CONTENT AI

CONTENT AI ***ОЦИФРОВАЛА*** БИОГРАФИЧЕСКИЙ ***ДВУХТОМНИК ЛЬВА ТОЛСТОГО*** ДЛЯ ЭЛЕКТРОННОГО ПУТЕВОДИТЕЛЯ ПО НАСЛЕДИЮ АВТОРА

«Слово Толстого» — цифровой путеводитель по наследию писателя, созданный на основе 90-томного собрания сочинений Льва Толстого.

Этот проект — результат многолетней работы группы Tolstoy Digital, филологов и специалистов по Digital Humanities.

> Такого полного и системного цифрового представления наследия писателя до нас никто не делал, и нам очень приятно, что этот путь первопроходцев с нами разделяют наши партнеры — также заинтересованные в создании нового, как и мы. Еще 10 лет назад мы сделали большой волонтерский проект «Весь Толстой в один клик» на базе технологий, которые использует Content AI. Его результатом стало выверенное цифровое издание 90-томного собрания сочинений Толстого. Сегодня благодаря коллегам из Content AI мы смогли использовать всемирно признанные технологии распознавания текста для быстрого и качественного перевода сложных научных книг в цифровой вид, сразу распознавая цитаты, даты и ссылки на книги.
>
> ***ФЕКЛА ТОЛСТАЯ***, инициатор проекта «Слово Толстого», руководитель группы Tolstoy Digital.

Оцифровать нельзя обработать вручную

«Слово Толстого» — это цифровой путеводитель по наследию писателя Льва Толстого, который создан на базе его многочисленных сочинений. Удобная навигация ресурса позволяет осуществлять поиск по 90-томному собранию автора, используя различные фильтры, а также получать дополнительную информацию по описанным им событиям.

Проект постоянно развивается и пополняется новой информацией, поскольку является важной частью глобальной программы подготовки к 200-летию со дня рождения Толстого в 2028 году. И двухтомное издание «Летопись жизни и творчества Л. Н. Толстого», написанное личным секретарем писателя Николаем Гусевым, стало очередным пополнением коллекции проекта. Оцифровать и разметить тегами почти 2 тыс. страниц удалось с помощью российского разработчика решений для интеллектуальной обработки информации Content AI.

Цели

- Оцифровать биографический двухтомник «Летопись жизни и творчества Л. Н. Толстого» Николая Гусева
- Выделить атрибуты для расстановки тегов по тексту
- Обеспечить возможность удобной навигации по тексту при обращении к справочным материалам

Решение

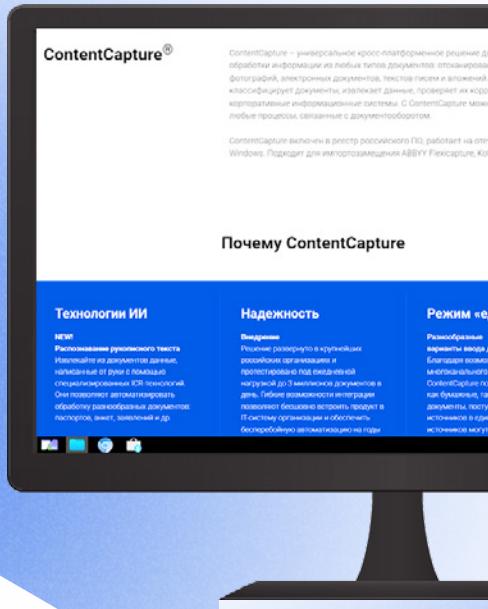
Универсальная платформа для интеллектуальной обработки информации ContentCapture

Результат

- Двухтомник «Летопись жизни и творчества Л. Н. Толстого» Николая Гусева переведен в электронный вид
- На основе размеченного текстового материала создан календарь, в котором в удобной форме можно читать биографию автора и соотносить эти данные с другими источниками
- Пользователи могут свободно перемещаться по многостраничному документу, уточняя по клику дополнительную информацию

CONTENTCAPTURE – ИЗВЛЕЧЕНИЕ И ОБРАБОТКА ДАННЫХ ИЗ ЛЮБЫХ ТИПОВ ДОКУМЕНТОВ

ContentCapture



Перед Content AI стояла задача не просто оцифровать двухтомник о писателе, но и выделить в тексте атрибуты для расстановки тегов, которые соотносятся с различными типами данных: датами и местами, где происходили важные события в жизни Толстого, или дополнительными комментариями автора.

Для реализации проекта решено было использовать IDP-платформу ContentCapture, которая с помощью признанных во всем мире OCR- и NLP-технологий умеет извлекать данные из любых типов и форматов документов и обрабатывать их по заданным сценариям.

Сначала команда Content AI разработала логику извлечения нужных полей в тексте, а также гибкие описания для выделения нескольких десятков необходимых атрибутов с дополнительными деталями по каждому событию.

Затем ContentCapture распознала отсканированные страницы издания и запустила процесс обработки и извлечения нужных атрибутов из текста. Для того чтобы этот этап прошел без ошибок, использовались скрипты автокоррекции и местозаполнители – с их помощью удалось проанализировать структуру документа, разобрать описание событий на структурные детали, восстановить пропущенные в тексте или представленные иносказательно данные (например, «в том же году»).

В результате полученный интерактивный многостраничный текст с размеченными данными лег в основу календаря событий из жизни автора.

ИИ и Толстой

ContentCapture с помощью технологий искусственного интеллекта смогла оцифровать и разметить почти 2 тыс. страниц двухтомника о жизни Льва Толстого. Результатом работы стал подробно размеченный текстовый материал, представленный в структурированном интерфейсе. На его основе был создан календарь, в котором в удобной форме можно читать биографию Л.Н. Толстого, написанную Н.Н. Гусевым, и соотносить эти данные с другими источниками. Также у пользователей есть возможность искать нужную информацию по хронике, используя многочисленные фильтры.

Создатели проекта планируют в будущем передать все собранные данные в Институт русского языка РАН для дальнейшего создания «Словаря языка Толстого», а также подготовить корпус файлов для создания подкорпуса Толстого в НКРЯ.

CONTENT AI ×

ООО «Контент ИИ»
office@contentai.ru
Юридический адрес: 127273, г. Москва, ул. Отрадная, дом 2Б, строение 6, помещ. 1/3.