Useful in Times of Crisis: Improved Access to Online Content

Dimitar Minev, Ivan Kratchanov

National Library "Ivan Vazov", 17 Avksentii Veleshki St., 4000 Plovdiv, Bulgaria
nbiv@libplovdiv.com, digitization@libplovdiv.com

Abstract. This paper gives an overview of the current efforts of National Library "Ivan Vazov" in the field of digitization. It focusses on the technical aspects of a digitalization project, involving OCR and provides insight into the library's joint efforts with its partners to develop relevant tools and methodologies.

Keywords: Digitization, Plovdiv, Cultural Heritage, Digital Library.

1 Introduction

In 2019 new functionalities were introduced in the online digital library[1] of the National Library "Ivan Vazov" (NLIV) - indexing and searching the contents of PDF files and related augmentations. By enabling optical character recognition (OCR), it is possible to take advantage of the possibilities offered by machine-readable text, thus making it easier for users to find and use the content and renders it more accessible to learners and scholars.

The digital library's enhanced functionalities served as an appropriate response to the COVID-19 crisis, offering free content with improved access, so that patrons may continue to read and learn without disruption. All of the usage statistics showed a significant increase during the State of Emergency in Bulgaria, when the library's building was closed to visitors and the tendency for increased usage of digital resources has remained. An indispensable feature of the digital library was its entirely online-based administration, which allowed the personnel of the Digitization Center to shift efficiently to remote work from home.

To continue improving, it is important to work closely and exchange ideas with the software developer of the digital library so that our intentions could be fully realized and adapted to the limitations of the Microsoft SharePoint software platform, which functions as the library's foundation.

A key strategic objective to the success of such projects is the sharing of digitization experience with other partners, working in the same field, with the aim to form a comprehensive strategy and collaborative solutions. The library recognized the importance

[1] The digital library is available online at the following web address: http://digital.plovdiv.bg/BG/Pages/LibIvanVazov.aspx

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 10, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2020. ISSN: 1314-4006, eISSN: 2535-0366

of cross-institutional, multidisciplinary cooperation, and an example of such involvement is the library's participation in the project CLaDA-BG.

2 Augmentation of the Digital Library's Interface and PDF Files' Migration

The updated capabilities of the software platform required changes in the user interface in order to accommodate the newly introduced functionalities. New search fields were added and content search is applied to both global and collection search. Searching is done in the indexed contents of the PDF files and the provided metadata, and is supported by the Boolean operators provided for by the portal's software platform SharePoint 2010. A new gallery tool for viewing and navigating PDF files was implemented, providing tools for searching and navigation.

The contract with the software developer included a one-time migration service for all collections, in order to replace the existing images in the platform with the corresponding PDF files. The migration was executed by a software tool, developed specifically for that purpose, which uploads and publishes PDF files to the respective destination, while simultaneously removing the existing image files. The PDF files had to be named in a specific way, so that the tool may establish their desired destination. The name included the title's system ID, and in the case of periodical publications, the year of publication and issue number as well. The purpose of the migration is to minimize manual labour and to decrease the time needed for the replacement. In mid-April 2020, the migration was completed and all text resources available in the digital library contain recognized, machine-readable text.

3 Specifics of OCR and PDF File Processing

ABBYY FineReader 14 was chosen as the most appropriate software for accomplishing the OCR task. Our surveys showed that it was the most widely used program to perform OCR and in our tests performed better at recognizing Cyrillic text, with a higher degree of recognition, than the other tested products.

Employing FineReader on master files obtained from well-preserved originals, written in modern Bulgarian language, yields very high OCR success rate, most often above 99%. However, the majority of the digitized texts, those of high cultural and historical significance and with expired copyright, are predominantly from the period before the Orthographic Reform of 1945. The archaic text and font, and the deteriorated condition of the originals are some of the factors that inevitably lead to a decrease in the accuracy of OCR and finding ways to achieve higher degree of recognition becomes essential.

Implementing the newest trends of datafication of library collections (Mahey & Al-Abdulla, 2019), the primary master files, stored on the library's servers and created for the purposes of long-term digital storage, will be used for OCR.

