Создание программного комплекса для исследовательской онлайн работы с архивом рукописей Ф. М. Достоевского

В.В. Захаркина, И.А. Мбого

Санкт-Петербургский государственный университет

Аннотация. Создан и развивается на протяжении ряда лет программный комплекс работу текстологов, изучающих рукописный архив Ф.М. Достоевского. Портал «Архив Ф.М. Достоевского» объединяет результаты нескольких проектов, использующих единый первичный массив данных – цифровые копий страниц девятнадцати рукописных тетрадей Ф. М. Достоевского с соответствующими текстовыми расшифровками (более 3000 страниц). Программная интерпретация разметки текстов позволяет в интерактивном режиме выявлять слои авторской правки. В настоящее время идёт работа над созданием корпуса графических образцов, реализованы программные решения, предоставляющее онлайн инструментарий для создания, веб-публикации и гибкого отображения библиотеки графических образцов. База данных графических образцов пополняется, развиваются возможности просмотра и фильтрации материалов. Результаты работы предоставят основу для дальнейшей исследовательской работы.

Ключевые слова: Достоевский, методы исследования рукописей, цифровой архив писателя, информационная система, программный анализ текстовой разметки

Creation of an online software for research work with the archive of F. M. Dostoevsky's manuscripts

V.V. Zakharkina, I.A. Mbogo

Saint-Petersburg State University

Abstract. A software package has been created and has been developing for a number of years, the work of textologists studying the handwritten archive of F.M. Dostoevsky. The portal "Archive of F.M. Dostoevsky" combines the results of several projects using a single primary data array – digital copies of pages of nineteen handwritten notebooks of F.M. Dostoevsky with corresponding text transcripts (more than 3000 pages). Software interpretation of text markup allows you to interactively identify layers of author's edits.

Currently, work is underway to create a corpus of graphic samples, software solutions have been implemented that provide online tools for creating, web publishing and flexibly displaying a library of graphic samples. The database of graphic samples is being updated, the possibilities of viewing and filtering materials are being developed. The results of the work will provide a basis for further research work.

Keywords: Dostoevsky, methods of manuscript research, digital archive of the writer, information system, program analysis of text markup

Введение

В течение ряда лет в Институте русской литературы РАН (Пушкинский дом) проводится работа в рамках грантов РНФ, посвящённая текстологическому анализу архивных рукописей Ф. М. Достоевского и нацеленная на онлайн публикацию результатов.

Первичный массив данных представлял собой отсканированные 19 рукописных тетрадей из собраний РГАЛИ и ОР РГБ и соответствующие текстовые расшифровки, работа над которыми шла ранее полтора десятка лет.

Задачей первых грантовых проектов был анализ текстовых расшифровок, их коррекция и формальная разметка в соответствии с принятыми принципами «динамической транскрипции». Специалистами-текстологами были выделены фрагменты, зачёркнутые либо дописанные автором, отмечены неразборчивые фрагменты и записи на полях (маргиналии), сформулированы обширные комментарии, значимые для дальнейшей исследовательской работы.

Для обеспечения публикации результатов был создан сайт и разработан программный инструментарий, обеспечивающий просмотр страниц рукописей и оригинальные интерактивные возможности работы с динамической транскрипцией. Важным технологическим решением была реализация «пакетной загрузки» на сайт массивов подготовленных данных.

Текущий грантовый проект направлен на выявление, каталогизацию и последующий анализ графических элементов рукописей (образцы почерка, каллиграфия, рисунки). В технологическом аспекте эти задачи потребовали создания новых структур данных, а также расширения функциональности веб-приложения (выделение фрагментов страницы рукописи, определение их места в принятой таксономии, пополнение базы данных и т.д.). Самым существенным здесь представляется новый аспект: обеспечение текущей совместной исследовательской работы в онлайн режиме. Созданный программный комплекс развивается в соответствии с вновь возникающими задачами, сформулированными коллективом исследователей.

Следует отметить, что текущая версия сайта «Архив Ф. М. Достоевского» (http://dostoevsky-archive.ru) имеет рабочий характер. В соответствии с политикой ИРЛИ, разделы, посвящённые рукописям (отсканированные страницы и их текстовые расшифровки с динамической транскрипцией) доступны лишь зарегистрированным пользователям. Библиотека графических образцов (более 48000 образцов почерка, рисунки, каллиграфия) доступна без авторизации.

