

## Подобрен достъп до дигитално съдържание в Народна библиотека „Иван Вазов“ – Пловдив

Иван Крачанов,  
ръководител Дигитален център  
Народна библиотека „Иван Вазов“, Пловдив

## Improved access to digital content in National Library „Ivan Vazov“ – Plovdiv

Ivan Kratchanov,  
Head of Digitization Centre  
National Library Ivan Vazov, Plovdiv

**Abstract:** *The report will present an overview of the ongoing efforts in the field of digitization in National Library „Ivan Vazov“ (NLIV) by focusing on how the library transforms its digital collections in accordance with current standards for accessibility of modern information resources. Ensuring easier access and more efficient use of digital content is particularly important now, marked by the COVID-19 pandemic.*

*In 2019, the digital library of NLIV introduced the ability to add PDF files and index their content, with the possibility to search in the text, which provided users with a very useful opportunity for more effective discovery and use of the content. Digitally created files only need to be added to the digital library to be fully accessible, but they represent a very small fraction of all digital objects, most of which are scanned paper sources, which in their original digital form are images (most often .tiff or .jpg files), i.e. the text is not machine-readable and can only be read on a computer screen. It is therefore necessary to take the next step in extracting machine-readable text from the images, which requires the use of optical character recognition (OCR) software. OCR technology is an extremely useful and accessible tool, but it should be noted that recognition is very rarely error-free and its success depends on a number of characteristics of the original document – condition, language, font and more. After an in-depth study of the available software solutions for performing OCR and the experience of libraries from Bulgaria, Russia, Ukraine and Serbia, NLIV decided to purchase the software product ABBYY FineReader. Using it with master files, obtained from well-preserved originals, in modern Bulgarian language, gives a very high success rate of OCR, most often above 99%. However, most of the digitized materials, those of high cultural and historical value and expired copyright, are predominantly from the period before the*

*Orthographic Reform of 1945, which poses a number of problems to the accuracy of OCR.*

*National Library „Ivan Vazov“ is currently participating in the CLADA-BG project, which is integrated within the European infrastructures CLARIN and DARIAH. The mission of CLADA-BG is to create a national technological infrastructure for resources for the linguistic, cultural and historical heritage and technologies that will provide public access to these resources, tools for processing Bulgarian language and tools for access and management of CHH datasets for various societal tasks, targeted at wide audience. Participants are grouped into two categories: content providers and technology partners. One of the areas of the project, in which the library participates as a content provider, is the development of the best methodology for OCR and the corresponding improvement of the search methods in text resources.*

*The result of our work on the project, in cooperation with the Institute of Information and Communication Technologies at BAS (IICT-BAS), was the preparation and testing of a dictionary of old Bulgarian word forms, the purpose of which is to aid OCR. The dictionary was named CLADABG-MODEL by its creators at IICT-BAS. The test was conducted in the period March–April 2020. Its purpose was to determine the extent to which the dictionary with old word forms helps the software program in performing OCR of printed texts in Bulgarian language before the Orthographic Reform of 1945. ABBYY FineReader relies on dictionaries for improving the quality of recognition by reinforcing hypotheses about „uncertain“ words by finding them in the dictionary. The program has a built-in dictionary only for the modern Bulgarian language, but allows the addition of personalized, external dictionaries. CLADABG-MODEL, which contains 1,121,872 words in use before the Orthographic Reform of 1945, was developed to test the hypothesis that its use would lead to a higher recognition rate. The amount of the increase, if any, also had to be determined. The main indicator is the percentage of unrecognized words in relation to the total number of words. The test results showed a slight increase in the success of OCR with the use of CLADABG-MODEL, which encouraged the team to continue working and possible future improvements were identified, such as the inclusion of personal names and surnames.*

**Keywords:** digitization, cultural heritage, digital library, optical character recognition, Bulgarian language

**Резюме:** Докладът ще представи преглед на текущите усилията в полето на дигитализацията в Народна библиотека „Иван Вазов“ (НБИВ), като акцентира на това как библиотеката трансформира своите дигитални колекции, отговаряйки на настоящите стандарти за достъпност на съвременните информационни ресурси. Осигуряването на улеснен дос-

търп и по-ефективно използване на дигиталното съдържание са особено важни в настоящото време, белязано от пандемията COVID-19.

