УДК 004.89:004.912 doi: 10.15827/0236-235X.149.039-046 2025. Т. 38. № 1. С. 39–46

Гибридный подход к выделению структурированных данных из «Летописи жизни и творчества А.С. Пушкина»

П.П. Кокорин [1], А.А. Котов [1], С.В. Кулешов [1], А.А. Зайцева [1]✉

[1] Санкт-Петербургский федеральный исследовательский центр РАН, г. Санкт-Петербург, 199178, Россия

Ссылка для цитирования
Кокорин П.П., Котов А.А., Кулешов С.В., Зайцева А.А. Гибридный подход к выделению структурированных данных из «Летописи жизни и творчества А.С. Пушкина» // Программные продукты и системы. 2025. Т. 38. № 1. С. 39–46. doi: 10.15827/0236-235X.149.039-046
Информация о статье
Группа специальностей ВАК: 2.3.1
Поступила в редакцию: 12.07.2024 После доработки: 13.08.2024 Принята к публикации: 23.08.2024

Аннотация. Статья посвящена решению проблемы создания программной инфраструктуры для систематизации, аннотирования, хранения, поиска и публикации рукописей и иных материалов, представленных в цифровом виде. Исследование построено на материалах, связанных с жизнью и творчеством А.С. Пушкина и составляющих важную часть научно-просветительского ресурса «Пушкин цифровой». Актуальность решаемой проблемы обусловлена необходимостью сохранения авторского наследия русских писателей в условиях цифровой трансформации предметной области филологических, источниковедческих и библиографических исследований их трудов, что является частью национальных проектов Российской Федерации «Образование», «Культура», «Наука и университеты». В данном контексте особую роль играет решение задачи извлечения структурированного текста из растровых изображений страниц томов «Летописи жизни и творчества А.С. Пушкина» для использования в разрабатываемых системах хранения, систематизации, публикации материалов библиотечных, архивных, музейных, фонографических и иных фондов и коллекций и частичной автоматизации филологических, источниковедческих и библиографических исследований. В работе предложен гибридный подход, основанный на использовании априорных данных о структуре элементов верстки страницы, технологиях OCR – распознавание текста на базе библиотеки Tesseract и методах верификации. Особенностью разработанных методов верификации является использование регулярных выражений для извлечения структурированных данных из предварительно распознанного текста и автоматизированного конвейера обработки текстов в сборочной системе GitLab. Приведены результаты применения предложенного гибридного подхода. Показано, что этот подход дает удовлетворительные результаты, обеспечивая минимизацию ручной постобработки полученных данных путем вычитки результатов, размещаемых на научно-просветительском ресурсе. Полученные результаты могут использоваться не только в разрабатываемом ресурсе «Пушкин цифровой», но и в других проектах, в основе реализации которых лежит необходимость распознавания и автоматизированной обработки больших объемов оцифрованных авторских текстов, архивных и других бумажных документов.

Ключевые слова: распознавание текста, извлечение данных, структурированные данные, обработка текста
Благодарности. Работа выполнена в рамках реализации Государственного задания № FFZF-2023-0001

Введение. Сохранение авторского наследия русских писателей в условиях цифровой трансформации предметной области филологических, источниковедческих и библиографических исследований их трудов является частью национальных проектов Российской Федерации «Образование», «Культура», «Наука и университеты». К 225-летнему юбилею А.С. Пушкина с целью сохранения культурного наследия великого поэта разрабатывался научно-просветительский ресурс «Пушкин цифровой». Важная составляющая этой работы – создание программной инфраструктуры для систематизации, аннотирования, хранения, поиска и публикации рукописей и иных материалов, представленных в цифровом виде, связанных с жизнью и творчеством А.С. Пушкина как одного из самых знаковых и узнаваемых во всем мире русских писателей [1–3].

Одним из ключевых источников данных для наполнения портала стала «Летопись жизни и творчества А.С. Пушкина» (далее – Летопись) [4]. К сожалению, исходных машиночитаемых структурированных данных издания не существует. В связи с этим задача извлечения структурированных текстовых данных, пригодных для использования в рамках научно-просветительского ресурса, трансформируется в задачу извлечения текста из растровых изображений страниц томов Летописи. Извлеченные текстовые данные должны быть пригодны для использования в разрабатываемых системах хранения, систематизации, публикации материалов библиотечных, архивных, музейных, фонографических и иных фондов, коллекций и частичной автоматизации филологических, источниковедческих и библиографических исследований.

