КРАЕВЕДЧЕСКИЙ ВЕБ-АРХИВ ЯКУТИИ: ИИ-КАТАЛОГИЗАЦИЯ В ПРОЕКТЕ «YAKUTIA.ONLINE»



doi 10.34920/2181-8207/2025/2-172



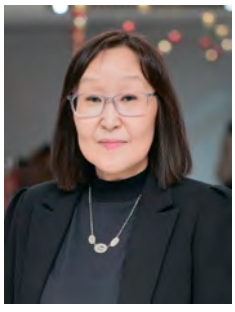
Саргылана Максимова, директор Национальной библиотеки Республики Саха (Якутия)

Annotation. The National Library of the Republic of Sakha (Yakutia) implements the «Yakutia. Online» project, whose primary objective is to preserve digital local history resources and ensure their broad accessibility. The project involves monitoring internet resources, employing artificial intelligence (AI) technologies to automate the processes of identifying, cataloging, and thematically classifying content, as well as developing methodological guidelines for selecting online local history materials. The project is integrated with the «ILIM» search service, which provides centralized access to materials through the library's official portal.

Keywords: web archiving, local history web archive, AI cataloging, online local history resources, bibliographic cataloging.

Аннотация. Национальная библиотека Республики Саха (Якутия) реализует проект «Yakutia.Online», основной целью которого является сохранение цифровых краеведческих ресурсов и обеспечение к ним широкого доступа. В рамках проекта осуществляется мониторинг интернет-ресурсов, применяются технологии искусственного интеллекта (ИИ) для автоматизации процессов выявления, каталогизации и тематической классификации контента, а также разрабатываются методические рекомендации по отбору сетевых краеведческих материалов. Проект интегрирован с поисковым сервисом «ИЛИМ», обеспечивающим централизованный доступ к материалам через официальный портал библиотеки.

Ключевые слова: веб-архивирование, краеведческий веб-архив, ИИ-каталогизация, сетевые краеведческие ресурсы, библиографическая каталогизация.



Яна Чабыева, заведующая Центром формирования интегрированных информационных ресурсов Национальной библиотеки Республики Саха (Якутия)

Национальная библиотека Республики Саха (Якутия), следуя своей миссии по созданию исчерпывающего национального документного фонда и библиографии, реализует стратегически важный проект «Yakutia. Online». Основная цель проекта - сохранение и обеспечение доступности цифровых краеведческих ресурсов, представляющих культурно-историческое наследие региона.

Ключевым компонентом проекта стало веб-архивирование, которое решает три важнейшие задачи:

1. Долгосрочное сохранение цифрового контента, имеющего культурную и научную ценность.
2. Расширение доступа к краеведческим ресурсам через интернет-пространство.
3. Популяризация уникального наследия Республики Саха (Якутия).

Реализация проекта позволяет не только сохранить цифровые свидетельства современной эпохи для будущих поколений, но и значительно расширить исследовательские возможности за счет систематизированного доступа к электронным краеведческим ресурсам. Это соответствует стратегическим задачам библиотеки по формированию комплексного национального документного фонда в цифровую эпоху.

Хронология реализации проекта «Yakutia.Online»

2023 год:

- проведен комплексный мониторинг интернет-ресурсов;
- сформирован реестр приоритетных источников;

- разработаны критерии отбора контента;
- создана технологическая карта проекта;
- осуществлена библиографическая каталогизация в Сводной базе краеведческих статей «Саха сирэ»;
- проведены работы по веб-архивации отобранных ресурсов;
- реализован комплекс мероприятий по продвижению сетевых краеведческих материалов.

2024 год:

- выполнена библиографическая каталогизация в Сводном каталоге электронных ресурсов Республики Саха (Якутия) (СКЭР РС (Я));
- разработаны методические рекомендации по библиографической обработке сетевых краеведческих ресурсов в формате RUSMARC (аналитическое и монографическое описание).
- внедрен экспериментальный модуль ИИ-каталогизации подсистемы «Веб-архивирование».