През 2019 г. в дигиталната библиотека на НБИВ беше въведена възможност за добавяне на PDF файлове и индексирание на съдържанието им, като по този начин стана възможно търсенето в текста, с което на потребителите беше предоставена изключително полезната възможност за по-ефективно разкриване и използване на съдържанието. За създадените в цифров вид файлове е необходимо само тяхното добавяне в дигиталната библиотека, за да бъдат използвани пълноценно, но те представляват много малка част от всички дигитални обекти, повечето от които са сканирани хартиени източници, които в първоначалния си дигитален вариант представляват изображения (най-често .tiff или .jpg файлове), т.е. текстът не е машинно четим и може да се използва само за четене на компютърния екран. За да бъдат използвани максимално ефективно е необходимо да се предприеме следващата стъпка по извличането на машинно четим текст от изображенията, която изисква използването на софтуер за оптично разпознаване на символи (OCR). OCR технологията е изключително полезен и достъпен инструмент, но трябва да се има предвид, че разпознаването много рядко е безгрешно и успеваемостта му зависи от редица характеристики на оригиналния документ – състояние, език, шрифт и др. След задълбочено проучване на достъпните софтуерни решения за извършване на OCR и на опита на библиотеки от България, Русия, Украйна и Сърбия, НБИВ взе решение за закупуването на софтуерен продукт ABBYY FineReader. Използването му с мазтър файлове, получени от добре запазени оригинали, на съвременен български език, дава много висок процент успеваемост на OCR, най-често над 99%. По-голямата част от материалите обаче, обект на дигитализация, тези с висока културна и историческа стойност и с изтекъл срок на авторско право, са предимно от периода преди Правописната реформа от 1945 г., което поставя редица проблеми пред точността на OCR.

Народна библиотека „Иван Вазов“ понастоящем участва в проект КЛаДА-БГ, който е интегриран в рамките на европейските инфраструктури CLARIN и DARIAH. Мисията на КЛаДА-БГ е да създаде национална технологична инфраструктура за ресурси за езиковото, културно-историческото наследство и технологии, които да осигурят публичен достъп до тези ресурси, инструменти за обработка на българския език и инструменти за достъп и управление на масивите от данни на културно-историческото наследство за различни обществени задачи, насочени към широката аудитория. Участниците са групирани в две категории: доставчици на съдържание и технологични партньори. Едно от направ-

ленията по проекта, в което библиотеката участва като доставчик на съдържание, е разработването на най-добрата методология за OCR и съответно подобряване методите за търсене в текстовия ресурс.

Резултат от нашата работа по проекта, в сътрудничество с Института за информационни и комуникационни технологии към БАН (ИИКТ-БАН), беше изготвянето и тестването на речник на старите български словоформи, целта на който е да се подпомогне OCR. Речникът е наречен CLADABG-MODEL от създателите му в ИИКТ-БАН. Тестът беше проведен в периода март–април 2020 г. Целта му беше да се определи до каква степен речникът със стари словоформи подпомага софтуерната програма при извършване на OCR на печатни текстове на български език преди Правописната реформа от 1945 г. ABBYY FineReader разчита на речници за повишаване на качеството на разпознаване чрез затвърждаване на хипотези за „несигурни“ думи, чрез намирането им в речника. Програмата разполага с вграден речник само за съвременния български език, но позволява добавянето и на персонализирани, външни речници. CLADABG-MODEL, който съдържа 1 121 872 думи, използвани преди правописната реформа от 1945 г., беше разработен с цел да се тества хипотезата, че използването му ще доведе до по-висок процент на разпознаване. Размерът на увеличението, ако има такова, също трябваше да бъде определен. Основният метричен показател е процентът на неразпознатите думи по отношение на общия брой думи. Резултатът от теста показва леко увеличение на успеха на OCR с използването на CLADABG-MODEL, което окуражи екипа да продължи с работата и бяха набелязани са някои възможни бъдещи подобрения, като например включването на лични имена и фамилии.