Оптическое распознавание символов (*Orticaл Caracter Recognition*, OCR) – это процесс преобразования изображений в редактируемый текст. OCR-системы используются для сканирования документов и книг, а также для обработки фотографий и других изображений с текстом. Они могут быть применены для различных целей, включая оцифровку архивов, создание электронных книг, перевод текстов на другие языки и многое другое, а также для автоматизации процессов обработки документов, таких как заполнение форм или проверка данных. Однако, несмотря на все преимущества, OCR-системы неидеальны. Некоторые из них сталкиваются с проблемами распознавания сложных шрифтов или рукописного текста.

Существует большое количество коммерческих [5] и открытых [6, 7] программных средств, позволяющих производить OCR для документов и книг. При этом большинство таких систем хорошо справляются с переводом языковых токенов (отдельных слов и фрагментов предложений) в редактируемый текст, но имеют большое количество ошибок при распознавании и воспроизведении структуры (таблиц, абзацев, колонок), связанных с особенностями типографской верстки. Причем большинство исследователей в качестве основного инструмента распознавания используют нейросетевые модели, применяемые с различной степенью качества. Например, в [8] приведена одна из возможных реализаций решения задачи автоматизированного распознавания сущностей, основанная на дообучении языковой модели на архитектуре BERT, подключенной к библиотеке Spacy с использованием Spacy Transformers. Подобные нейросетевые инструменты ограниченно применимы в ситуациях, когда требуется точное соответствие извлекаемых данных структуре документа, поскольку разрабатываемый портал должен отвечать требованиям энциклопедичности и академичности.

Похожая проблема возникает при формировании машиночитаемых электронных словарей, например, в [9] распознавание макроструктуры и микроструктуры словарей основано на выделении границ словарных статей, границ зон внутри словарных статей в исходном тексте и их классификации. При этом авторы делают вывод о необходимости постоянной доработки алгоритмов под конкретную верстку словаря либо проверки распознанных данных вручную. Для облегчения разработки таких словарей в [10, 11] определены правила формирования терминологических кластеров и сетей.

Некоторые варианты атрибутирования текстов на основе анализа структуры рукописей по их отсканированным изображениям, а также возникающие проблемы описаны в [12].

В работе [13] сделана попытка перейти от посимвольного распознавания русскоязычных рукописных текстов к распознаванию строк текста с использованием нейросетевых подходов. Наиболее интересным с точки зрения сохранения структуры распознаваемых документов представляется описанный в [14] подход, направленный на распознавание русских рукописных текстов целыми абзацами. При этом результатом работы предложенных алгоритмов все равно является неструктурированный текст без выделения необходимых в отдельных случаях структурных элементов.

Для сокращения количества ошибок при автоматизированной обработке предлагается гибридная схема, использующая одновременно результат оптического распознавания текста всего документа в целом, не отражающий сведения о структурной разметке текста, а также совокупность результатов распознавания текста, отдельных фрагментов текста согласно шаблону страницы (рис. 1). Результаты объединения текстов сохраняются в БД, содержащей пригодные для использования на портале поля.

Особенности исходного материала

Текст Летописи имеет сложную двухколоночную типографскую верстку: 1-я колонка – временной отрезок, место события и описание события Летописи; 2-я колонка – примечания, ссылки на библиографию и другие временные отрезки Летописи.

При этом в форматировании используется чередование (зеркальное отображение) колонок на четных и нечетных страницах (http://www.swsys.ru/uploaded/image/2025-1/Kokorin.html). Смена текущего года отмечается в верстке книги появлением отдельного абзаца с цифрами года (http://www.swsys.ru/uploaded/image/2025-1/23.jpg).

Кроме того, в верхнем и нижнем колонтитулах содержится дополнительная информация:

– текущий год Летописи;

– текущее место пребывания Пушкина А.С. в это время;

– номер страницы в пределах тома;

– сквозной номер страницы в пределах всех томов.