Организационное обеспечение:

- утверждено временное «Положение о веб-архивации сетевых краеведческих ресурсов в НБ РС(Я)»;
- четко регламентированы механизмы взаимодействия между участниками проекта;
- оптимизированы рабочие процессы между подразделениями.

Каждый этап реализации сопровождался мониторингом эффективности и корректировкой рабочих процессов для достижения оптимальных результатов.

Технологическая платформа проекта позволяет эффективно решать задачи выявления, обработки, архивации и предоставления доступа к цифровому краеведческому ресурсу республики.

В основе проекта лежит многоуровневая технологическая платформа. Ее ключевой компонент - интеллектуальная система поиска и каталогизации, созданная специалистами Отдела цифровых ресурсов (ОЦР) и использующая передовые методы машинного обучения для идентификации релевантных ресурсов. Особое значение имеет модуль ИИ-каталогизации, который генерирует предварительные библиографические описания в системе «OPAC-Global» в полном соответствии с современными нормативными требованиями.

Архивация ресурсов осуществляется в международном формате WARC (Web ARChive), обеспечивающем долгосрочную сохранность цифровых объектов любого типа - от PDF-документов до мультимедийного контента (аудио, видео, изображений).

Доступ пользователей организован через единый поисковый сервис «ИЛИМ», который предоставляет: полнотекстовый поиск по веб-архивированным материалам, удобный интерфейс для работы с ресурсами, интеграцию с официальным порталом библиотеки.

На начальном этапе реализации проекта был проведен комплексный мониторинг интернет-источников, публикующих материалы о Республике Саха (Якутия). Критериями отбора стали: соответствие краеведческому содержанию, географическая привязка к региону, культурная и историческая значимость, многообразие языков публикации, востребованность информации у потенциальных пользователей.

Выявленные ресурсы охватывают 26 республиканских СМИ: «ЯСИА», «Якутия24», «Sakhalife», «SakhaNews», «SakhaPress», «Ulus.Media» и другие профильные издания. В перечень вошли федеральные российские ресурсы (32 источника): «Региональная Россия», «Российская Арктика», «Культура.РФ» и др. Среди российских сетевых СМИ статьи о регионах чаще всего встречаются на сайтах «РИА Новости», «ФедералПресс», «EastRussia» («Восток России»), где есть специальные разделы с новостями по регионам. Дополнительно включены 6 зарубежных сайтов, публикующих статьи о Якутии и Арктике. (см. *рис.1)*

В основу отбора ресурсов положены следующие принципы: использование только открытых источников, строгое соблюдение авторских прав с обязательным указанием авторов и первоисточников, работа исключительно с материалами, не требующими специальных разрешений на архивирование. Такой подход гарантирует правомерность хранения и предоставления доступа к материалам, а также позволяет создать полноценный электронный фонд, который достоверно и разносторонне представляет информацию о регионе из различных проверенных источников.

Отслеживаемые электронные ресурсы для веб-архивирования основываются на принципах, схожих с отбором печатных документов, но с учетом их цифровой специфики. Ключевые аспекты оценки включают:



Ольга Софронеева, заведующая Отделом каталогизации ресурсов Национальной библиотеки Республики Саха (Якутия)

В основу отбора ресурсов положены следующие принципы: использование только открытых источников, строгое соблюдение авторских прав с обязательным указанием авторов и первоисточников, работа исключительно с материалами, не требующими специальных разрешений на архивирование

