Преглед НЦД 30 (2017), 7–17

Nikola Smolenski, Nemanja Zikic, Matea Milosevic, Milena Kostic, dr Vasilije Milnovic, dr Adam Sofronijevic
University Library "Svetozar Markovic" Belgrade, Serbia

NEW HORIZON OF DIGITIZATION IN SERBIA IMPROVEMENT OF DIGITIZATION THROUGH COOPERATION WITH WORLD LEADING INSTITUTIONS AND THE IN-HOUSE DEVELOPMENT OF DIGITAL TOOLS

Abstract. In 2016 the University Library "Svetozar Markovic" in cooperation with eight partner institutions started a project entitled "New Horizon of Digitization in Serbia" which was funded by the Ministry of Culture and Information of the Republic of Serbia. The aim of the project was to improve the main segments of digitization which were developed properly and to apply new digitization methods in as many institutions in Serbia in a standardized way. New digitization methods and approaches used in this project mostly come from the cooperation of the University Library "Svetozar Markovic" with the world leading institutions such as the British National Library or result from the in-house development of digital tools which improve segments of digitization process. During 2016 the University Library "Svetozar Markovic" in cooperation with the British National Library carried out the project entitled "Safeguarding the fragile collection of the private archive of the Lazic family" within which it acquired rich experience in enhancing digitization process. This paper presents several segments which show how the digitization process is improved within the project. Possibilities and practice of creation of master files of digital copies of important documents which contain colour and size references which began at the University Library "Svetozar Markovic" are described in the paper. Importance and selectivity of the application of this method and its consequences for digitization practice are also portrayed especially with regard to processes of long-term storing of digital copies. The need for the application of publicly available mechanisms for authentification of digital materials is defined and types of digital materials to which this method should be applied are considered as well as possibilities for the application of the model in practice. The paper outlines the need for the full searchability of digital text in materials which are presented to patrons and which portray the University Library "Svetozar Markovic" experience on the example of a searchable collection of historical newspapers. Two methods for creating searchable texts are outlined: via the creation of METS-ALTO files in case of printed materials and via automatic recognition of the handwritten text. The paper defines steps in the creation, long-term preservation and presentation of METS-ALTO files based on the University Library "Svetozar Markovic" experience and describes digital tools necessary for carrying out these steps, docWorks programme, a professional repository for long-term preservation and display of digital materials Therefore and a system for searching and display of searchable texts developed by the University Library "Svetozar Markovic" in cooperation with the National Library of Luxembourg. The scope and possibilities for automatic recognition of the handwritten text based on the present experience coming from the participation of the University Library "Svetozar Markovic" in Horizon2020 READ are described and the digital tool Transcribus which is applied in the process. Finally, the needs of patron groups in working with digital materials and possibilities for their display and promotion in the modern business environment of libraries are analyzed. The paper describes in detail the experience of using the promotional device Magic Box at the University Library "Svetozar Markovic" and outlines analysis and scenarios of its application in cultural institutions. This paper is an overview of the possibilities and first experiences of improving digitization process which will mark the following development of the field in Serbia and the region.

Keywords. digitization process, digital tools, University Library "Svetozar Markovic"

I Introduction

The University Library in Belgrade (ULB) as the central academic library in Serbia – which has gained invaluable experience in versatile EU funded digitization projects, the

most notable being “Europeana Libraries” and “Europeana Newspapers” – coordinated the digitization project “Safeguarding the fragile collection of the private collection of the Lazić family” in the framework of Endangered Archives Programme (EAP) of the British Library. With these endeavours the ULB has positioned itself as the main innovative content provider for researchers in Serbia and one of the leading institutions in digitization.

It is anticipated that the information needs of researchers relating to research will be met through the information available via digital technology (Newton, 2000). So, the ability to access and use digital technologies is becoming a critical aspect of the contemporary science. Academic libraries are the places where researchers discover new information and where information literacy skills of researchers are being developed (Julien et al, 2017). The role of such a library is to become an information hub where one can experiment with new technologies and where new concepts can be adopted (Jerkov et al, 2015). This is especially important for humanities and social sciences researchers in Serbia who represent a deprived group which often struggles to acquire up-to-date information and knowledge.