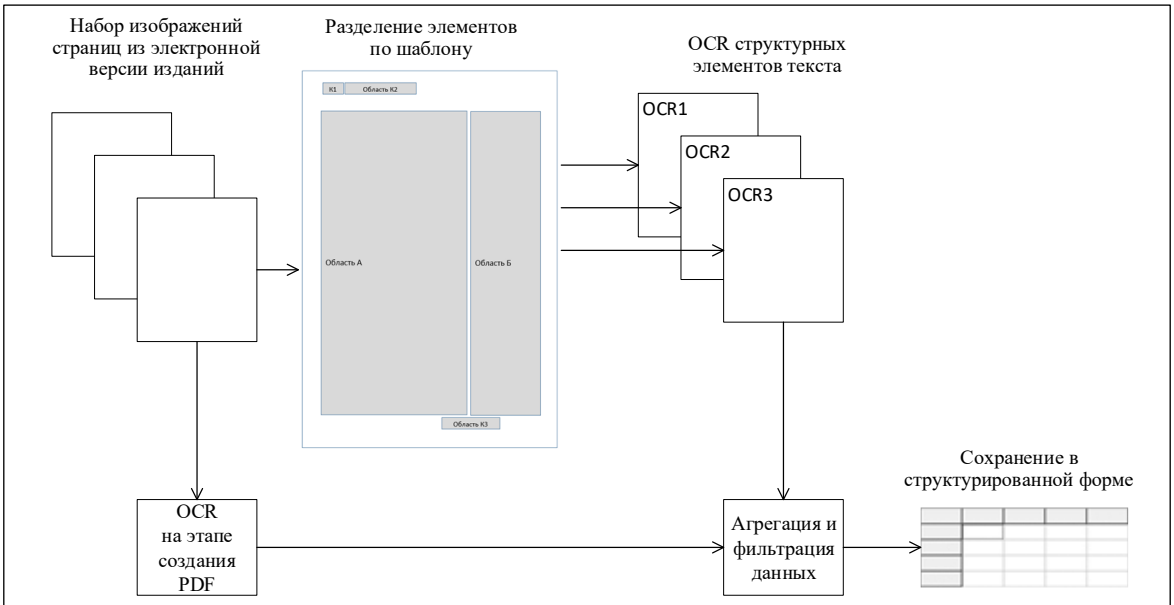

Рис. 1. Место оптического распознавания в рамках задачи выделения структурированных данных

Fig. 1. Place of optical recognition within the structured data extraction task

Все эти поля, а также соответствующий им текст, описывающий события, должны быть сохранены в БД для портала.

Оптическое распознавание текста

Для предварительного извлечения структурированных данных из текста используются априорные знания о структуре страницы, задаваемые шаблоном верстки. Пример шаблона расположения областей документа для четных страниц приведен на рисунке 2. Для каждой области вызывается внешний модуль оптического распознавания (библиотека Tesseract), работающий с изображениями отдельной области. Конкретные размеры и положения областей распознавания корректируются эвристическим алгоритмом поиска областей, не заполненных текстом, между областями распознавания. После получения результата от модуля оптического распознавания текста записи Летописи и текста примечания (рис. 2) сопоставляются запись Летописи (1-я колонка, область А) с примечанием (2-я колонка, область Б); место, время и номера страниц выделяются из областей колонтитула (К1, К2, К3).

Такая верстка текста ориентирована на чтение человеком и малопригодна для автоматизированной компьютерной обработки, что и составляет основную сложность.

Сложность сопоставления записи Летописи и комментария определена различными вариантами взаимного выравнивания абзацев текста (http://www.swsys.ru/uploaded/image/2025-1/24.jpg): по верхнему краю текста Летописи и текста примечания, по середине – текста Летописи.

Извлечение структурированных данных из текста

После первичного оптического распознавания формируется текстовый документ, содержащий ошибки распознавания и разметки.

Задача извлечения структурированных данных сводится к извлечению из распознанного текста списка записей Летописи с текстом примечания. Каждая запись Летописи должна сопровождаться фиксированным набором атрибутов:

– номер страницы в томе Летописи;
– сквозной номер страницы в пределах всех томов;
– год;
– год и место из верхнего колонтитула страницы;
– диапазон датировок записи;
– текст записи Летописи;
– текст связанного примечания.

