Sustainable development of the practices of digitization in National Library "Ivan Vazov" - Plovdiv

Ivan Kratchanov

Sustainable development of the practices of digitization in National Library “Ivan Vazov” – Plovdiv

Ivan Kratchanov

Introduction: National Library Ivan Vazov – history, holdings and traditions

1 National Library Ivan Vazov in Plovdiv is the second national repository of Bulgarian textual heritage. The library has played an essential role in the preservation of Bulgarian culture and history. It is the first cultural institution in Southern Bulgaria, established as a Regional Library and Museum of Eastern Rumelia in 1879.

2 After the Unification of Bulgaria and Eastern Rumelia, the established legislature regulated equal rights and obligations for the two national libraries - in Sofia and Plovdiv. That is why the National Library in Plovdiv has developed as an archive of Bulgarian books and periodicals, a historical archive, a rich repository of manuscripts and Revival literature, unique collections of rare and valuable publications.

3 Today, National Library Ivan Vazov in Plovdiv is a dynamic cultural institute that continues to enrich and develop the traditions of its prominent founders. The library enjoys over 120,000 visits annually loaning some 300,000 library documents. The library’s holdings are comprehensive and amount to over 1,900,000 library units – scientific, fiction, manuscripts, old-printed, rare and valuable publications, Bulgarian and foreign periodicals, photographs, maps, audiovisual and electronic documents, original works of art, personal libraries.

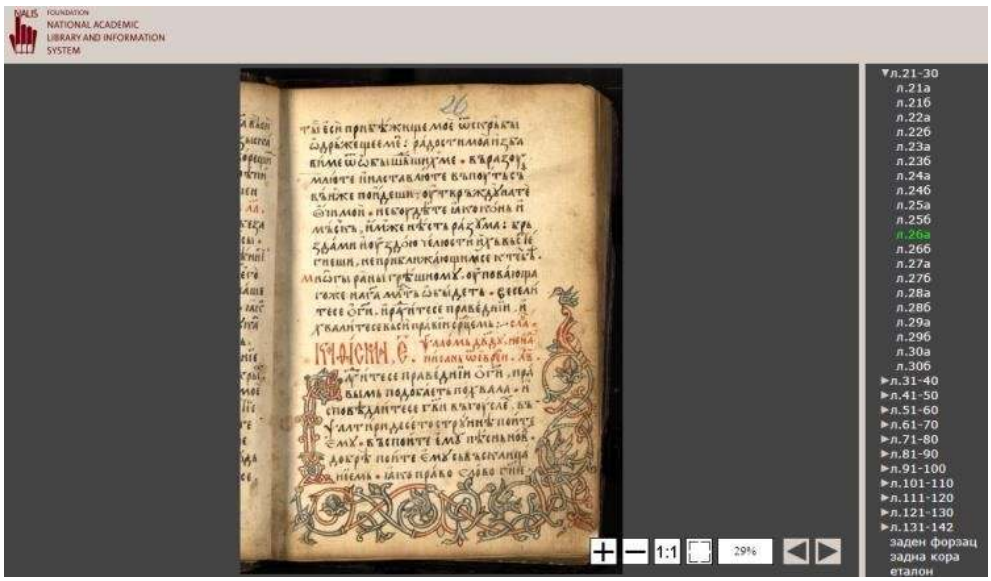
Traditions in the field of library automation and the digital display of holdings

Stages in the technological development

4 The library traditionally follows and implements the latest trends in automation. In 1979 the library's first computer system IZOT-0310 was introduced, together with the information search devices UPDML 9002-02 and IPU IZOT-0320. The first subscription to the AGRIS (FAO) database, stored on magnetic tapes, was done in 1980. The local library-information network was established in 1994.