The master files are 24-bit color images, scanned at 300 ppi or higher, from the original paper source. (Fig. 1.) In this way, the degree of recognition will be approaching its maximum.

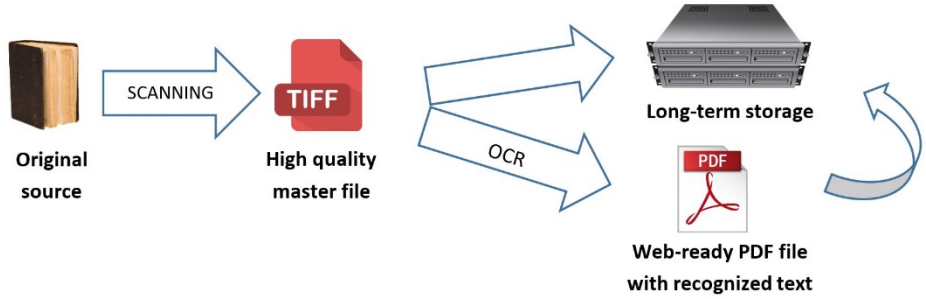

Fig. 1. Generation of files in the process of digitization

The resultant PDF file will have appropriately lowered image quality, with a size suitable for online display. Its images will be aligned with a hidden layer of machine-readable text, product of the recognition. Mixed raster content (MRC) compression, a method for compressing compound images, containing both binary text and continuous-tone images (de Queiroz, Buckley, & Xu, 1998), will not be used. Although MRC compression produces PDF files with much smaller size, its radical alteration of the original appearance was not deemed appropriate for the purpose of presenting cultural heritage objects.

After the software performs OCR, there are a number of ways to improve the resulting recognized text, such as performing manual corrections. This is a labor-intensive, time-consuming task, ultimately inefficient for large-scale digitization projects.

Another method, first implemented by the National Library of Australia in their newspaper digital collection, is to design the digital library in such a way as to allow the users of the online resources to correct OCR mistakes. It is considered a successful practice and has been incorporated by other institutions[2]. NLIV decided not to pursue this option, because it requires a massive, and therefore very expensive, overhaul of the online platform and because of the uncertainty concerning the feasibility of the outcome and the extent of the potential results, especially considering the possibility of intentionally wrongful acts.

4 CLaDA-BG Project

NLIV currently participates in the CLaDA-BG project, which is integrated within the European CLARIN and DARIAH infrastructures. The mission of CLaDA-BG is to establish a national technological infrastructure of language, cultural and historic heritage

[2] An example of this practice is the Estonian newspaper archive DIGAR, available at: https://dea.digar.ee/cgi-bin/dea?l=en

(CHH) resources and technologies which to provide public access to language and CHH resources, tools for Bulgarian language processing and tools for access and management of CHH datasets for various societal tasks, targeted at wide audience (CLaDA-BG, n.d.). The participants are grouped into two distinct categories: content providers and technological partners. The library's efforts, as a content provider, are focused in two main directions:

1. To develop the best methodology for optical character recognition and consequently to enhance the methods of searching in the text. The library's role would be in providing texts from its comprehensive holdings, from different periods, with different fonts, formats, etc., and also to test and apply the developed resources, such as thesauri, search-engine-complementing instruments and others. An especially important aspect is that of the recognition and access to printed text before the Orthographic Reform (1945). The goal is to use the tools developed by the technological partners in CLaDA-BG to minimize and correct errors in the machine-readable text, acquired by OCR, and also to allow normalization of the text (to convert it into modern spelling) in order to aid the user, so that he/she would not have to search for a word or expression twice, in the new and old spelling. The retrieved search results would include simultaneously both.

2. To increase Bulgarian content in Europeana[3] by developing software tools to improve metadata submission. Currently, the ability of the software platform to export XML files in EDM (Europeana Data Model) format is problematic and limited. The XML files do not fully meet the requirements and cannot be sent directly. We are keen to have a stronger presence in Europeana, but the great amount of manual work currently needed disrupts our potential to contribute in a meaningful and visible way. The CLaDA-BG technological partner Ontotext is currently working on the correction of the XML EDM files, but also on the enrichment of the provided metadata with associated open data and on its linking into knowledge graphs. We can provide our online resources for the development of the tools, as well as participate in tests. Once the work is complete, the library will be able to submit to Europeana its entire collection of digitized cultural heritage, currently amounting to a total of more than 33,000 objects.