Авторы данной статьи участвуют в проектах как специалисты по информационным технологиям и веб-разработчики. Наша роль: создание и развитие сайта и его структур; создание веб-приложений, обеспечивающих исследовательскую работу в онлайн режиме и пополнение базы данных; алгоритмы обработки динамической транскрипции и соответствующие интерфейсные решения; представление на сайте результатов исследований в их динамике на уровне текущего состояния базы данных.

Результаты многолетней работы по созданию программного комплекса публикуются впервые.

1. Текстовые расшифровки с динамической транскрипцией и их онлайн представление

В обширном раздел «Рукописи» (доступен лишь для авторизованных пользователей в соответствии с политикой ИРЛИ) опубликованы 19 рукописных тетрадей. Для каждой страницы выводится отсканированное изображение и (по выбору) варианты текстовых расшифровок: в оригинальной либо современной орфографии, в формате HTML либо PDF.

Способом соотнесения расшифровки с её рукописным источником является динамическая транскрипция – особая система знаков, отражающая характер авторского письма и правок. Несмотря на то, что эта система довольно проста для понимания, её восприятие ухудшается вместе с тем, как возрастает сложность рукописи. Это стало первой предпосылкой для создания дополнительного инструментария, визуализирующего отдельные семантически существенные элементы разметки и дающего возможность работы в интерактивном режиме.

Одним из основных онлайн инструментов является возможность автоматической обработки транскрипций. Разработан и реализован онлайн инструмент анализа и визуализации сущностей динамической транскрипции в соответствии с выбранным режимом просмотра расшифровки в формате HTML.

В результате программного анализа текста с редакторской маркировкой создаётся HTML-разметка, обеспечивающая манипулирование элементами DOM (Document Object Model) на уровне языка описания стилей визуального отображения CSS (Cascading Stile Sheets) и языка сценариев JavaScript.

Так, например, фрагменты, помеченные как вычеркнутые автором, могут быть оформлены перечёркнутым шрифтом либо скрыты; текст, дописанный автором, может быть выделен определённым цветом; редакторские пометы могут быть временно убраны из визуального представления. По умолчанию в выбранном типе расшифровки отображается редакторская разметка. Разработанный инструментарий имеет ряд режимов интерактивного отображения элементов транскрипции: редакторская разметка, цветовое/стилевое оформление транскрибированных элементов, скрытие редакторской разметки в зоне просмотра.

Рис. 1 иллюстрирует одновременное включение двух режимов: выделение дописанных и зачёркнутых автором фрагментов. Визуализация элементов транскрипции сразу позволяет увидеть, что в результате авторской правки некоторые добавленные фрагменты были позже вычеркнуты.

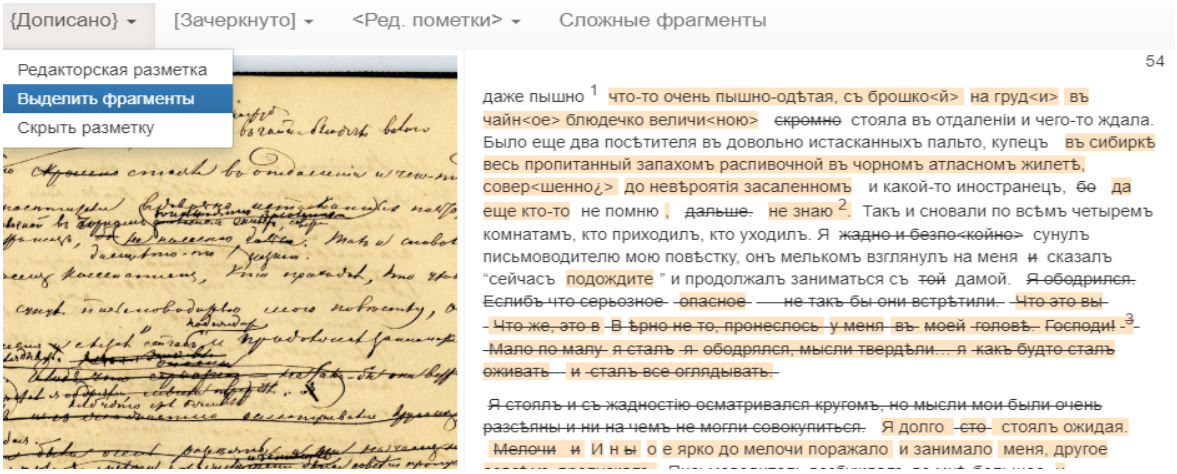

Рис. 1. Дописанные фрагменты выделены фоновым цветом, зачёркнутые – соответствующим стилем шрифта