5 The Digitization Centre was founded in 2008. Since then the library has participated in various projects, on a national and international level, concerning the digitization and online display of its valuable holdings. Some noteworthy projects are *Europeana Photography*[1], EMBARK[2], BG08 "Digital Cultural and Historical Heritage of Plovdiv Municipality" and others. Digital copies of 95 Slavonic manuscripts from the 12th – 18th centuries are currently accessible at the Manuscript collection of NALIS Repository[3], a prestigious national academic digital library, founded by the Central Library of the Bulgarian Academy of Science, Sofia University "St. Kliment Ohridski" and the American University in Bulgaria.

Figure 1. An example of a folio from Markovski psalter, 1638 (No. 5(207)) displayed in the interface of NALIS Repository



(This image is public domain, from the collection of the National Library Ivan Vazov)

Digital library: current content and potential for growth

6 In 2017 the Digital Library became available online at http://digital.plovdiv.bg/BG/Pages/LibIvanVazov.aspx. It is a part of a web portal, which unites seven of the most significant cultural institutions in Plovdiv. The library offers nine collections:

- BOOKS: providing access to 194 items in the time of writing, predominantly Statutes of professional organizations from Plovdiv and the region, from the end of the 19th century and

the beginning of the 20th century. They are of interest to researchers because they provide insight into the way of life of the era. The plans are to continue uploading the most valuable and compelling holdings of the library.

- PERIODICAL PUBLICATIONS: offering access to newspapers and magazines currently containing 223 full-text titles, with approximately 20,000 separate issues. Particularly valuable are the newspapers and magazines from the East-Rumelian period: "Maritsa" newspaper – the first Bulgarian newspaper after the Liberation, "Narodniy glas", "Nauka", "Zora", etc., as well as the Revival-period collection of periodical publications.
- MANUSCRIPTS: this rich collection includes Slavonic, Greek, Ottoman and Persian manuscripts on parchment and paper created between the 11th and 19th centuries. This type of content is a priority for being ingested and made accessible within the NALIS Repository. The digital library currently offers access to 16 manuscripts, with more to be uploaded in the near future. An example is presented on Figure 1.
- GRAPHIC PUBLICATIONS is a collection of art prints, lithographs, etchings, engravings, posters, original paintings, etc. Especially interesting are the projects for monumental works – a total of 95 projects for large-scale wall murals, frescoes, ceramic tilework, most of them realized in many Bulgarian towns. The artworks were created by renowned Bulgarian artists such as Dimitar Kirov, Yoan Leviev, Encho Pironkov (see for an example Figure 2).

Figure 2. Yoan Leviev, Anna Grebenarova. Design for a ceramic piece, 1966

(This image is public domain, from the collection of the National Library Ivan Vazov)

- CARTOGRAPHIC PUBLICATIONS presents valuable possessions of the library such as the oldest map of Bulgaria, created by a Bulgarian and printed in the Bulgarian language – "Map of the present Bulgaria, Thrace, Macedonia and the adjacent lands".
- The PHOTOGRAPHS digital collection currently hosts 163 photographs (see for an example Figure 3) and will be expanded over time. The library has a physical collection of

approximately 4,000 photographs and postcards, portraits, events and sites of historical significance.

Figure 3. Dimitri Ermakov. Plovdiv, Sahat Tepe, 1876

(This image is public domain, from the collection of the National Library Ivan Vazov)

- The ARCHIVES DIGITAL COLLECTION also has a substantial potential for development. The Bulgarian historical archives in the National Library Ivan Vazov are of national value, casting light on key moments of the political, economic and cultural development of Bulgaria. The documents cover the period from the 12th to the 20th centuries, the most numerous being from the second half of the 19th century.
- In the future, we will also develop our AUDIOVISUAL collection. The software platform of our digital library allows to make accessible audio and video files. The library has a rich collection of classic films on 35mm reels – masterpieces from the invention of the Cinématographe in 1895 to the early 40's of the 20th century, including classic films by Lumiere brothers, David Griffith, Charles Chaplin, Buster Keaton, Fritz Lang, John Ford, Dziga Vertov, Sergei Eisenstein, Luis Bunuel, Orson Wells and others. The library also has an extensive collection of recorded music with approximately 16,000 vinyl records.
- Last but not least, the library also has a rich collection of SHEET MUSIC PUBLICATIONS. Of particular interest is the manuscript "Bulgarska kitka", which is already accessible from the Digital library. It was created in 1881 by the Czech composer Franz Schwestka to celebrate the newly established brass musical ensemble of Plovdiv.

