

DIGITIZATION OF LIBRARY MATERIALS: REVIEW, THE PROBLEMS AND DERIVED SOLUTIONS

Aleksandra Adzic, sandra_adzic@yahoo.co.uk,
Stevan Sremac Public Library, Borivoja Gojkovica 9, 18 000 Nis

Abstract

Libraries are important centers for creating information. The improvement of their services is based on opportunities provided by the rapid development of information and communication technologies. The process of creating a digital library is one of the responses to the demands of users who want to find information in the manner of today. Specifically, we are not talking about new data, but a radical improvement of the way in which data are made available to end users, with respect to time as the only category that is absolutely irreversible in the process of creating information.

Digitization of library materials is a very complex process, regardless of all the technical equipment an institution has at its disposal. Communicating with colleagues about the problems that occur during the initial phases in creating a digital library, difficulties appearing because of the many recommendations and guidelines in this area, the need to detect crucial aspects during the implementation of the digitization process – all of these require to be presented in one place and in an understandable way with possible solutions resulting from practice.

Keywords

digitizing, scanning, digital library, a comprehensive approach

1. Introduction

Until recently, digitization of library materials in many libraries in Serbia has been considered a necessity of the future, something that can be done once the financial, human and technological conditions are met. That future time has already

come and, in Serbian libraries we can already share our own experiences about the process of presenting library collections in a digital format.

Those who need to start digitization are mostly interested in the clear definition of the problems that occur during the initial phases of

the process and possible solutions for overcoming these problems as much as possible. This paper presents the most common problems for librarians in the process of digitization, as well as some of the possible solutions for overcoming these problems.

2. The first phase of digitization of library materials

Before the beginning of the process of digitization it is necessary to provide several things – appropriate technical equipment, the software that will allow the presentation of digitized material and trained staff that is able to carry out the process. All of this implies certain financial investments, so this should be taken into consideration when deciding on the method and pace of digitization. But, first of all, depending on the needs and goals that the library wants to achieve with the digitization, a detailed plan must be developed.

Digitization of library materials is mainly done for two reasons: preservation and access to library materials. Depending on those reasons, every library should carefully make a selection of the library materials that will be digitized. Is the library going to digitize damaged library materials, the material that would be completely destroyed by further usage, or the most precious and most complete material first, or material that is sensitive to damage and that users will rather use in electronic form, or material relating to an anniversary of the library or the town in which the library is located, or native book and non-book materials, or material for which there is a great interest of users – it all depends on the priorities of each library, so the decision about these things should be made in the first, initial phase of digitization of library materials.

When selecting library materials, there is a need to take special care about respecting the Law on Copyright and Related Rights Act, i.e. the right of the creators of literary, scientific and

artistic works. Among other things, the authors are protected from unauthorized copying and piracy by the Law on Authorship Rights. Librarians should be aware of the following: “Copyright lasts for the life of an author and seventy years after his death, regardless of the time when the work was lawfully published. The copyright on an anonymous work that was published under a pseudonym lasts seventy years after the work was lawfully published. After the termination of the validity of copyright, copyright will become public property and may be freely used with the obligation of recognition of an authorship, with the respect of author’s work and honor, as well as the reputation of the author.”¹

3.3. The necessary hardware equipment in implementing library material digitization

The worldwide market offers several types of scanners, but in Serbia there are two types: flat bottom scanners and scanners with a cradle. Depending on the type of the material, the combined use of these two types of scanners is a solution that turned out to be very effective in practice.

The main part of the scanner with a cradle is a V-shaped base, the so-called cradle, and over it there is an affixed transparent panel. When scanning, the publication is placed in the cradle and manually pressed with transparent panel on the top, so that they function as a V-shaped sandwich. In this way, the resulting image is completely identical to the scanned pages: since pages are totally straight, the digital copy does not have distorted lines or shadows. At the same time, publications scanned in this way are protected from wear and destruction, because the publication is not opened completely in the process of scanning, but rather in an angle dictated by the V-shaped stand. The base, constructed in this way, is ideal for digitizing old and rare library materials that had already been damaged by the

passage of time, and the paper is already battered and crispy making complete opening and the pressure of conventional scanner just another step in their physical decay. On the other hand, these publications cannot bear to be handled by the library users anymore, so digitization of this type of material is a priority.

Above the base, on the metal frame a special lever is located with two high-resolution cameras fixed on the lever. On the left side of the frame there is a camera that takes photos on the right side of the publication and on the right side there is a camera that takes photos on the left side of the publication, in an interval of one second. The camera is one of the main components of this entire system, so it is desirable to be of a high quality. The cameras are fixed to the lever and they are moving across the lever by using the ball bearings. The camera position on the lever depends on the format of the scanned picture. On the lever, labeled graphically and in writing, there is a mark for each format separately, so the camera should be fixed manually to the designated place. Furthermore, the camera can be rotated until the optimal position for the scan is set. Cradle is illuminated by special reflectors which are attached to the metal frame and the whole structure is protected by plasticized fabric that does not allow illuminating the cradle with another source of light. If even the thinnest ray of light from another source cuts the trajectory of camera shooting, shadows that look like grease stains appear on the digital copy, while in those places the text is blurry and unreadable. Therefore, it is important to position the screen properly.