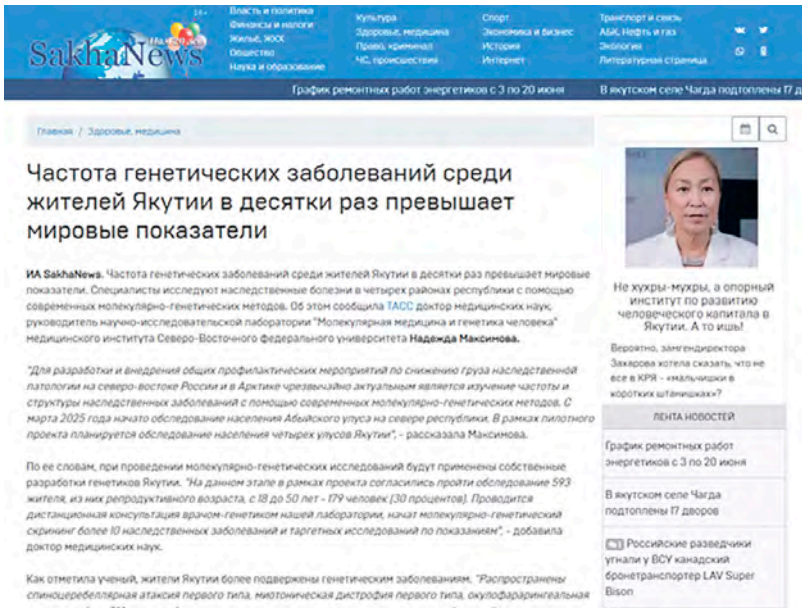


Рис.1. Веб-страница сайта «SakhaNews»

1. Содержательные характеристики:
   - качество контента (информационная ценность, достоверность, глубина раскрытия темы);
   - язык и страна происхождения;
   - авторитетность источника.
2. Технические параметры:
   - стабильность веб-страницы (частота изменений URL, риск исчезновения);
   - удобство навигации и чтения (оформление текста, графики, таблиц).
3. Объем и значимость информации:
   - достаточность материалов для комплексного освещения темы;
   - долговременная ценность контента.

Отобранные материалы классифицируются по форматам: текстовые (статьи, документы, публикации), мультимедийные (аудио- и видеозаписи), визуальные (изображения, инфографика).

После завершения отбора сетевых краеведческих ресурсов начинается этап их библиографической обработки, который осуществляется по следующим направлениям:

1. Отдел каталогизации составной части ресурсов (ОКСЧР) выполняет: каталогизацию документов из электронных периодических изданий, создает библиографические записи на статьи и размещает обработанные материалы в Сводной базе краеведческих статей «Саха сирэ».

2. Отдел каталогизации ресурсов (ОКР) специализируется на каталогизации сетевых СМИ, формирует монографические библиографические записи на ресурсы и включает обработанные материалы в СКЭР РС (Я).

Системы веб-архивирования обеспечивает управление работой робота, который ежедневно осуществляет автоматический мониторинг веб-ресурсов, поиск соответствующих краеведческих материалов и их отбор по заданным тематическим критериям

Такой распределенный подход позволяет обеспечить комплексную обработку всех типов электронных ресурсов с соблюдением единых стандартов каталогизации. Оба отдела работают в тесной координации, что гарантирует целостность и системность формируемого электронного краеведческого фонда.

В июле 2024 года для оптимизации обработки сетевых краеведческих ресурсов (СКР) был внедрен экспериментальный модуль ИИ-каталогизации в составе подсистемы «Веб-архивирование», который включает два ключевых компонента:

- веб-интерфейс для управления настройками робота;
- веб-интерфейс для работы с заявками.

Данное решение позволило автоматизировать трудоемкие процессы, сократить временные затраты на библиографическое описание на 40% и повысить точность каталогизации благодаря интеллектуальным алгоритмам, одновременно снизив объем рутинных операций, улучшив качество метаданных и упростив взаимодействие между специалистами, при этом система продолжает развиваться.

Веб-интерфейс «Настройки робота» системы веб-архивирования обеспечивает управление работой робота, который ежедневно осуществляет автоматический мониторинг веб-ресурсов, поиск соответствующих краеведческих материалов и их отбор по заданным тематическим критериям. На текущем этапе тестирования подключены три ключевых республиканских СМИ: ЯСИА (Якутское Саха Информационное Агентство), Sakhalife и Yakutia-daily, при этом каталогизаторы имеют возможность добавлять новые источники для мониторинга, настраивать и обновлять тематические фильтры. Простой и удобный интерфейс позволяет гибко настраивать поиск, просматривать результаты и при необходимости корректировать их.