Формат диапазона дат имеет большое количество вариантов:

– год (например, 1799);
– год, месяц, день (например, 1800. Декабрь, 2);
– год, месяц, день и время дня;

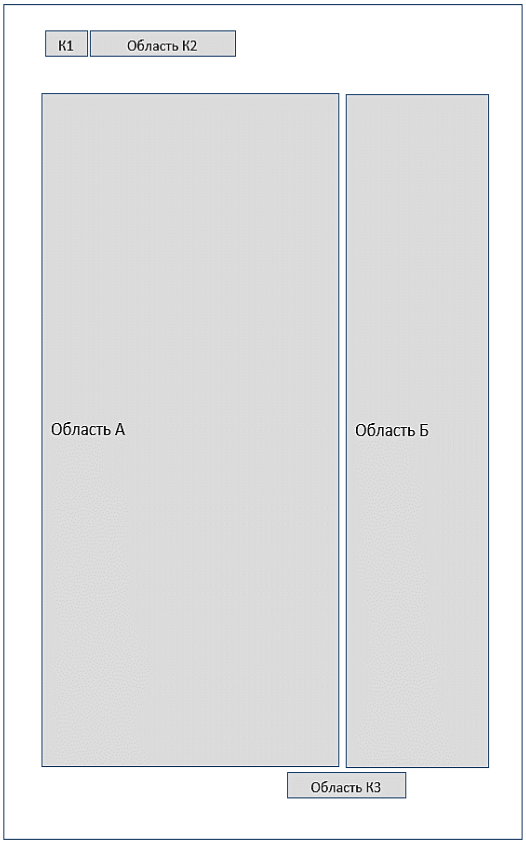

Рис. 2. Шаблон областей для применения OCR на четных страницах

Fig. 2. Area pattern for applying OCR on even-numbered pages

– месяц, день (например, Январь, 9).

Спецификаторы и сложные структуры дат:

– спецификатор неточной датировки; любой элемент датировки может сопровождаться признаком неточной датировки – (?) (например, 1825 (?). Февраль, 1...3, Январь (?), Февраль, 1(?)...13(?), 1825(?). Февраль(?)...Май(?));

– слово «или» – два и более вариантов дат (например, Сентябрь, до 17 или Октябрь, после 14.);

– спецификаторы верхней/нижней границы дат – слова «не ранее, не позднее, до, конец, начало, после, первая половина» и т.п.; обозначают диапазон дней в месяце (например, Январь, после 18...Март, до 9, Ноябрь, перв. пол. (?));

– символ «/» – используется для разделения даты в григорианском и юлианском календарях (например, 1825. Декабрь, 21/1826. Январь, 2(?));

– символы «...», «–» – диапазон дат (например, 1829–1830 (?), Февраль, 1...10, Апрель, 4–5, Сентябрь, 25 (?)...Октябрь, 5 (?));

– спецификатор времени суток – утро, день, вечер, ночь (например, Январь, ночь с 27-го на 28-е).

Строки с датой могут содержать несколько значений, например:

– 1825. Декабрь, 21/1826. Январь, 2(?)...1826. Июнь, 24/Июль, 6(?);

– Август, перв. пол. ...Сентябрь, до 17 или Октябрь, после 14...Декабрь.

Для извлечения из текста записи Летописи строки, содержащей датировку, разработан набор регулярных выражений. Приведенные правила вывода строк достаточно хорошо покрываются возможностями регулярных выражений [12], благодаря чему удалось избежать разработки и применения специализированных грамматик.

При реализации алгоритмов выделения текста датировок возник ряд сопутствующих проблем, связанных с ошибками OCR:

– цифры: 1 – !, 3 – З, 3 - } и т.п.;

– скобки: (– <;

– знаки пунктуации: точки и запятые;

– месяцы: Сеитябрі.

Наиболее частые ошибки OCR были устранены путем их внесения в регулярные выражения для автоматической замены.

Ограничения на диапазоны с вариантами типовых ошибок распознавания:

– год Летописи: 1799–1836. r'1799|18[0-2][0-9Зз]|18[0-9Зз][0-7]';

– месяц: r'(?:Январь|Февраль|Март|Апрель|
Май|Июнь|Июль|Авгу[се]т|Сентябрь|Октябрь|
Ноябрь|Декабрь)';

– день месяца: 1 – 31. r'[0-9!$}Зз]|[1!][0-9!$Ф}Зз]|2[0-9!$Ф}Зз]|[3}Зз][0-1!]'.