7 In 2019 a new functionality was introduced in the Digital Library of National Library Ivan Vazov – indexing and searching in PDF files' text contents. The machine-encoded text is obtained through the use of optical character recognition (OCR) software. By including the OCR step, full-text indexing and Internal document search capability can be applied, making it easier for users to discover and use the materials. Digital-born PDF files do not need to be processed.

8 The main activities within the scope of the new functionality are as follows:

- Development of the software platform to upload and display PDF files with the possibility to search the contents of the file (if OCR had been implemented).
- One-time migration service for all existing collections in the Digital Library in order to replace the existing images in the platform with corresponding PDFs.
- Purchasing of ABBYY FineReader 14 software product for OCR and processing of PDF files.

9 It was important to work closely and exchange ideas with the software developer, so that our intentions could be realized as fully as possible and to adapt to the limitations of Microsoft's software platform SharePoint, on the basis of which our Digital Library was created.

Developments of the Digital Library aimed at content-based retrieval

10 The updated capabilities of the software platform required changes in the user interface in order to accommodate the new 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 augmented by the boolean operators supported by SharePoint 2010.

11 A help box is a useful addition to the search fields; it displays static text with short search instructions, as well as a link, leading to an external page with comprehensive guidelines.

12 Especially significant are the changes in the collection "Periodical publications", because of the necessity to display a multitude of search results from many issues. So far, this was not necessary because the search was done only with respect to the metadata of the title.

13 A new gallery tool for viewing and navigating PDF files was implemented. It offers the following capabilities:

- Search by keyword in the contents of the file, mark the matches found with distinctive colour, display the number of matches and provide controls to move to matches.
- Suitable navigation of the pages of the document, with instruments such as thumbnail view of individual pages and a blank field where the desired page number may be entered.
- Fullscreen view.
- Help menu with additional tools – navigate to the first or last page of the file, change the orientation of the page and hand-tool.

14 The administrative part of the software platform did not require a significant overhaul and the most important addition was a section for PDF file upload in the metadata entry form.

15 The contract with the software developer also includes a one-time migration service of PDF files for all collections in order to replace the existing images in the platform. The purpose is to decrease the time needed for the replacement. The personnel of the Digitization Center are currently working on the preparation of the complete batch of files for the replacement. The name of each PDF file must include the unique ID number of the corresponding record in the Digital Library system, which it will replace.

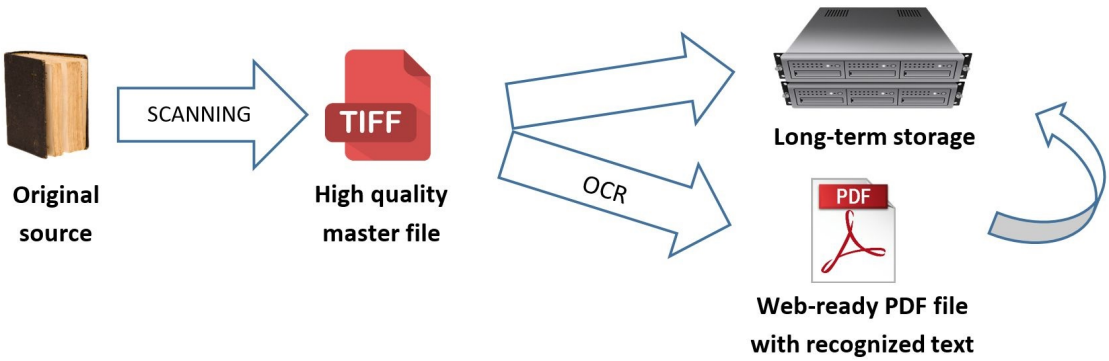
Particulars of OCR and PDF file processing

16 Implementing the newest trends of datafication of library collections (Mahey et al. 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 (see Fig.4.).