**Ключови думи:** дигитализация, културно-историческо наследство, дигитална библиотека, оптическо разпознаване на символи, български език

## **Въведение**

Основна задача на Националната библиотека „Иван Вазов“ — Пловдив (НБИВ) е да осигури достъп на учени и читатели до дигитализирано съдържание. Дистанционният достъп става особено важен в настоящия момент, белязан от ограниченията, които налага COVID-19 пандемията. С увеличаването на потребността от достоверни е-източници дигиталните библиотеки се превръщат в ключово средство за набавяне на висококачествени електронни

книги, периодика и образователни материали – тенденция, която се потвърждава от статистиката на водещи световни електронни библиотеки (Falt, E., Das, P. P., 2020).

Докладът ще представи преглед на текущите усилията в полето на дигитализацията в Народна библиотека „Иван Вазов“ (НБИВ), като акцентира на това как библиотеката трансформира своите дигитални колекции, отговаряйки на настоящите стандарти за достъпност на съвременните информационни ресурси.

През 2019 г. в дигиталната библиотека на НБИВ беше въведена възможност за добавяне на PDF<sup>1</sup> файлове и индексирание на съдържанието им, като по този начин стана възможно търсенето в текста, с което на потребителите беше предоставена изключително полезната възможност за по-ефективно разкриване и използване на съдържанието. Важно предимство на PDF (напр. пред .docx формата на MS Word) е това, че представя документите по един и същи начин, независимо от хардуерните и софтуерните характеристики на системата. Също така, един файл може да съдържа различни по вид елементи, като текст, изображения и др. Това е от особено значение за целите на дигитализацията, защото разпознатият чрез OCR текст може да бъде скрит зад изображението, под формата на невидим заден слой, подравнен със сканирания видим текст. Полезно е и, че има възможност за настройване на степента на компресия на данните с цел получаване на файлове с желан размер.

За създадените в цифров вид PDF файлове е необходимо само тяхното добавяне в дигиталната библиотека, за да бъдат използвани пълноценно, но те представляват много малка част от всички дигитални обекти, повечето от които са сканирани хартиени източници, които в първоначалния си дигитален вариант представляват изображения (най-често .tiff или .jpg файлове), т.е. текстът не е машинно четим и може да се използва само за четене на компютърния екран. За да бъдат използвани максимално ефек-

---

<sup>1</sup> PDF (Portable Document Format) е отворен стандарт за обмен на документи. Създаден е от Adobe Systems през 1993 г., като през 2008 г. става отворен стандарт и е публикуван от Международната организация по стандартизация под номер ISO 32000 – 1:2008.

тивно е необходимо да се предприеме следващата стъпка по извличането на машинно четим текст от изображенията, най-ефективният метод за което е използването на софтуер за оптично разпознаване на символи (OCR). OCR технологията е изключително полезен и достъпен инструмент, но трябва да се има предвид, че разпознаването много рядко е безгрешно и успеваемостта му зависи от редица характеристики на оригиналния документ – състояние, език, шрифт и др. След задълбочено проучване на достъпните софтуерни решения за OCR и на опита на библиотеки от България, Русия, Украйна и Сърбия, НБИВ взе решение за закупуването на софтуерен продукт ABBYY FineReader. Използването му с мазтър файлове, получени от добре запазени оригинали, на съвременен български език, дава много висок процент успеваемост на OCR, най-често над 99%. По-голямата част от материалите обаче, обект на дигитализация, тези с висока културна и историческа стойност и с изтекъл срок на авторско право, са предимно от периода преди Правописната реформа от 1945 г., което поставя редица проблеми пред точността на OCR.