Веб-интерфейс «Заявки на веб-архивирование» - ИИ-каталогизатор автоматически создает заявку и заготовку RUSMARC-записи для каждого выявленного сетевого краеведческого ресурса, предварительно проверяя наличие дублетных записей. На следующем этапе каталогизатор

с соответствующими правами распределяет заявки между специалистами, которые через веб-интерфейс проводят верификацию ресурсов: проверяют соответствие критериям отбора, содержание и качество материала, а затем либо создают библиографическую запись с помощью кнопки «Создать БЗ», либо отклоняют ресурс кнопкой «Отклонить» в случае несоответствия установленным требованиям веб-архивирования. (см. *рис. 2)*

Обработка БЗ каталогизатором. При нажатии кнопки «Создать БЗ» система автоматически формирует заготовку RUSMARC-записи в Сводном каталоге электронных ресурсов Республики Саха (Якутия) и открывает в новой вкладке браузера интерфейс АИБС «OPAC-Global» с данной заготовкой.

Каталогизатор выполняет следующие действия:

- проверяет автоматически сгенерированную заготовку библиографической записи в формате RUSMARC;
- дополняет запись, проставляя предметные рубрики, классификационные индексы (УДК, ББК) и заполняя поля 7 Блока интеллектуальной ответственности.

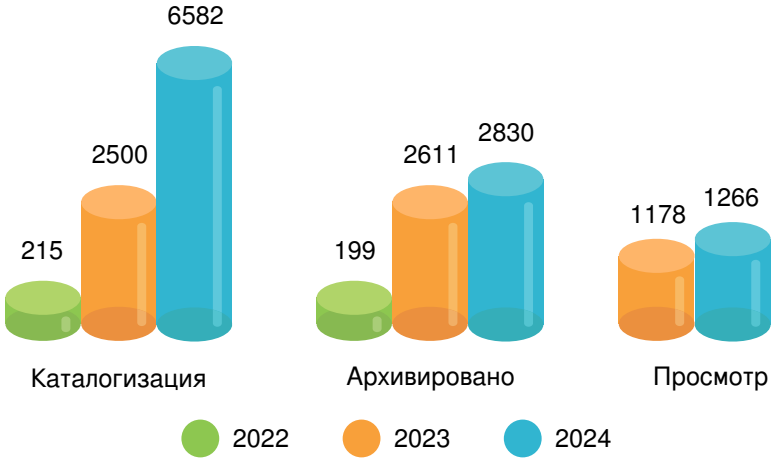
При проведении библиографической каталогизации обязательно используется авторитетная запись из Национального авторитетного файла Республики Саха (Якутия) (НАФ РС(Я)). В случае отсутствия необходимой авторитетной записи каталогизатор создает новую. (см. *рис. 3)*

После финализации библиографической записи робот осуществляет веб-архивирование страницы путем её сохранения на сервере библиотеки и создания электронного объекта. Для этого используется уникальный идентификатор записи (OPAC ID), который увязывает запись и копию страницы. Сохранение веб-страницы производится в формате WARC, принятом международным стандартом для долгосрочного хранения веб-контента. Перед сохранением контент проверяется на целостность и отсутствие вредоносного кода, после чего помещается на защищенные серверы библиотеки. Доступ пользователей к архивированным ресурсам организован через единый поисковый сервис «ИЛИМ» и официальный сайт библиотеки: веб-архивы представлены в виде структурированных электронных объектов с полным библиографическим описанием. (см. *рис. 4)*

Интеграция с поисковой системой «ИЛИМ» обеспечивает пользователям быстрый доступ к архивированным ресурсам.

(Статистика)

Динамик развития проекта «Yakutia.Online»



«ИЛИМ» (поисковая строка) предоставляет пользователям единую точку доступа ко всем оцифрованным материалам и веб-архиву, удобный поиск по каталогам, тематическим коллекциям и возможность работы с архивами через портал библиотеки. В результате создана единая экосистема, в которую включаются библиографические записи, краеведческие веб-ресурсы и цифровые копии из документного фонда библиотеки. *(Статистика)*

Технологии искусственного интеллекта доказали свою эффективность в процессе библиографической каталогизации, обеспечивая: существенное ускорение обработки ресурсов, повышение точности формирования метаданных, оптимизацию трудозатрат за счет автоматизации рутинных операций.