17 Predominantly, the master files are 24-bit colour images, scanned at minimum 300 dpi from the original paper source. Scanning from the originals is generally acknowledged to produce higher quality master images (Klijn, Edwin, 2008). In this way, the degree of recognition will be approaching its maximum.

18 The master files will be processed to a single PDF file for each unit of cultural heritage. For example, a single PDF file should be a monograph or a newspaper issue. The resultant file will have appropriately lowered image quality, with a size suitable for online display. The PDF will have its images aligned with hidden machine-readable text, the product of the recognition. MRC compression method will not be used.

Figure 4. Generation of files in the process of digitization



(This image is public domain)

19 After a thorough review of the available software, which included sampling the experience and opinions of libraries from Bulgaria, Russia, Ukraine, Serbia and other countries, where the recognition of Cyrillic text is of special relevance, we decided to purchase a licensed version of ABBYY FineReader 14, which was the most widely used software to perform OCR and in our tests was the best at recognizing Cyrillic text. Other options were considered and tested as well (such as Adobe Acrobat XI Pro) but the results achieved were not as satisfactory, because the degree of recognition was lower.

20 Using ABBYY FineReader 14 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 texts that we have to deal with, those of high cultural and historical value and with expired copyright, are predominantly from the period before the Orthographic Reform of 1945. The accuracy of OCR software is language dependent: alphabet; old letters without the coding tables; old grammar, obsolete words, phrases and idioms; dictionaries; multi-lingual documents (Andreev,

Andrey and Kirov, Nikolay, 2009). At the same time, despite the care that has been taken by the library to preserve the originals, old textual documents present challenges to OCR, such as the natural darkening of the paper, faded print, in-library binding in proximity to the text, etc.

21 There are a number of ways to improve the accuracy of the recognized texts. A straightforward solution is to include post-OCR manual corrections as another stage to the digitization process. However, with the human efforts needed to correct OCR errors, it becomes quite a tedious job (Andreev, Andrey and Kirov, Nikolay, 2009). Considering the large amount of time necessary, to achieve high levels of accuracy (around 98%), the labour-intensive cleaning required to remove OCR errors means the two-step process may be no more efficient than manually inputting texts from scratch, a procedure that suits small- to medium-scale projects (Strange, C., McNamara, D., Wodak, J., Wood, I., 2014). A way to mitigate this issue, while still using manual labour, is to design the digital library software platform in such a way that it involves the users of the resources and allows them to correct OCR mistakes. This solution was first implemented by the National Library of Australia in their newspaper digital collection. It is considered a successful practice (Holley, Rose, 2009) and has been incorporated by other institutions[4]. The National Library Ivan Vazov 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.

22 Another method of improving the quality of OCR, relevant to the software ABBYY FineReader, is the option to train the program to patterns, the interpretation of which the software deems uncertain. This is useful in cases of non-standard fonts and is especially important for Cyrillic texts, where the training to recognize specific letter symbols is essential. Such are the letters ѣ, Ѧ, Ѣ, Ѧ, Ѩ, etc., which were gradually removed from the modern written language, eventually reducing the number of letters in the alphabet to the current 30.

23 When reviewing the practices of libraries in Bulgaria, which offer PDF files with recognized text in their digital libraries, it is evident that there is no uniform standard for the ways in which OCR of documents created before the Orthographic Reform of 1945 is performed. For instance, some of the reviewed libraries decided to replace the archaic letter symbols with their modern equivalents, which is done in order to aid the search of the contents of older texts, so that users would not have to write the required search expression twice – in the old and new spelling. However, this cannot be considered a good practice, because the spelling conversion is not fixed in the sense that old letter symbols are often replaced by more than one modern letter symbol. For instance, the letter "ѣ" is replaced by modern "е" or "я" and "Ѧ" is replaced by modern "ъ" or "а".