Одной из важных задач, которая может быть решена на основе алгоритмического анализа расшифрованного и размеченного текста рукописи, является определение слоёв авторской правки. Отметим, что в простых случаях возможность выявления некоторых слоёв даёт включение очевидных режимов отображения отдельных элементов размеченного текста. На рис 2 представлена ситуация, когда зачёркнутые фрагменты временно скрыты, выделение фоновым цветом дописанных фрагментов сохранено. В результате можно видеть последний слой авторской правки с акцентом на дописанный текст.

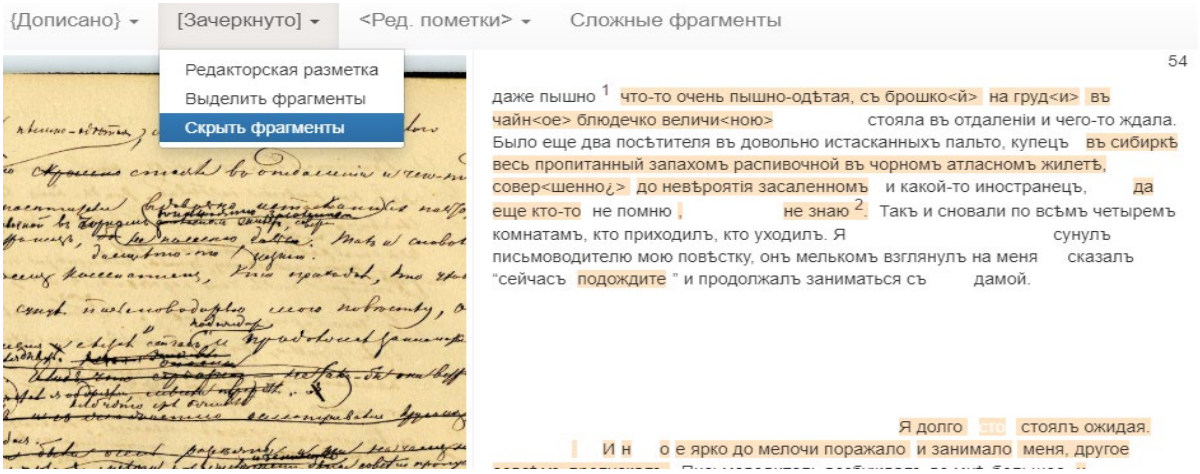

Рис. 2. Последний слой авторской правки с акцентом на дописанный текст

В случае же многократных последовательных авторских вычёркиваний и дописываний (на некоторых страницах выделено до 5 слоёв) задача программной обработки расшифрованного и размеченного текста становится нетривиальной. С формальной точки зрения она может быть сведена к анализу вложенных фрагментов двух типов («дописано/вычеркнуто») с определением возможной последовательности действий.

В настоящее время реализован подход, использующий рекурсивный алгоритм. В рамках текущего интерфейсного решения при просмотре страниц рукописи с авторской правкой более двух уровней вложенности появляется возможность просмотра «сложных фрагментов». Программный сценарий на основе анализа текста генерирует встроенный навигационный элемент, который обеспечивает интерактивный просмотр этапов авторской правки «сложного фрагмента». Функция визуализации «сложных фрагментов» демонстрирует количество этапов авторской правки и их результат. В окне текстовой расшифровки дописанные фрагменты пошагово проявляются, а вычеркнутые – выделяются зачёркнутым шрифтом (рис. 3). Слои авторской правки можно просмотреть в интерактивном режиме.