Ограничения библиотеки Tesseract не позволяют обращаться к пользовательским словарям на этапе оптического распознавания, поэтому такие контекстные ошибки также пришлось учесть в регулярных выражениях.

Реализация регулярных выражений

Реализовать механизм регулярных выражений для извлечения структурированных данных из текста позволяют язык программирования Python 3 и набор стандартных библиотек [15].

Особенности технологии реализации данного механизма отражены во фрагменте кода:

```
   _regex_years_range = r'1799|18[0-
2][0-9Зз]|18[0-9Зз][0-7]'
_regex_years_range_group =
```

```
f'(?:{_regex_years_range})'
_regex_months_range_group = r'(?:Ян-
варь|Февраль|Март|Ап-
рель|Май|Июнь|Июль|Авгу[се]т|Сен-
тябрь|Октябрь|Ноябрь|Декабрь)'
_regex_day_in_month_range = r'
[0-9!$}Зз]|[1!][0-9!$Ф}Зз]
|2[0-9!$Ф}Зз]|[3}Зз][0-1!]'

# Октябрь, 22.
# Ноябрь. 28(?).
# Март, 2/14.
# Январь, после 18.
# Ноябрь, 28(?).
# Ноябрь (?), 28(?).
_regex_day_in_month_prefix_text =
r',|/|до|ок\.|ок\.\s*\(не\s*позднее\)
|по[се]ле|не\s*ран(?:е|ьш)е|не\s*позже'
_regex_month_part_group =
r'(?:нач\.?|начало|конец|втор(?:ая|\.)\
s*пол\.?|перв(?:ая|\.)\s*пол(?:овина|\.
?)|перв(?:ые|\.)\s*чи[се]ла|(?:ок\.|до)
?\s*сер\.?|посл\.\s*числа|не\s*позд-
нее)'
_regex_day_in_month_suffix_text = f'-
е\s*{_regex_question_date_group}?\
s*числа'
```

Разработка методики оценки ошибок и построение автоматизированного конвейера обработки текстов

Для оценки качества работы алгоритмов оптического распознавания и алгоритмов выделения структурированного текста создана методика оценки ошибок и построен автоматизированный конвейер обработки текстов в сборочной системе GitLab [16]. Конвейер реализует следующие шаги.

1. На вход конвейера подаются наборы растровых изображений страниц томов Летописи.
2. На следующем этапе для каждого тома Летописи производится оптическое распознавание растровых изображений в текст файла в формате CSV (ocr\timeline\tesseract\T1out.csv, ocr\timeline\tesseract\T2out.csv, ocr\timeline\tesseract\T3out.csv, ocr\timeline\tesseract\T4out.csv, колонки: text, comment_text, year_place, page_num_global, page_num_local, year).
3. Далее из поля text с помощью регулярных выражений выделяется диапазон дат.
4. Обработанные записи сохраняются в файлах в формате CSV (T1.csv, T2.csv, T3.csv, T4.csv, колонки: page_num, page_num_continuous, year, year_place, timeline_range, text, comment_text).
5. Дополнительно вычисляется статистика обработки текста, которая сохраняется в файл статистики:

– подсчет общего количества записей Летописи в файле;

– подсчет количества записей с пустым полем диапазона дат;

– подсчет количества записей, где в поле текста Летописи встречаются дополнительные вхождения по регулярному выражению диапазона дат; эта проверка нужна для оценки качества структуризации текста Летописи и вычисления количества ошибочно объединенных записей;

– подсчет количества записей, где длина поля диапазона дат превышает заданный порог; необходим для оценки слишком длинных диапазонов дат, вероятно, с ошибочно выделенным текстом диапазона дат;

– запись в файл статистики выявленных аномальных записей.

6. Автоматически сравниваются результирующие файлы T1.csv, T2.csv, T3.csv, T4.csv и файлы статистики с результатами предыдущей успешной сборки конвейера, результаты сохраняются в файлы T1.diff, T2.diff, T3.diff, T4.diff, T1_stat.diff, T2_stat.diff, T3_stat.diff, T4_stat.diff.