Народна библиотека „Иван Вазов“ понастоящем участва в проект КЛаДА-БГ, който е интегриран в рамките на европейските инфраструктури CLARIN и DARIAH. Мисията на КЛаДА-БГ е да създаде национална технологична инфраструктура за ресурси за езиковото, културно-историческото наследство и технологии, които да осигурят публичен достъп до тези ресурси, инструменти за обработка на българския език и инструменти за достъп и управление на масивите от данни на културно-историческото наследство за различни обществени задачи, насочени към широката аудитория. Участниците са групирани в две категории: доставчици на съдържание и технологични партньори. Едно от направленията по проекта, в което библиотеката участва като доставчик на съдържание, е разработването на най-добрата методология за OCR и съответно подобряване методите за търсене в текстовия ресурс.

### **Подход**

Работата по техническите аспекти на процеса по дигитализация, който включва оптично разпознаване на символи (OCR)

и изисква правилно обработване на стари текстове на кирилица, особено такива, публикувани преди последната голяма правописна реформа от 1945 г., се провежда в рамките на проект КЛА-ДА-БГ и съвместните дейности на библиотеката с Института по информационни и комуникационни технологии към Българската академия на науките (ИИКТ-БАН) за разработване на съответните инструменти и методологии. Целите на сътрудничеството са две: (1) да се извърши правилен OCR на стари текстове и (2) те да се нормализират, т.е. да бъдат приведени към актуалната правописна норма. Първата цел е от съществено значение за онлайн публикуването на оригиналните документи (вестници, списания, книги и др. текстови ресурси). Втората е важна по две основни причини: нормализацията ще даде възможност на потребителите да търсят в корпус, който включва документи от различни периоди с различна ортография, без това да изисква познаването на историческите правописни норми и подаването на няколко отделни заявки за една и съща дума или израз, изписан по различни правила. Освен това създаването на нормализирани копия на старите текстове ще позволи върху тях да се използват NLP инструменти, създадени за съвременния български език.

За постигане на първата цел бяха планирани няколко експеримента. Най-добрият вариант е, разбира се, да се обучи професионален OCR софтуер, който да изпълнява OCR задачи за стар български правопис по възможно най-добрия начин. Затова първият експеримент беше да се обучи системата ABBYY FineReader върху морфологичен речник, предоставен от ИИКТ-БАН. Той съдържа множество от словоформи на думите от основния речников състав на съвременния български, преобразувани в съответствие с предходни правописни норми. Преобразуването е направено с регулярни изрази, които отчитат позицията на дадената буква, кандидат за замяна, нейният непосредствен ляв и десен контекст, позицията на ударението, както и някои релевантни граматически характеристики. Новата „състарена“ версия на морфологичния речник съдържа 1 121 872 словоформи. След като НБИВ приключи с обучението на ABBYY FineReader върху речника, беше направена и оценка на ефективността на модела. Тестове на ефективността

бяха проведени в периода март–април 2020 г. Програмата ABBYY FineReader (вер. 14 и 15) беше използвана за разпознаване на 20 страници от брой 1/1881 г. на списание „Наука“ от фонда на НБИВ, със сигнатура П РЦ-9. Всички страници са цветно сканирани със i2S CopyBook A2 скенер с разделителна способност 300 ppi, 24-битов, TIFF формат, без компресия.

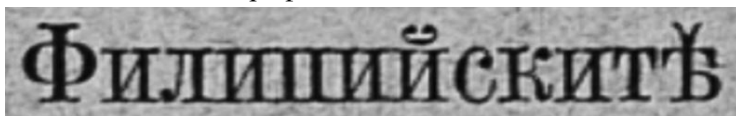
### ***Експерименти и резултати***

Целта на тестовете беше да се определи до каква степен речникът CLADABG-MODEL подпомага софтуерната програма при извършване на OCR на печатни български текстове преди правописната реформа от 1945 г. Както вградените, така и потребителските речници, използвани от ABBYY FineReader, представляват списъци от словоформи на думите, налични в езика на сканираните документи. Програмата разчита на тях за повишаване на качеството на разпознаване, като при наличие на няколко конкуриращи се версии за състава на дадена дума, се предпочита тази, която съвпада с дума, включена в речника. Речниците са особено полезни в случаите, когато текстът съдържа рядко срещани думи (ABBYY Technology Portal, б.д.). ABBYY FineReader има вграден речник само за съвременния български език, докато CLADABG-MODEL съдържа словоформи, част от които съдържат отпаднали буквени символи като **ъ**, **ж** и т.н. съгласно Дринов-Иванчевския правопис; в него са включени и фонетични и словообразователни архаизми.