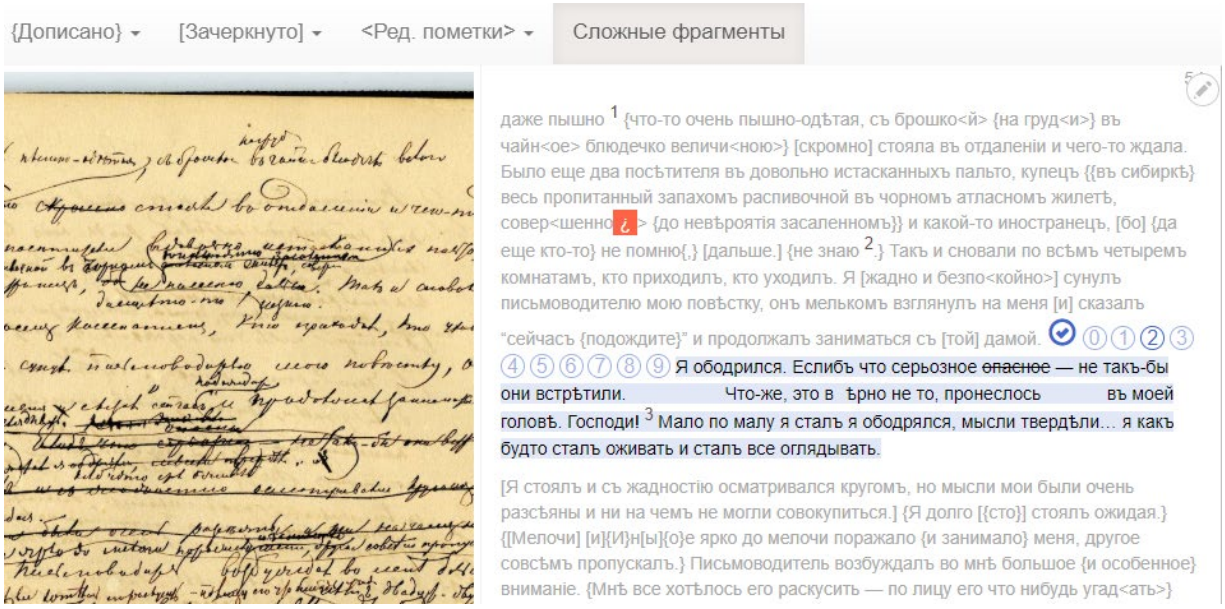

Рис 3. Интерфейсная реализация просмотра многоуровневой авторской правки. На скриншоте 2 слой

Текстовые расшифровки и их динамическая транскрипция доступны для дальнейшего редактирования участниками проектов.

2. Проблемы пополнения базы данных и реализованные решения

Нетривиальным аспектом технологической разработки стал тот факт, что к моменту подключения авторов статьи к работе над проектом текстологами – участниками проекта – был разработан оригинальный вариант разметки текстов в соответствии с их идеей динамической транскрипции. К моменту проектирования онлайн представления уже был подготовлен значительный объём материалов.

Текстовые материалы, подлежащие публикации на сайте, изначально готовились в формате MS Word, разметка проводилась на уровне заранее согласованных комбинаций специальных символов. Каждая тетрадь была представлена единым файлом MS Word. На сайте же необходимо публиковать каждую страницу тетради в виде отдельных веб-документов в формате HTML и PDF, причём в двух вариантах транскрипции: оригинальном и современном.

Мы согласились с реалиями уже налаженного процесса. Действительно, предварительная обработка принципиально позволяет «разрезать» большой файл на страницы и автоматически конвертировать их в форматы HTML и PDF. В дальнейшем эти файлы могут быть использованы для «пакетной загрузки» страниц тетради на сайт.

Были разработаны программные сценарии, обеспечивающие анализ исходных текстов MS Word, очистку ненужного форматирования

(характерного для работы в MS Word разных участников проекта) и конвертацию их фрагментов в необходимые форматы.

Созданы программные модули, обеспечивающие пакетную загрузку материалов на сайт (тексты в форматах HTML и PDF в оригинальной и современной орфографии, а также соответствующие сканы страниц архивных рукописей).

Таким образом, на сайт были загружены полные материалы по 19 рукописным тетрадям (более 3000 страниц). Текстовые расшифровки и их динамическая транскрипция доступны для дальнейшего редактирования участниками проектов.

3. Создание онлайн-платформы для проектов с единой первичной информационной базой

Текущий проект в рамках гранта РНФ, посвящённый анализу графических элементов рукописей, развивается на уже созданной первичной базе: страницы рукописей с текстовыми расшифровками.

В основе идеи создания платформы для веб-представления проектов, связанных между собою лишь первичным информационным массивом, лежат следующие соображения.

3.1. С точки зрения концептуальной представляется весьма эффективной возможность сконцентрировать функциональность различных проектов на одном портале. Так, например, при работе над текущим проектом уже реализован просмотр и выбор страницы рукописи для выделения графических элементов с использованием тех же интерфейсов, что и для просмотра текстовых расшифровок. Далее планируется интегрировать новые возможности, обеспечивающие интерактивный просмотр графических образцов, в объединённый инструментарий ряда веб-приложений.