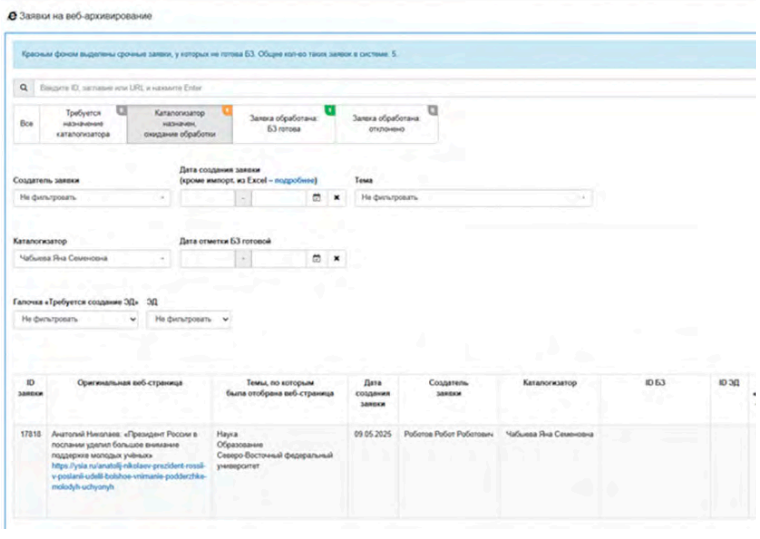


Рис.2. Заявки на веб-архивирование

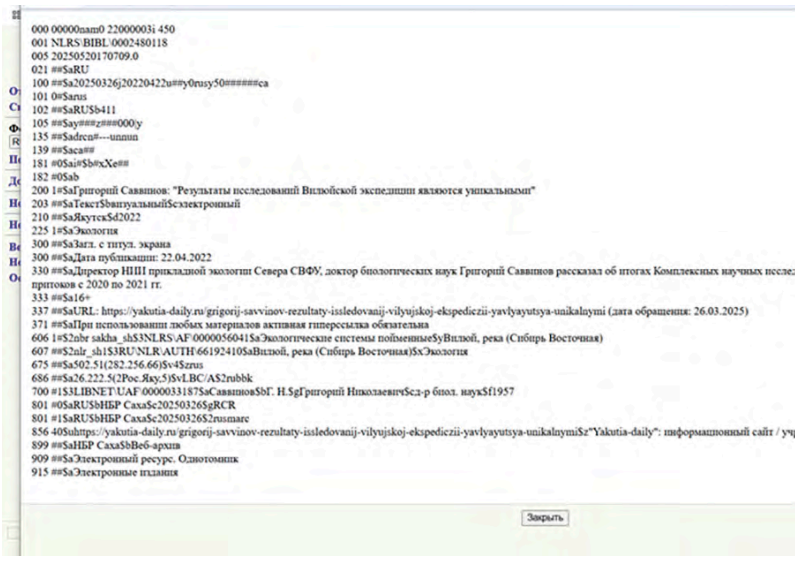


Рис.3. ИИ-запись веб-архивированного ресурса в формате RUSMARC

Комбинация веб-архивирования и ИИ-каталогизации позволяет:

1. Создавать масштабируемую базу сетевых ресурсов, включающую: текстовые материалы, мультимедийный контент, комплексные цифровые объекты.
2. Обеспечивать долгосрочное сохранение цифрового культурного наследия.
3. Расширять доступ к краеведческим знаниям через современные поисковые инструменты и интерактивные форматы представления данных.

