., Estrin D., Franklin M., Martonosi M., Raghavan P., Stodden V., Szalay A. S. Realizing the potential of data science. *Communications of the ACM*, 2018, 61(4): 67–72. https://doi.org/10.1145/3188721

Brown M. A., Gruen A., Maldoff G., Messing S., Zanderson Z., Zimmer M. Web scraping for research: Legal, ethical, institutional, and scientific considerations. *ArXiv*, 2024. https://doi.org/10.48550/arXiv.2410.23432

Chani T., Olugbara O. O., Mutanga B. The problem of data extraction in social media: A theoretical framework. *Journal of Information Systems and Informatics*, 2023, 5(4): 1363–1384. https://doi.org/10.51519/journalisi.v5i4.585

Díaz de la Paz L., Crispí A. T., Mederos A. A. L. Model for the evaluation of metadata quality: Proposal for open science management in Cuba. *Advanced Notes in Information Science*, 2024, 6: 100–113. https://doi.org/10.47909/978-9916-9974-5-1.97

Edara P., Pasumansky M. Big metadata: When metadata is big data. *Proceedings of the VLDB Endowment*, 2021, 14(12): 3083–3095. https://doi.org/10.14778/3476311.3476385

Foerderer J. Should we trust web-scraped data? *ArXiv*, 2023. https://doi.org/10.48550/arXiv.2308.02231

Holom R.-M., Rafetseder K., Kritzinger S., Sehrschön H. Metadata management in a big data infrastructure. *Procedia Manufacturing*, 2020, 42: 375–382. https://doi.org/10.1016/j.promfg.2020.02.060

Huang Y.-N., Munteanu V., Love M. I., Ronkowski C. F., Deshpande D., Wong-Beringer A., Corbett-Detig R., Dimian M., Moore J. H., Garmire L. X., Reddy T. B. K., Butte A. J., Robinson M. D., Eskin E., Abedalthagafi M. S., Mangul S. Perceptual and technical barriers in sharing and formatting metadata accompanying omics studies. *Cell Genomics*, 2025, 5(5). https://doi.org/10.1016/j.xgen.2025.100845

Moreno-Ortiz A., García-Gámez M. Strategies for the analysis of large social media corpora: Sampling and keyword extraction methods. *Corpus Pragmatics*, 2023, 7: 241–265. https://doi.org/10.1007/s41701-023-00143-0

Ohme J., Araujo T., Boeschoten L., Freelon D., Ram N., Reeves B. B., Robinson T. N. Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, 2024, 18(2): 124–141. https://doi.org/10.1080/19312458.2023.2181319

Park J.-R., Tosaka Y. Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly*, 2010, 48(8): 696–715. https://doi.org/10.1080/01639374.2010.508711

Park J.-R., Tosaka Y., Maszaros S., Lu C. From metadata creation to metadata quality control: Continuing education needs among cataloging and metadata professionals. *Journal of Education for Library and Information Science*, 2010, 51(3): 158–176.

Pretorius K. A simple and systematic approach to qualitative data extraction from social media for novice health care researchers: Tutorial. *JMIR Formative Research*, 2024, 8: 1–9. https://doi.org/10.2196/54407

Skluzacek T. J., Chen M., Hsu E., Chard K., Foster I. Models and metrics for mining meaningful metadata. *International Conference on Computational Science. Computational Science – ICCS 2022*: Proc. 22nd Intern. Conf., London, UK, 21–23 Jun 2022. Springer, 2022, 417–430.

Subramaniam P., Ma Y., Li C., Mohanty I., Fernandez R. C. Comprehensive and comprehensible data catalogs: The what, who, where, when, why, and how of metadata management. *ArXiv*, 2021. https://doi.org/10.48550/arXiv.2103.07532

Wilkinson M. D., Dumontier M., Aalbersberg I. J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L. B., Bourne P. E., Bouwman J., Brookes A. J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C. T., Finkers R., Gonzalez-Beltran A., Gray A. J. G., Groth P., Grethe J. S., Mons B. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 2016, 3(1). https://doi.org/10.1038/sdata.2016.18

Yang W., Fu R., Bilal Amin M., Kang B. The impact of modern AI in metadata management. *Human-Centric Intelligent Systems*, 2025, 5: 323–350. https://doi.org/10.1007/s44230-025-00106-5

Yulfitri A., Sensuse D. I., Ulum M. B., Achmad Y. F. Metadata management to accelerate Big Data implementation. *Journal of Informatics and Communication Technology*, 2025, 6(2). https://doi.org/10.52661/jict.v6i2.362

Zachlod C., Samuel O., Ochsner A., Werthmüller S. Analytics of social media data – state of characteristics and application. *Journal of Business Research,* 2022, 144: 1064–1076. https://doi.org/10.1016/j.jbusres.2022.02.016