The University Library “Svetozar Markovic” consolidated its digitization activities in the project entitled “New Horizons of Digitization in Serbia” carried out with eight partner institutions (National Library of Serbia, Military Archives, University Library Kragujevac, Biblical Institute of the Faculty of Orthodox Theology (University of Belgrade), City Library Novi Sad, Library “Milutin Bojic”, Journalists’ Association of Serbia and BITEF). This project was supported by the Ministry of Culture and Information of the Republic of Serbia. New digitization methods were applied in the project. New approaches to digitization were gradually adopted in the following ways: from the University Library “Svetozar Markovic” cooperation experience with the British National Library and National Library of Luxembourg, as a result of in-house development of digital tools necessary for the application of certain improvements in the digitization process, and through the cooperation with partner institutions on the project and with other institutions in Serbia and the region.

As regards improvements of the digitization process, it should be noted that different segments of the process, from scanning to user interface, are enhanced in daily and project activities at the University Library. These segments are: implementation of colour and size references, user authentification of digital materials, full searchability of digital text via METS/ALTO files, docWorks, a system for storing digital copies Therefore, a system for searching and display of METS/ALTO files and a new presentation device of digital materials Magic Box. Additionally the system is improved via the participation in HORIZON 2020 “*READ*” project.

II Colour and size references

So far in Serbia within the process of digitization as a rule, international standards such as *Technical Guidelines for Digitizing Cultural Heritage Materials* [1] or standards recommended by UNESCO [2] have been consulted when drafting general recommendations until a wholesome national standard is tailored. In coordination with the British Library, the University Library “Svetozar Markovic” has adopted new digitization concepts which contain new standards.

The recommendations for creating digital copies of physical objects refer to the resolution, minimum being 300 dpi (*dots per inch*), or format tiff (*Tagged Image File Format*). In addition to these widely known and broadly applied standards, by meeting the demands of the Endangered Archives Programme [3], we were introduced with the new practices and standards which helped us improve our own digitization guidelines.

Above all, work on this project, has brought about two new dimensions: colour and length.

As a geographical map is closely determined by dimensions and colour, a scan or a photograph (a picture of a digitized object) is more precisely determined with a ruler and a color calibration card. Thereby every scan or every photograph provides more precise data about the genuine physical characteristics of an object, by giving unquestionable information about its precise dimensions and colors. Simultaneously, calibration card and colour scheme give a more precise insight into the colours of an object. This is especially prominent with old archival and physically damaged or endangered materials such were digitized in the EAP project.

Different sources of light have different temperatures. Photographs which are taken under different conditions do not portray precise colours of an object. To avoid this we use White Balance, which is a source based correction. Colours corrected in such a way change balance between red, green and blue curves (RGB *curves*), but not their shape nor position. Therefore, what is changed in the photograph is the light not colour shades. Moreover, when photographs are taken with different devices we do not get identical colours (Understanding color management, 2016). That is why colour management was developed to make such conversions more subtle and to improve the quality of the photograph.

One of the main colour management tools is a calibration card or a colour scheme. Adjusting colours on the photograph to portray genuine colours of an object is a challenge in digital and analogue photography. A Swedish company from Gothenburg *QP Cards AB* developed a cost-effective and efficient solution which is based on an open correction software and calibration cards which we was used within the EAP project. There are several versions available.

Figure 1 – QP Card

QP Card is only one of the accepted models of calibration cards. These cards are industrially acquired, i.e. manufactured in factories, usually made of cardboard and cannot be printed through one's own efforts, especially if they should satisfy a particular standard. They usually contain a ruler and a color scheme.

Color correction in pictures with a QP Card is done by calibration software, *QPcolorssoft 501,* which can be downloaded from the manufacturer's website. This software and a QP Card set up in the scanning surface create a reference profile. They should be set up indirectly to the camera sensor, neither at an angle nor in the shade during the scanning. As the white balance is fixed and a suitable colour profile with

given parameters created, a reference correction profile is created and all other pictures taken with the QP Card can be calibrated.

Calibration consists of the following steps: the QP Card is selected, it is adjusted to the colours of the card so that every colour takes the right place in the pattern, then a specific colour profile and a reference calibration profile are created. When the profile is created all the pictures taken under the determined conditions and the same white balance can be corrected as a group.

Figure 2 – Example 1

Figure 3 – Example 2

If in addition to tiff, as a suitable format for pictures, 300dpi resolution, as the basic minimum resolution, sufficient and necessary for OCR, completely covered surface of the scanned object, from edge to edge, a colour scheme with a ruler was included. A digital object created in such a way would represent an almost ideal picture of the physical object.