Применение описанного автоматизированного конвейера позволяет контролировать влияние изменений в алгоритмах на всех этапах обработки на качество конечного результата обработки на корпусе входных данных (томов Летописи), добиваясь минимального количества ошибок в выходных данных.

Таким образом, разработчик-исследователь может принять решение о добавлении новых форматов в регулярные выражения, внесении изменений в процедуру коррекции промежуточных результатов оптического распознавания и повторном запуске конвейера, тем самым обеспечивая повышение скорости обработки Летописи и точности получаемых результатов.

Выводы

В работе рассмотрена задача извлечения структурированных текстовых данных, пригодных для использования в рамках научно-просветительского ресурса, особенностью которой является необходимость использования управляемого оптического распознавания текста с последующей агрегацией полученных данных. Предложена гибридная схема, использующая одновременно результат оптического распознавания текста всего документа в целом, а также совокупность результатов распознавания фрагментов текста согласно шаблону страницы. Для извлечения структурированных дан-

ных из текста использован механизм регулярных выражений. Предложенная методика автоматизации позволяет в процессе работы оценить реальный процент ошибок, связанных с неправильным оптическим распознаванием символов. Например, при ошибке в распознавании месяца или года метод оценки не сможет обнаружить факт повторного вхождения диапазона дат в текст Летописи, что является признаком неверного разделения ее записей. Кроме того, нет возможности автоматической оценки корректности сопоставления записи Летописи и текста комментария. Для оценки качества работы алгоритмов оптического распознавания и алгоритмов выделения структурированного текста разработана методика оценки ошибок и построен автоматизированный конвейер обработки текстов в сборочной системе GitLab.

Список литературы

1. Kassab K., Teslya N. An approach to a linked corpus creation for a literary heritage based on the extraction of entities from texts, Applied Sci., 2024, vol. 14, no. 2, art. 585. doi: 10.3390/app14020585.

2. Тесля Н.Н., Жарков В.М. Структурирование библиотеки произведений А.С. Пушкина через создание базы данных для научно-просветительского портала «Пушкин цифровой» // Наука и технологии в трансформации социально-экономического ландшафта: сб. тр. Междунар. конф. TECHNOPERSPECRIVE 2023. 2024. С. 168–171.

3. Тесля Н.Н., Сиповский Г.В. Разработка алгоритма сопоставления сущностей по описательным характеристикам их названий // Наука и технологии в трансформации социально-экономического ландшафта: сб. тр. Междунар. конф. TECHNOPERSPECRIVE 2023. 2024. С. 127–134.

4. Летопись жизни и творчества А.С. Пушкина: в 4 т. М., 1999.

5. Tereshchenko V., Rybkin V., Shamis A., Yan D. Principles of handwriting recognition in FineReader. Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications, 1998, vol. 8, no. 3, pp. 456-457. URL: https://www.elibrary.ru/item.asp?id=16316347 (дата обращения 11.07.2024).

6. Tafti A.P., Baghaie A., Assefi M., Arabnia H.R., Yu Z., Peissig P. OCR as a service: An experimental evaluation of google docs OCR, Tesseract, ABBYY FineReader, and Transym. In: LNIP. Proc. ISVC, 2016, vol. 10072, pp. 735–746. doi: 10.1007/978-3-319-50835-1_66.

7. Нейфельд О.А., Рудникович А.С. Возможности OCR-библиотеки Tesseract при распознавании текста по видеоизображениям // Сб. избранных статей науч. сессии ТУСУР. 2019. № 1-2. С. 16–19.

8. Prelevikj M., Žitnik S. Multilingual named entity recognition and matching using BERT and Dedupe for Slavic languages. Proc. 8th Workshop on Balto-Slavic Natural Language Processing, 2021, pp. 80–85.

9. Беляева Л.Н., Ефремова А.Н. Автоматическая компиляция базы данных комплексного электронного словаря // Вестн. ВГУ. Сер.: Лингвистика и межкультурная коммуникация. 2015. № 3. С. 42–45.

10. Мальковский М.Г., Соловьев С.Ю. Правила формирования терминологических кластеров // Открытые семантические технологии проектирования интеллектуальных систем. 2014. № 4. С. 169–172.

11. Мальковский М.Г., Соловьев С.Ю. От терминологических сетей к толковым словарям // Открытые семантические технологии проектирования интеллектуальных систем. 2015. № 5. С. 281–284.