The library's involvement in the CLaDA-BG project is important, because it will allow wide dissemination of the results of the project, providing a direct link to the information users. We see our role in providing texts from the library's rich holdings, from different periods, with different fonts, formats, etc. We will share our experience and we will test and apply the resources developed by the project partners.

The first major outcome of our work on the project, carried out in collaboration with the Institute of Information and Communication Technology at the Bulgarian Academy of Sciences (IICT-BAS), was the preparation and testing of a dictionary of old Bulgarian spelling word forms, to be used for the purpose of assisting OCR. The dictionary was named CLADABG-MODEL by its creators at IICT-BAS. Testing was conducted in the period March-April 2020. The program ABBYY FineReader (ver. 14 and 15) was used to carry out recognition of 20 pages from issue 1/1881 of the magazine "Наука" ("Science") from the holdings of NLIV, with call number П РЦ-9. The pages

[3] https://www.europeana.eu/portal/en

are color scanned with an i2S CopyBook A2 scanner at a resolution of 300 ppi, 24-bit, TIFF format, no compression.

The purpose of the test was to determine to what extent the dictionary with old word forms assists the software program in performing OCR of printed texts in Bulgarian language before the Orthographic Reform of 1945. The dictionaries used by FineReader are a list of words available in a specific language. The program relies on dictionaries to increase the quality of recognition by reinforcing hypotheses about words found in a dictionary. Custom dictionaries may help in cases when text contains many non-common words (ABBYY, n.d.). The program has a built-in dictionary only for the modern Bulgarian language. CLADABG-MODEL contains 1,121,872 words in use before the Orthographic Reform of 1945, including words no longer used and containing symbols gradually removed from the modern written language, such as ѣ, ѫ, ѧ and others. Much of the digitized valuable library possessions have text that is pre-1945 and the purpose of developing CLADABG-MODEL was to test the hypothesis that its use will lead to a higher recognition rate. The amount of the increase, if any, also had to be determined. The main indicator used was the percentage of misrecognized words in relation to the total number of words. The counting was done manually.

In the course of the test, two other characteristic features of the OCR process and of the software program were measured: the degree of recognition of images in grayscale (as opposed to those in color) and whether and how the FineReader parameter "Low-confidence characters" (expressed in percentage) can serve as an indicator of the success of OCR.

The original paper version of the journal "Наука" is very well preserved, respectively, the resulting scanned files are close to the optimal characteristics recommended for OCR. However, darkening of the paper is observed, which reduces the contrast and distinctiveness of the letters. Also, the chosen font (widely used back then) makes it difficult for the program to distinguish letters with dominant vertical lines (such as И, П, Н, Ш, Л) - Fig.2. The horizontal lines converge, the letters fuse together and further complicate the work for the recognition algorithm.

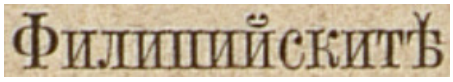

Fig. 2. Example of a word with merged letter symbols

To test the degree of recognition of CLADABG-MODEL, 20 identical pages were scanned, with uniform text and font. The total number of words is 5485 and their average number per page is 274.25.

Minimal training was done, to aid the recognition of traditionally problematic symbols such as "ѫ", which without prior training always becomes a "ж".

A testing was included also for the simultaneous, combined use of two dictionaries - the FineReader built-in Bulgarian dictionary and CLADABG-MODEL, with recognition performed using two base languages: 1) Bulgarian with a standard set of characters with the built-in dictionary and 2) language set based on the Bulgarian language,

with added old letters, such as ѣ, ѫ, ѧ , etc., and with the CLADABG-MODEL dictionary. The inclusion of the combined dictionary test was done due to the consideration that when the program works only with CLADA-MODEL, there is a risk of greater recognition failure in words still in use in modern Bulgarian.

The results are as follows:

Table 1. Summary – mean percentage of misrecognized words[4] for 20 color scanned pages, 300 ppi, 24-bit, TIFF format, no compression.