Желаемый результат – возможность эффективной работы исследователя, который сможет выбрать страницу рукописи и увидеть всю сопутствующую информацию (в текстовом либо графическом виде), которая следует из проведённых исследований по всем проектам (и текстовую расшифровку, и выделенные графические элементы, и иные сущности).

3.2. С точки же зрения технологической можно обратить внимание на следующие моменты:

- нецелесообразность дублирования первичного массива данных (в нашем случае – цифровых копий страниц рукописей Ф. М. Достоевского из собрания ИРЛИ РАН) на нескольких веб-ресурсах;
- обеспечение возможности развития инструментальной платформы для интеграции дальнейших проектов.

4. Развитие платформы. Создание инструментов, обеспечивающих онлайн работу коллектива исследователей над новым проектом, посвящённым графическим элементам рукописей

Существенная особенность текущего проекта состоит в том, что что вся работа, связанная с пополнением библиотеки графических образцов, происходит исключительно онлайн (через браузер). Созданы соответствующие программные модули и разработан специальный инструментарий, доступный авторизованному участнику проекта. Отметим лишь некоторые важные моменты:

Онлайн режим работы продиктован необходимостью в любой момент видеть, оценить и (при необходимости) скорректировать полный массив данных, подготовленный всеми участниками проекта. При работе коллектива исследователей над единым информационным массивом возникает целый ряд технических проблем, которые могут быть решены лишь в рамках онлайн формата.

Обеспечить просмотр, синхронизацию и коррекцию данных возможно, по существу, лишь двумя путями. Первый заключается в разработке специализированного приложения, которое обеспечит требуемую функциональность и должно будет быть установлено на все компьютеры, на которых предполагается работать. Мы выбрали второй путь – создание веб-приложения, которое даст возможность работы участникам проекта, которые успешно прошли авторизацию на любом компьютере. Для работы необходим лишь браузер.

Каждый участник проекта, таким образом, в любой момент видит текущие общие результаты в полном объёме. Есть возможность корректив и планирования собственной работы с учётом результатов коллег. Доступны различные варианты просмотра, достаточно гибкая фильтрация данных, вывод результатов в слоях. Планируется дальнейшее развитие интерфейсных возможностей.

Таким образом, разработанное веб-приложение обеспечивает совместную работу через браузер и сохранение результатов в серверной базе данных.

Созданный инструментарий позволяет выделить графический образец на выбранной странице рукописи (рис. 4), сопоставить ему соответствующие категории, заполнить иные поля. Выделенный фрагмент сохраняется в отдельном файле. Помимо этого, в базу данных заносятся координаты и размеры фрагмента, что позволит в дальнейшем отмечать отобранные образцы на скане страницы рукописи в интерактивном режиме. Для работы в старой орфографии используется виртуальная клавиатура.

Рис 4. Фрагмент рабочего интерфейса исследователя

Каждому участнику проекта доступен не только просмотр текущего наполнения базы данных (рис. 5), но и возможность скорректировать графические результаты, сопоставив их с результатами коллег.