12. Чевтаев А.А. Формирование цифровых баз данных рукописей: проблемы и текстологические перспективы. Статья 1 // Новый филологический вестн. 2019. № 1. С. 28–43.

13. Coquenet D., Chatelain C., Paquet T. End-to-end handwritten paragraph text recognition using a vertical attention network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, vol. 45, no. 1, pp. 508–524. doi: 10.1109/tpami.2022.3144899.

14. Mohammed S., Teslya N. Handwritten paragraph recognition using spatial information on Russian notebooks dataset. Proc. FRUCT, 2023, pp. 108–113. doi: 10.23919/FRUCT60429.2023.10328173.

15. Фридл Дж. Регулярные выражения; [пер. с англ.]. СПб: Символ-Плюс, 2008. 608 с.

16. Преображенская Т.В., Симонов В.И. Автоматизированное тестирование в среде GitLab CI/CD // Флагман науки. 2023. № 4. С. 803–806.

Software & Systems doi: 10.15827/0236-235X.149.039-046 2025, 38(1), pp. 39–46

Extracting structured data from the Chronicle of the Life and Work of A.S. Pushkin: Hybrid approach

Pavel P. Kokorin [1], Aleksandr A. Kotov [1], Sergey V. Kuleshov [1], Aleksandra A. Zaytseva [1]✉

[1] St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, 199178, Russian Federation

For citation

Kokorin, P.P., Kotov, A.A., Kuleshov, S.V., Zaytseva, A.A. (2025) 'Extracting structured data from the Chronicle of the Life and Work of A.S. Pushkin: Hybrid approach', *Software & Systems*, 38(1), pp. 39–46 (in Russ.). doi: 10.15827/0236-235X.149.039-046

Article info
Received: 12.07.2024 After revision: 13.08.2024 Accepted: 23.08.2024

Abstract. The paper discusses the problem of creating a software infrastructure for systematization, annotation, storage, search and publication of manuscripts and other digital materials. The research focuses on the materials related to the life and work of A.S. Pushkin. These materials form an important part of the scientific and educational resource “Pushkin Digital”. The problem is relevant due to the need to preserve the Russian author's heritage under conditions of digital transformation of philological, source and bibliographic studies into their works. This is a part of the national projects of the Russian Federation “Education”, “Culture”, “Science and Universities”. It is especially important to extract a structured text from bitmap images of pages from A.S. Pushkin's Chronicle of Life and Work volumes to use it in the developing systems of storage, systematization, publication of library, archival, museum, phonographic and other funds and collections and partial automation of philological, source and bibliographic research. The paper proposes a hybrid approach based on the a priori data about the structure of page layout elements, OCR technologies (text recognition based on Tesseract library) and verification methods. The peculiarity of the developed verification methods is using regular expressions for extracting structured data from pre-recognized text and automated text processing pipeline in the GitLab assembly system. The paper demonstrates satisfactory results of the proposed hybrid approach. The approach minimizes the manual post-processing of the obtained data by proofreading the results posted on the research and education resource. The results are useful not only in the research and educational resource Pushkin Digital under development, but also in other projects, which require recognition and automated processing of large volumes of digitized author's texts, archival and other paper documents.
Keywords: text recognition, data extraction, structured data, text processing
Acknowledgements. The work was carried out within the framework of implementing State Task no. FFZF-2023-0001

References

1. Kassab, K., Teslya, N. (2024) ‘An approach to a linked corpus creation for a literary heritage based on the extraction of entities from texts’, *Applied Sci.*, 14(2), art. 585. doi: 10.3390/app14020585.

2. Teslya, N.N., Zharkov, V.M. (2024) ‘Structuring the library of works by A.S. Pushkin through the creation of a database for the scientific and educational portal "Pushkin Digital"’, *Proc. Int. Conf. TECHNOPERSPECRIVE 2023*, pp. 168–171 (in Russ.).

3. Teslya, N.N., Sipovsky, G.V. (2024) ‘Development of an algorithm for matching entities based on the descriptive characteristics of their names’, *Proc. Int. Conf. TECHNOPERSPECRIVE 2023*, pp. 127–134 (in Russ.).

4. (1999) *Chronicle of the life and work of A.S. Pushkin: In 4 vol*. Moscow (in Russ.).