III User authenitification of the digitized materials

Quality control and evaluation have to be carried out over the whole digitization process and the potential future standard. In the current project, the University Library followed the EAP guidelines and took the following measures:

- at the end of every workday it is necessary to carry out quality control of the scans;
- scans are copied to the external hard drive and stored at separate locations (back up);

- when the pictures are stored one needs to check if they are rotated properly so that the content can be read;
- prior to permanent storing of the material *MD5 checksum* is applied to detect errors.

The *MD5 checksum* for a file is a 128-bit value, something like a fingerprint of the file. There is a very small possibility of getting two identical checksums of two different files. This feature can be useful both for comparing the files and their integrity control [4]. To understand how this value works, one needs to imagine that there are two physically separated huge files for which it should be determined whether they are the same or different, but which cannot be joined or compared directly. With the *MD5 checksum* it is sufficient to calculate control sums for both files and then to compare them and determine whether the files are the same or different. One should note that the algorithm *MD5 checksum* which was used at first and which is a common practice at the British Library was replaced with the algorithm *SHA2* which is mostly used by institutions in America. The reason for this is security. Namely, American algorithm is more secure and less prone to hacking.

IV METS/ALTO

Full searchability of digital texts is achieved via METS/ALTO files. An institution can adapt its digital objects, pictures, to searchable documents via these files. PDF documents offer very limited search possibilities by keywords. Important metadata such as title, subtitle, author or geometric lines of certain articles are not built into the PDF (Geiger et al, 2011). Searching a PDF results in a marked search term in the text, but not in the picture of the page, which makes locating the search term even harder. Therefore, METS/ALTO standard was developed in newspaper and library communities. METS and ALTO standards were established for easier description of printed materials. The idea was to separate descriptive information from the content of materials in order to manipulate digital objects more easily because when all data are in an XML file (as with TEI – *Text Encoding Initiative* format), the XML file is huge. This standard has already proven its quality when it comes to preservation of digitized newspaper collections and it is widely supported today. The University Library obtained the first such files via *Europeana Newspapers* project. After further development of such files in cooperation with the National Library of Serbia, the University Library can produce boundless files.
METS (*Metadata Encoding and Transmission Standard*) is an XML based open standard established by the Congress Library in Washington in 2001. It is used for permanent storing files which describe digital objects, printed media (books, newspapers, journals), audio and video material etc. METS usually contains several types of metadata standards: descriptive, administrative, structural information, standards regarding physical and logical structure and links to other digital objects, pictures, audio-visual and textual files.

ALTO (*Analyzed Layout and Text Object*) is an XML based open standard also established by the Congress Library in Washington in 2001. It is used for digital description of the printed page layout so that the original page could be reconstructed. This file comprises content of an individual page of a digital document and can contain tags with more data about the very object. It describes styles, layout and the type of

information blocks. All METS/ALTO are grouped in the system for browsing and displaying METS/ALTO files. The collection of these files is searchable and it is displayed via open software. This was achieved in cooperation with the National Library of Luxembourg.

Digital objects structured in such a way will be much more operative and will provide a unique search – in the physical space when it comes to new technologies *Magic Box* and online when it comes to a specialized digital repository – with the results that will provide a detailed overview of collection contents to the user and fast and easy search by the keyword. In addition to the content of the digitized object, ready metadata and expert literature accompanying the theme of the object will be provided for users. Thereby the book is not only digitized but also datafied. Books become *data sets*, i.e. *text corpora*, and words become *data points*. Hence, machines become *readers* (Murrel 2014).

V *docWorks*

The aforementioned transition of images into searchable documents is achieved via *docWorks* application, which we implemented with the support of the Ministry of Culture and Information of the Republic of Serbia. Software *docWorks,* which is the main model of a programme for organizing contents in *Magic Box,* will be used for the preparation of the material.

Preparation of material in *docWorks* [5] consists of the following steps:

- cropping page surfaces of some digital objects;
- zoning objects by segmenting pages into blocks and columns with surfaces defined for OCR (*Optical Character Recognition*) and determining their type as regards the function in the object: titles, text, author, pictures etc.;
- arranging the structure of the object (bullet, chapter, article) by connecting titles and contingent text;
- correcting text and metadata;
- creating ready objects in the form of METS/ALTO files suitable for display in *Magic Box* and the repository.