Percentage of misrecognized words (FineReader built-in dictionary)	Percentage of misrecognized words (CLADA-MODEL)	Percentage of misrecognized words (Combined)
4,90	4,40	4,50

The results show that the recognition with CLADA-MODEL is improved. However, the improvement is not so significant - on average with 0.5% fewer misrecognized words. Surprisingly, the use of combined dictionaries did not lead to higher recognition success.

The second test concerned the parameter "Low-confidence characters" - a percentage number, calculated and reported by FineReader for each recognized page.

Table 2. Summary – Mean percentage of parameter “Low-confidence characters” for 20 color scanned pages, 300 ppi, 24-bit, TIFF format, no compression.

Parameter “Low-confidence characters” (FineReader built-in dictionary)	Parameter “Low-confidence characters” (CLADA-MODEL)	Parameter “Low-confidence characters” (Combined)
3,85	2,70	3,00

Despite there being a significant numerical discrepancy between the parameter "Low-confidence characters" and the percentage of misrecognized words (counted manually), the trend of slight improvement using CLADA-MODEL is being demonstrated as well. This means the parameter can be used in need of an overall estimation, and may be employed in future tests as a replacement of the time-consuming task of manual word counting.

The "Low-confidence characters" parameter was then used to study the recognition difference between color and grayscale pages. The recognition success of the greyscale pages was only slightly better, which does not justify prioritizing the greyscale scanning mode or unnecessary file conversion.

[4] Misrecognized are the words in which there is a discrepancy of the letter symbols between the scanned primary word in image form and the recognized, derivative machine-readable word. It is not considered incorrect recognition if the primary word is spelled incorrectly and the derived word has correctly recognized letter characters, thus duplicating the spelling error.

Overall, the benefits of CLADABG-MODEL have been proven and its use is highly recommended. The work on the dictionary will continue it order to streamline and improve it, and to achieve higher recognition success.

5 Conclusion

This paper gave an overview of the digitization efforts in the National Library "Ivan Vazov", highlighting how the library transforms its digital collection answering the current demand for datafied collections, which improve the accessibility and use of the digital resources.

Advancements in this area are especially important in the current times, marked by the COVID-19 pandemic. Indeed, as the demand for credible e-resources surges, digital libraries have emerged as vital pathways to high-quality e-books, journals and educational content. Statistics from the world's leading e-libraries testify to their cultural significance (Falt & Das, 2020).

An important part of the future digitization work at National Library "Ivan Vazov" will be the mastery and long-term establishment of OCR-related activities as the first step towards datafication of the digital collection, aiming to ensure the highest possible level of resource usability. Cooperation with CLaDA-BG technological partner Ontotext is ongoing as well, with the aim to enhance the digital library's algorithm for creating EDM XML files, which also includes the enrichment the provided metadata with associated open data and for linking into knowledge graphs.

It is important to share digitization experience with partners, often in the form of multidisciplinary associations, with the aim to form a comprehensive strategy and collaborative solutions.

References

ABBYY. (n.d.). *Dictionaries and OCR*. Retrieved 6 17, 2020, from ABBYY Technology Portal: https://abbyy.technology/en:features:ocr:dictionary_support

CLaDA-BG. (n.d.). *Mission*. Retrieved 6 17, 2020, from CLaDA-BG: https://clada-bg.eu/en/mission/

de Queiroz, R., Buckley, R., & Xu, M. (1998, 12 28). *Mixed Raster Content (MRC) Model for Compound Image Compression.* doi:https://doi.org/10.1117/12.334618

Falt, E., & Das, P. P. (2020, 4 8). *Digital libraries can ensure continuity as Covid-19 puts brake to academic activity*. Retrieved 6 17, 2020, from UNESCO: https://en.unesco.org/news/digital-libraries-can-ensure-continuity-covid-19-puts-brake-academic-activity

Mahey, M., & Al-Abdulla, A. (2019, 9). *Quatar University - QSpace Institutional Repository.* doi:http://hdl.handle.net/10576/12115

Received: June 16, 2020
Reviewed: June 30, 2020
Finally Accepted: July 15, 2020