5. Tereshchenko, V., Rybkin, V., Shamis, A., Yan, D. (1998) ‘Principles of handwriting recognition in FineReader’, *Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications*, vol. 8, no. 3, pp. 456-457, available at: https://www.elibrary.ru/item.asp?id=16316347 (accessed July 11, 2024).

6. Tafti, A.P., Baghaie, A., Assefi, M., Arabnia, H.R., Yu, Z., Peissig, P. (2016) ‘OCR as a service: An experimental evaluation of google docs OCR, Tesseract, ABBYY FineReader, and Transym’, in *LNIP. Proc. ISVC*, 10072, pp. 735–746. doi: 10.1007/978-3-319-50835-1_66.

7. Neifeld, O.A., Rudnikovich, A.S. (2019) ‘Capabilities of the Tesseract OCR library in text recognition from video images’, *Proc. Collection of Selected Papers from the Scientific Session of TUSUR*, (1-2), pp. 16–19 (in Russ.).

8. Prelevikj, M., Žitnik, S. (2021) ‘Multilingual named entity recognition and matching using BERT and Dedupe for Slavic languages’, *Proc. 8th Workshop on Balto-Slavic Natural Language Processing*, pp. 80–85.

9. Belyaeva, L.N., Efremova, A.N. (2015) ‘Automatic compilation of the database of a comprehensive electronic dictionary’, *Proc. of VSU. Ser.: Linguistics and Intercultural Communication*, (3), pp. 42–45 (in Russ.).

10. Malkovsky, M.G., Soloviev, S.Yu. (2014) ‘Rules for terminological clusters creations’, *OSTIS*, (4), pp. 169–172 (in Russ.).

11. Malkovsky, M.G., Soloviev, S.Yu. (2015) ‘From terminological networks to the explanatory dictionaries’, *OSTIS*, (5), pp. 281–284 (in Russ.).

12. Chevtaev, A.A. (2019) ‘Formation of digital databases of manuscripts: Problems and textological perspectives. Article 1’, *The New Philological Bull*., (1), pp. 28–43 (in Russ.).

13. Coquenet, D., Chatelain, C., Paquet, T. (2023) ‘End-to-end handwritten paragraph text recognition using a vertical attention network’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), pp. 508–524. doi: 10.1109/tpami.2022.3144899.

14. Mohammed, S., Teslya, N. (2023) ‘Handwritten paragraph recognition using spatial information on Russian notebooks dataset’, *Proc. FRUCT*, pp. 108–113. doi: 10.23919/FRUCT60429.2023.10328173.

15. Friedl, J.E.F. (2006) *Mastering Regular Expressions*. O'Reilly Media Publ., 535 p. (Russ. ed.: (2008) St. Petersburg, 608 p.).

16. Preobrazhenskaya, T.V., Simonov, V.I. (2023) 'Automated testing in the GitLab CI/CD environment', *Flagship Science*, pp. 803–806 (in Russ.).

Авторы

Кокорин Павел Петрович [1], к.т.н.,
старший научный сотрудник,
kokorin.p@iias.spb.su
Котов Александр Александрович [1],
младший научный сотрудник,
alexanderkotovspb@gmail.com
Кулешов Сергей Викторович [1], д.т.н.,
профессор РАН, главный научный сотрудник,
kuleshov@iias.spb.su
Зайцева Александра Алексеевна [1],
к.т.н., старший научный сотрудник,
cher@iias.spb.su

[1] Санкт-Петербургский федеральный
исследовательский центр РАН,
г. Санкт-Петербург, 199178, Россия

Authors

Pavel P. Kokorin [1],
Cand. of Sci. (Engineering),
Senior Researcher, kokorin.p@iias.spb.su
Aleksandr A. Kotov [1],
Junior Researcher,
alexanderkotovspb@gmail.com
Sergey V. Kuleshov [1],
Dr.Sci. (Engineering), Professor RAS,
Chief Researcher, kuleshov@iias.spb.su
Aleksandra A. Zaytseva [1],
Cand. of Sci. (Engineering),
Senior Researcher, cher@iias.spb.su

[1] St. Petersburg Federal Research Center
of the Russian Academy of Sciences,
St. Petersburg, 199178, Russian Federation