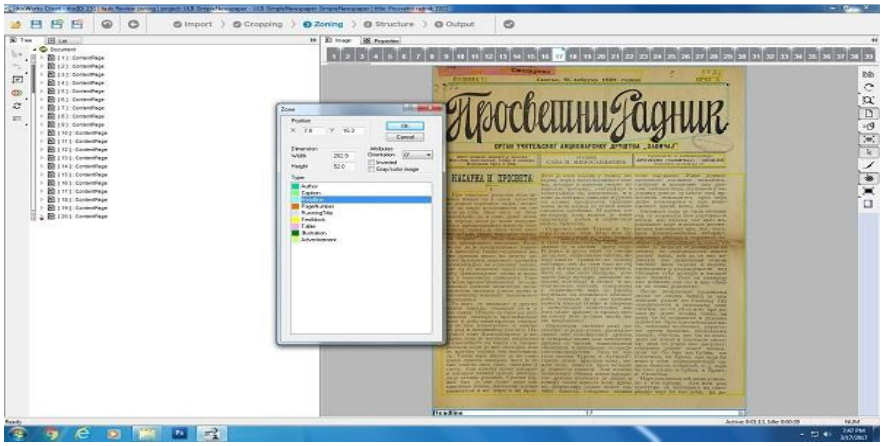

Figure 4 - docWorks

All aforementioned steps imply an automatic analysis and then manual correction.

VI Therefore

During 2014 the University Library “Svetozar Markovic” carried out a project entitled

"University Library "Svetozar Markovic" for the Network of Serbian Public Libraries: Knowledge, Content and Programme Delivery". Within the project we developed a business solution for advancing library activity in the field of working with digital documents. The solution is based on the professional software platform *Therefore* and enables functional connection of the scanning station, regardless of scanner type and software, with the system for automation of library work (*COBISS* or *BISIS*), via repositories of digital objects with advanced functions. Therefore system was obtained in cooperation with the University Library Kragujevac and City Library Novi Sad.

We developed interface for automatic manipulation of metadata, both within the system where metadata for automation of library work are automatically paired with scanned objects and for connecting with entities outside this system via OAI-PMH protocol. In addition, the system facilitates management of business documentation and all other digital documents in the library. Within the system, one can create a timeline, map names and places which appear in metadata and create virtual exhibitions by linking objects after a certain criteria.

Within *Therefore* there are 7 applications:

- *Navigator*: a client application which represents metadata basis and allows for their search;
- *Case viewer*: an application which facilitates working with documents, creating templates, full text search, copying documents etc.;
- *Viewer*: serves for importing and editing metadata and also for downloading and overview of documents;
- *Capture Client*: this application imports scanned material;
- *Solution*: this application facilitates system administration, it is located on the server, where the Library does not have access;
- *Console*: administration of user accounts;
- *Loader*: serves for creating scripts and importing scans and metadata into *Therefore*.

This system is a part of a bigger picture and our idea to form a network of repositories which would introduce new library services such as, for instance, digital interlibrary loan.

VII MagicBox

The final phase of digitization process is presentation of digital documents. That is why the University Library acquired very attractive cutting-edge technology called *Magic Box*. This device is the first in east Europe and our library is the third in the world to own it. It provides transparent browsing of digital material on a touchscreen while the physical object can be seen behind the screen.

Figure 5 – Magic Box

Being aware of the importance of availability of information and open access, especially regarding cultural and scientific heritage, for achieving knowledge society and the role of academic libraries in their dissemination, the University Library's digital collections will be presented to the patrons via the up-to-date technologies.

Trying to keep up with the new technological trends in librarianship, and bearing in mind the protection of rare books, the University Library bought Magic Box with the support of the Ministry of Culture and Information. This is a cabinet display which provides virtual, yet very real experience of leafing through the digitized and protected library collections.

This kind of technology has already become a pulsating window into the world of interests of researchers and experts in the field of history, philology, sociology and also of the wider public thereby encouraging their active participation. This smart device is suitable for interactive display of materials which are too fragile to leaf through. In that way, permanently stored digital copies are less physically used, therefore less destructed, which ensures that they last longer (de Stefano, 2000). Patrons can search through digital content in a completely transparent way, and in addition to digitized print media, Magic Box displays photo galleries, 3D objects and videos.

This means that all interested researchers will be able to search through invaluable archival and library collections in a sophisticated digital way as physical access to these materials is limited due to their fragility. For instance, one shall consider certain materials within the EAP collection. Majority of them were published, but the largest part of the collection was published in small circulation, periodicals and calendars in particular. There were not as many readers at the time and many of these documents were destroyed after reading. That is why they were printed on a low quality paper prone to fraying. On the other hand, when it comes to Serbian war periodicals, they were also published in small circulation due to war and exile. In other words, the majority of digitized collections in the project, and many other from the rich University Library collection are labeled as rare and endangered and cannot be easily accessed. This display will provide a completely new and unique experience to all those interested in rare publications which have limited access which is not adapted to mass use.