Много от дигитализираните ценни библиотечни притежания съдържат текст отпреди 1945 г. и целта на разработването на CLADABG-MODEL беше да се тества хипотезата, че използването му ще доведе до по-висок процент на разпознаване. Степента на подобрене, ако има такова, също трябваше да бъде определена. Като основен показател беше използван процентът на погрешно разпознатите думи<sup>2</sup> спрямо общия брой думи. Преброяването се

<sup>2</sup> Погрешно разпознати са думите, при които има несъответствие между буквен символ в първичната дума в сканираното изображение и същия буквен символ в производната машинно четима дума. Не се счита за неправилно разпознаване, ако в първичната дума има правописна грешка и производната дума е с правилно разпознати буквени символи, като по този начин се дублира грешката.

извърши ръчно. В хода на тестването бяха измерени два допълнителни показателя на OCR процеса и на софтуерната програма: степента на разпознаване на изображения в сивата скала (за разлика от тези в цвят) и дали и как отчитаният от ABBYY FineReader параметър „Low-confidence characters“ („Символи с ниска степен на увереност“ – изразен в проценти) може да послужи като индикатор за успешен OCR. Оригиналният хартиен екземпляр на списание „Наука“ е много добре запазен и съответно получените сканирани файлове са с характеристики, близки до оптималните, препоръчвани при оптичното разпознаване. Наблюдава се обаче известно потъмняване на хартията, което намалява контраста и отличителността на буквите. Също така, избраният шрифт (широко използван тогава) затруднява програмата при различаването на букви с доминиращи вертикални линии, като **и**, **п**, **н**, **ш**, **л** (вж. **Фиг. 1**). Хоризонталните линии се събират, буквите се сливат и това допълнително усложнява задачата за алгоритъма за разпознаване. За да се измери успеваемостта на разпознаване с използването на CLADABG-MODEL, бяха сканирани 20 идентични страници, с еднакъв текст и шрифт.



*Фигура 1. Пример за дума със сливащи се буквени символи.*

Общият брой думи е 5485, а средният им брой на страница е 274.25. Направено беше минимално обучение за графично съответствие, което да подпомогне разпознаването на традиционно проблематични символи като **ж**, който без предварително обучение винаги се интерпретира като **ж**. Измерена беше също така успеваемостта на разпознаване при едновременно, комбинирано използване на двата речника – вграденият във ABBYY FineReader български речник и CLADABG-MODEL, като бяха подготвени и заредени два езикови пакета: (1) „Български“ със стандартен, съвременен набор от символи, с вградения във ABBYY FineReader български речник и (2) „Български преди 1945 г.“ с включени старите буквени символи, като **ѣ**, **ж**, **л** и др., и с речника CLADABG-MODEL.

Процент погрешно разпознати думи спрямо общия брой думи		
Вграден речник на FineReader	CLADABG-MODEL	Комбинирани речници
<b>4,90%</b>	<b>4,40%</b>	<b>4,50%</b>

**Таблица 1.** Среден процент на неправилно разпознати думи от 20 цветно сканирани страници, 300 ppi, 24-битов цвят, формат TIFF, без компресия.

Тестването с комбиниране на речниците беше породено от опасението, че когато програмата работи само с CLADABG-MODEL, има риск от по-голям неуспех при разпознаването на думи, налични в съвременния български. Резултатите са обобщени в **Таблица 1**. Те показват, че разпознаването с CLADABG-MODEL е по-добро, въпреки че подобрението не е толкова съществено – средно с 0,5% по-малко грешно разпознати думи. Все пак това показва, че си струва работата в тази посока да продължи.