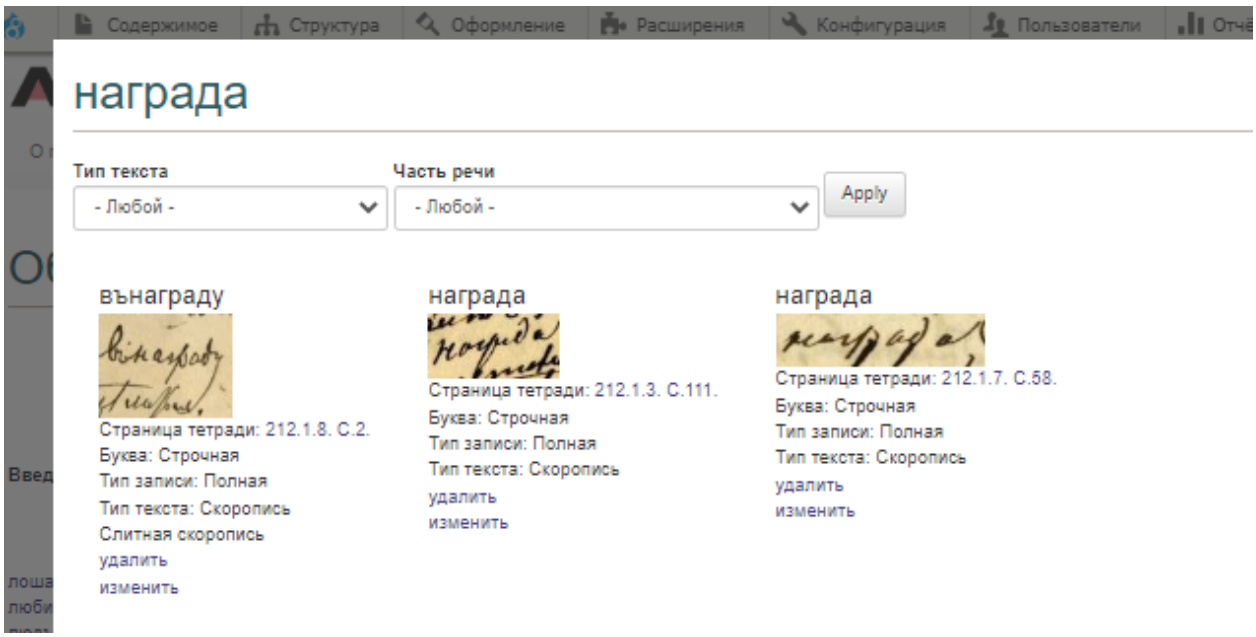

Рис. 5. Просмотр графических образцов, соответствующих основной словоформе

Для обеспечения коллективной работы разработаны дополнительные возможности: просмотр результатов личной и совместной работы, сохранение черновиков, дискуссии по заявленным вопросам и т.д.

Рискнём сказать, что разработанный программный комплекс предоставляет исследователям достаточно удобный и гибкий инструментарий: к настоящему моменту в базе данных представлено более 40000 образцов почерка, отобранных, сохранённых и структурированных посредством разработанного веб-приложения, а также сотни образцов каллиграфии и рисунки.

База данных графических образцов пополняется, развиваются возможности просмотра и фильтрации материалов. Результаты работы предоставят основу для дальнейшей исследовательской работы.

Заключение

Создан и развивается на протяжении ряда лет программный комплекс, обеспечивающий под эгидой ИРЛИ РАН работу текстологов, изучающих рукописный архив Ф.М. Достоевского.

Портал «Архив Ф.М. Достоевского» объединяет результаты нескольких проектов, использующих единый первичный массив данных – цифровые копий страниц девятнадцати рукописей Ф.М. Достоевского из собраний РГАЛИ и ОР РГБ с соответствующими текстовыми расшифровками.

На портале представлены 19 рукописных тетрадей (доступ к материалам требует авторизации), программная интерпретация разметки текстов позволяет в интерактивном режиме выявлять слои авторской правки.

В рамках текущего проекта реализованы программные решения, предоставляющее инструментарий для создания, веб-публикации и гибкого отображения библиотеки графических образцов. Все работы связанные с пополнением корпуса графических образцов, происходит исключительно онлайн.

Работы по созданию программного комплекса, развитию портала и его контентному пополнению поддержаны грантами РНФ:

- Рабочие тетради Ф.М. Достоевского: первая полнотекстовая публикация автографов в их динамической транскрипции. Грант РНФ № 16-18-10034. 2016—2018 (руководитель: С. А. Кибальник)
- Рабочие тетради Ф.М. Достоевского: первая полнотекстовая публикация автографов в их динамической транскрипции. Грант РНФ № 16-18-10034-П. 2019—2020 (руководитель: С. А. Кибальник)
- Новые методы изучения рукописного наследия Ф.М. Достоевского. Грант РНФ № 21-18-00333. 2021—2023 (руководитель: Н.А. Тарасова)

Литература

1. Тарасова Н. А., Мбого И. А., Захаркина В. В. Новые методы изучения творческого наследия Ф. М. Достоевского: на материале цифрового

архива писателя // Неизвестный Достоевский. 2021. Т. 8. № 3. С. 193–248. DOI: 10.15393/j10.art.2021.5662. – URL: https://unknown-dostoevsky.ru/files/redaktor_pdf/1633701239.pdf

References

1. Tarasova N. A., Mbogo I. A., Zakharkina V. V. New Approaches to the Creative Heritage of F. M. Dostoevsky: Based on the Materials from the Writer's Digital Archive. In: Neizvestnyy Dostoevskiy [The Unknown Dostoevsky], 2021, vol. 8, no. 3, pp. 193–248. DOI: 10.15393/j10.art.2021.5662