VIII Horizon 2020

Participation in international projects is an invaluable experience. In that sense, we didn't have second thoughts about accepting participation in HORIZON 2020 *READ* project[6], even though we don't receive funding. The University Library acquired

TRANSRIBUS within the project and developed methods and techniques for automatic recognition of handwritten text.

This is especially beneficial to social sciences and humanities researchers who somehow always fall behind colleagues in natural and technical sciences due to lack of use of innovative technologies in their research so they struggle to obtain up-to-date information.

TRANSKRIBUS programme enables researchers to upload documents and pictures which are the subject of research and to manage them via aforementioned digital tool, to segment pictures and blocks of text which are the subject of research and to link text with appropriate pictures. Simultaneously, technically speaking, after retyping first 20 pages of text, the algorithm recognizes the rest of the handwritten text. This project will help our library to offer its patrons extremely useful work with valuable materials and yet to be able to control it. Our patrons will get an opportunity to work with one of the most advanced transcription systems which is very important. Created transcriptions of concrete handwritten documents (whose access is limited) will not be lost in private computers of our patrons, but will enrich our digital collections with the possibility of integration into existant repositories in a standardized way. This is an ideal platform for crowd-sourcing projects bearing in mind that within this project there is a crowd-sourcing interface which is easily managed.

IX Conclusion

Multilingual and multicultural Europe creates a challenge for those who provide users with easy and seamless browsing experiences when exploring digital collections. To meet this challenge is to translate digital collections into a live vivid experience in a transparent and objective way via new technologies. All things listed in the text were first implemented at the University Library. We plan to transfer these achievements and results to the academic institutions network and to all interested institutions in public and private sector and NGOs.

The importance of this undertaking is portrayed in richer and more productive analysis of local historical and cultural events. By using advantages of new technologies, the University Library aims to create innovative digital objects online and in physical space, which will attract patrons' attention and focus it to good quality historical content. This content is connected with the need to establish wider and more defined scientific cooperation in the Balkans and in Europe which will be based on global digitization process.

Our future goal and big professional challenge will be connecting displays of digital and physical objects and patrons' experience in virtual and digital space.

New technologies should offer a big breakthrough into different scientific disciplines via open access to cultural and scientific heritage. This is in accordance with the noticeable tendency to use e-infrastructure and innovative technologies as a clear indicator of progress in the field of cultural and national heritage leading towards their full integration into the open science concept (Fresa, 2014).

This could contribute to the development of highly qualified young workforce. Moreover, when it comes to global presentation of library and archival material wider implications are of great importance.

References

[1] http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image_Tech_Guidelines_2016.pdf (accessed 27 Oct 2016).
[2] http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/digitization_guidelines_for_web.pdf (accessed 27 Oct 2016).
[3] Endangered Archives Programme.Guidelines for photographing and scanning archival material.Retrieved from http://eap.bl.uk/downloads/guidelines_copying.pdf, (accessed 01.07.2016).
[4] http://www.fastsum.com/support/md5-checksum-utility-faq/md5-checksum.php
[5] http://content-conversion.com/#docworks-2
[6] https://transkribus.eu/Transkribus/ (accessed March 27th 2017)
[7] De Stefano, P. *Selection for digital conversion*, Moving theory into practice: Digital imaging for libraries and archives, Mountain View, CA: Research Libraries Group, 2000, 11–23
[7] Fresa, A., *Digital Cultural Heritage Roadmap for Preservation*, Journal Of Humanities & Arts Computing: Journal Of Digital Humanities, 8(2014), 107–123.
[8] Geiger, B, Snyder, H, Zarndt, F., *Preserving and Accessing Born Digital Newspapers*, Newspapers: Legal Deposit and Research in the Digital Era, Berlin: De Gruyter Saur, 2011, 31–36.
[9] Jerkov A., Sofronijevic A., & Kavaja D., *Smart and Sustainable Library: Information Literacy Hub of a New City*, Information Literacy: Moving Toward Sustainability, 2015, Springer International Publishin, 22–30
[10] Julien, H., Gross, M., & Latham, D., *Survey of Information Literacy Instructional Practices in US Academic Libraries*, College & Research Libraries, 2017, 17–1024.
[11] Murrell, M. (2014). *The datafied book and the entanglements of digitization*. Anthropology Today, 30:5(2014), 3–6.
[12] Newtown, L., *Data-logging in practical science: research and reality*, International Journal of Science Education, 22:12(2000), 1247–1259.