Вторият тест е свързан със способността на FineReader да разпознава тирето, използвано за означаване на сричкопренасяне (вж. **Таблица 2**). В случай на успешно разпознаване, пренесените думите се запазват цели, което позволява тяхното търсене, копиране и др. Тенденцията за леко подобрение на успеваемостта на разпознаването, когато то е подпомогнато от CLADABG-MODEL, се затвърждава и от резултатите на втория тест.

Брой неразпознати сричкопренасяния		
Вграден речник на FineReader	CLADABG-MODEL	Комбинирани речници
<b>6,90</b>	<b>5,55</b>	<b>6,05</b>

**Таблица 2.** Среден брой думи на страница, при които малкото тире не е разпознато като знак за сричкопренасяне.

Що се отнася до разликата в разпознаването между цветните страници и тези в сивата скала (greyscale), успехът при разпознаването на вторите беше много подобрен, което не оправдава приоритетно сканиране в режима на скала на сивото или ненужно конвертиране на файлове.

Отчитаният от ABBYY FineReader параметър „Low-confidence characters“, въпреки че има съществено разминаване с процентът ръчно преброени погрешно разпознати думи, също демонстрира тенденцията за леко подобрение при използване на CLADABG-MODEL. Това означава, че може да бъде използван при

нужда от обща преценка, като може да замести трудоемкото ръчно броене на грешните думи при бъдещи тестове.

### ***Заключение и бъдещи планове***

Предимствата на CLADABG-MODEL са доказани и използването му е препоръчително. Работата по морфологичния речник ще продължи, за да се оптимизира процесът като цяло, както и неговата ефективност по отношение на постигането на по-висок успех при разпознаване. Две са основните причини за сегашните скромни резултати. В периода от средата на XIX век до 1945 г. са съществували или са били въвеждани различни правописни системи, докато „старият“ речник представя само една от тях, макар и широко приета. Това ограничение може да бъде преодоляно чрез включването на словоформи, които да отразяват различни правописни модели и тяхната кодификация в едноезични речници, граматика и други документи от различни периоди. В тази връзка започна сканирането на правописни речници от по-стари периоди, с цел обогатяването на CLADABG-MODEL. Друго решение би могло да бъде обогатяването на „стария“ речник така, че да обхваща няколко правописни варианта за всяка словоформа, подобно на многоезичен речник. Вторият фактор, който има негативен ефект върху резултатите, е липсата на лични и географски имена, и названия на организации или продукти. Планираме да разрешим този проблем като добавим лексикален материал, извлечен от ръчно коригирани OCR текстове. Освен обучение на софтуера за OCR предвиждаме да внедрим проверка на правописа чрез невронна мрежа за стари текстове. Моделът ще разчита на по-широк контекст, за да предскаже грешно разпознатите думи. За обучението на модели планираме автоматично да създадем паралелен корпус със стари и съвременни текстове, използвайки „старите“ речници и предварително обучени модели.

Екипите от НБИВ и ИИКТ-БАН, работещи по оптичното разпознаване на символи, с готовност ще споделят опита си с други институции, които се сблъскват с такава проблематика. Идеята на проекта КЛАДА-БГ е именно такава – да се осигури публичен достъп до създаваните по него ресурси и инструменти. Въпре-

ки че разработването му е все още в начален стадий, речникът CLADABG-MODEL може да бъде предоставен за провеждане на тестове. НБИВ е в готовност да сподели опита си и използваните методи за подобряване на OCR процеса, както и начините за повишаване на неговата ефективност.

### **Библиография:**

1. Falt, E., Das, P. P. Digital libraries can ensure continuity as Covid-19 puts brake to academic activity. [онлайн] 08.04.2020 [прегледан на 29 октомври 2020]. Достъпен на <https://en.unesco.org/news/digital-libraries-can-ensure-continuity-covid-19-puts-brake-academic-activity>
2. АBBY Technology Portal: Dictionaries and OCR. [онлайн] б.д. [прегледан на 02 ноември 2020]. Достъпен на [https://abbyy.technology/en:features:ocr:dictionary\\_support](https://abbyy.technology/en:features:ocr:dictionary_support)