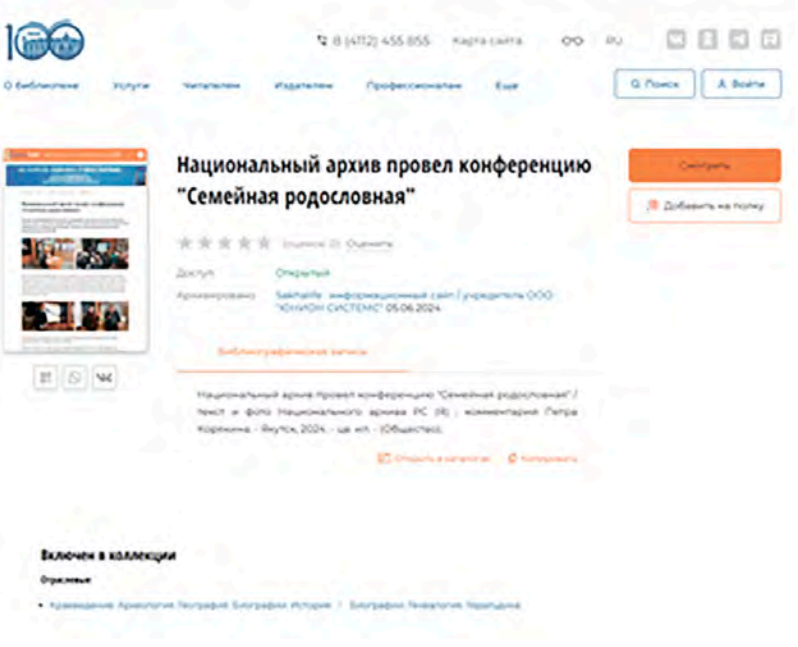


Рис.4. Веб-архивированная публикация на сайте НБ РС (Я)

Данные решения открывают новые возможности для систематизации и популяризации историко-культурного наследия Республики Саха (Якутия) в цифровой среде.

Список источников

1. *Балацкая Н. М. Краеведческий веб-архив в структуре информационных ресурсов библиотеки: модель и возможности реализации / Н. М. Балацкая, М. Б. Мартиросова. – Санкт-Петербург : Российская Национальная библиотека, 2023. – 207 с.*

2. *Ершова, В. В. Краеведческие ресурсы интернета как объект веб-архивирования : опыт реализации проекта Yakutia.Online в Национальной библиотеке Республики Саха (Якутия) / В. В. Ершова // Краеведческие ресурсы сети как объект веб-архивирования и исследования : сборник статей. – Санкт-Петербург, 2024. – С. 98-114.*

3. *Максимова, С. В. Новая стратегия развития цифровой интегрированной системы Национальной библиотеки Республики Саха (Якутия) / С. В. Максимова // Научные и технические библиотеки. – 2024. – № 3. – С. 127-141.*

4. *Национальная библиотека научила ИИ говорить и писать на якутском языке. – 13.02.2025. – URL: https://1sn.ru/nacionalnaya-biblioteka-naucila-ii-govorit-i-pisat-na-yakutskom-yazyke (SakhaNews : сетевое издание / учредитель: ООО «ЮНИОН СИСТЕМС» ; дата архивирования: 24.02.2025).*

5. *Национальная библиотека Якутии первой в России внедрила в свои сервисы GigaChat : за первые шесть месяцев работы AI-помощник обработал более 33 тысяч запросов пользователей. – 13.12.2024. – URL: https://ysia.ru/natsionalnaya-biblioteka-yakutii-pervoj-v-rossii-vnedrila-v-svoi-servisy-gigachat (ЯСИА : сетевое издание / Учредители: АО «РИИХ Сахамедиа», Министерство инноваций, цифрового развития и инфокоммуникационных технологий РС(Я) ; дата архивирования: 18.03.2025).*

6. *Об ИИ-продуктах НБ РС(Я) : [видеозапись] / Национальная библиотека РС (Я). – Якутск : НБ РС (Я), Медиацентр, 2025. – 1 видеофайл (mp4 ; 353 Мб ; 2 мин 20 с) : цв., зв. – URL: https://l.nlrs.ru/open/124841 (дата обращения: 16.04.2025).*