24 National Library "Ivan Vazov" currently participates in the CLaDa-BG[5] project, which is integrated within the European CLARIN[6] and DARIAH[7] infrastructures. The mission of CLaDA-BG is to establish a national technological infrastructure of language, cultural and historic heritage (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 Mission, n.d.). The participants are grouped in 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. 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 by developing software tools to improve metadata submission. Currently, the ability of the software platform to export XML files in EDM (Europeana Data Model) is problematic and limited. The XML files do not fully meet the requirements and cannot be sent directly. The CLaDA-BG technological partner Ontotext will create a tool for correction of the XML EDM files, but also for enriching the provided metadata with associated open data and for linking into knowledge graphs.

25 The library's involvement in the CLaDA-BG is essential because it will allow wide dissemination of the results of the project, providing a direct link to the information users.

Conclusion

26 In this paper we made 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. The new file format will lead to changes in the work cycle of the Digitization Centre, which in terms of web presentation focused exclusively on the .jpg image format. An essential part of the future digitization work 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. Therefore, it is important to share digitization experience with other partners, working in the same field, with the aim to form a comprehensive strategy and collaborative solutions.

27 Involvement in developing tools to automate the submission of data to Europeana[8] is of top priority as well. Stronger presence there is mandatory, but the great amount of manual work currently needed to prepare the collection for ingestion to Europeana delays our potential to contribute in a meaningful and visible way.

BIBLIOGRAPHY

Andreev, Andrey, and Kirov, Nikolay. "Hausdorff Distances for Searching in Binary Text Images." *Serdica Journal of Computing* 3.1 (2009): 23-46, eudml.org/doc/11442.

CLaDA-BG Mission, clada-bg.eu/mission/

Holley, Rose. *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*, 2009, www.nla.gov.au/content/many-hands-make-light-work-public-collaborative-ocr-text-correction-in-australian-historic.

Klijn, Edwin. "The current state-of-art in newspaper digitization : a market perspective." *D-Lib Magazine*, no. 14(1/2), 2008, www.dlib.org/dlib/january08/klijn/01klijn.html

Mahey, Mahendra et al. *Open a GLAM Lab*. QUPress, 2019, qspace.qu.edu.qa/handle/10576/12115

Strange, Carolyn et al. "Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers. " *Digital Humanities Quarterly*, no. 8(1), 2014, www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html.

NOTES

1. https://pro.europeana.eu/project/europeanaphotography
2. https://www.libplovdiv.com/index.php/en/enhance-manuscriptorium-through-balkan-recovered-knowledge-embark-en

3. http://digilib.nalis.bg/xmlui/
4. An example of this practice is the Estonian newspaper archive DIGAR, available at: https://dea.digar.ee/cgi-bin/dea?l=en
5. National Interdisciplinary Research E-Infrastructure for Bulgarian Language and Cultural Heritage Resources and Technologies. https://clada-bg.eu/
6. CLARIN: European Research Infrastructure for Language Resources and Technology. https://www.clarin.eu/.

7. DARIAH: Digital Research Infrastructure for the Arts and Humanities. https://www.dariah.eu/.
8. https://www.europeana.eu/portal/en

ABSTRACTS

The National Library Ivan Vazov in Plovdiv is the second largest library in Bulgaria. It serves as the second national legal depository of Bulgarian printed works. In addition, it has contributed significantly to the preservation and the digital accessibility of the national cultural and historical heritage. This article offers an overview of the library's history and current developments in the field of automation and digitization.

INDEX

Keywords: digitization, Plovdiv, cultural heritage, digital library

AUTHOR

IVAN KRATCHANOV

National Library Ivan Vazov, 17 Avksentii Veleshki Str. Plovdiv 4000, Bulgaria
ivankra@gmail.com