Besides the cameras, the position of the V stand can be adjusted depending on the publication. This scanner has an auto-centering stand, therefore during the scan, the center of the book remains in the same place. The ideal position for scanning of a publication of any format is achieved by an optimal combination of

positioning for the V- stand and the cameras.²

The basic technical scanning equipment also includes a computer which will perform the digitization of materials. The software that supports the work of a scanner and the software supporting the work of the cameras must be installed on the computer. With the combination of software settings of the scanner and the camera, depending on the digital images that need to be stored on the computer, the gained result is quality, size and other performances of the picture. The computer is connected to the scanner. Since digitization is a process that takes a long time and the format of images is demanding, in terms of memory, there is a recommendation that the computer should be of such a configuration that will not only support the work of the scanner, but it would also make scanning quick and efficient. It should also be mentioned that, over time, for storing digital copies more and more memory space will be required, therefore upgrading the computer with ordinal adding of hard disks, depending on the quantity of material that will be stored, should be planned on time. If the financial situation allows, for this purpose it is best to use NAS (Network Attached Storage), devices in which hard disks are placed in special containers.

It is desirable that two monitors are connected to a computer. The monitor beside the scanner would be used for monitoring the scanning process, while the second monitor would be used for reviewing the scanned collection and eventual additional work on actual digital copies.

Scanner, computer and monitors should be extra protected from voltage variations using an uninterruptible power supply (UPS), to diminish the possibility that the changes in power system affect or damage the digitizing equipment.

Besides this equipment, another scanner – flat bottom scanner, is used for scanning library materials. This scanner makes great images in scanning materials up to A3 size, primarily

photographs, leaflets and posters, newspapers, publications without binding and old and rare monographic publications, i.e. some samples that are in the form of a paper sheet, so the images in digital form are completely legible. This scanner can be installed on a computer of average characteristics, and it is also necessary to provide hardware protection.

From our experience with technical equipment, digitization of library materials is optimally carried out by combining these two types of scanners: it is recommended that books and all materials with a binding should be scanned using a scanner with a cradle, while the rest of the materials should be scanned by a flat-bottom scanner. With the combined usage of these two types of scanners, the time of digitization is shorter, while materials remain physically undamaged and digital copies have excellent quality.

4. The software support necessary for implementation of the digitization process

For efficient scanning of library materials, hardware equipment must be accompanied by appropriate software solutions. Software packages, that are necessary when digitizing library collections, are installed on the computer that is physically connected to the scanner. An important recommendation is that the hard drive of the computer should have only one partition (there is no justified reason for creating two or more partitions, since this computer serves for digitization purposes only. This recommendation is supported by the rational usage of space on the hard disk. When separating partitions on which required programs will be installed (partition :C), it is always necessary to leave extra space for software updates, but it is impossible to calculate exactly how much space should be left. Actually, the existence of two or more partitions on the hard disk does not interfere with the work process, but such partitioning of the disk has no

particular function).

With the purchase of a scanner with a cradle, two installation discs are included: one for the software that will support the work of scanners, and the other to support the camera. Software that enables the work of a scanner should be installed first, after which options should be set very carefully, and the initial settings should be performed. Then the cameras should be installed following the recommended values and leaving enough space for a number of manipulations. For each following publication, depending on the format of the publication, color of the paper, print quality, sharpness of the image, thickness of the paper, etc... it is necessary to perform an additional configuration to obtain the highest quality of digital copies, i.e. to avoid additional software reworking of scanned images.

Software that supports the work of the flat bottom scanner should be installed using the installation disc, following the manufacturer's instructions. The extent of this software settings is less complicated, so relatively little time is needed to find the optimal resolution, size, etc...

Apart from the above mentioned programs, it is desirable to install a program for image processing. In very few cases the scanned image has desirable characteristics (sharpness, contrast, color, vertical bagging or horizontal position ...), so there is a need to process some digital copies further. In addition, through this program it is possible to determine the size of the image, so installing this software is of great importance in order to obtain the desired characteristics of digital copies. Currently, there is a wide selection of this software, from Adobe Photoshop⁴, which is the most famous one, to the free version of this software, like Artweaver⁵. Usually, it is necessary to adjust the image just a bit, and all software of this type has basic options (which are mostly used for this purpose), and each library can choose software according to their habits or capabilities.

Besides this, it is desirable to install the OCR software (Optical Character Recognition Software for converting images into searchable text) which helps to convert scanned images into text. After testing of several different types of OCR software, we found that many of these software solutions have problems with our individual letters: č, ć, š, đ, ž. For instance software OmniPage® 18⁶ does a great job of translating the Latin alphabet, as well as for tabular displays of data, but it did not perform well with Cyrillic letters. Cvision⁷ is software that works very quickly – it translates up to twenty pages in a second, while frequently used software from the Belgian manufacturer Iris – RadIris⁸ proved to be an excellent software, but both of these software solutions have problems with bad pictures of Cyrillic letters č and đ. Software for OCR, Fine Reader by Abby⁹, showed great results in testing, especially in the interpretation of these letters, but this software has a difficulty to interpret illegible signatures and seals.

The habits and needs of modern users require that the publication must be searchable by keywords, and the search should be done throughout the text, which can be enabled only by using OCR software.

5. Software for presenting digitized library materials on the Internet

Considering the process of digitization of library materials, software that enables the presentation of digitized materials on the Internet is a very important resource. Often, this part of the digitization process is the greatest challenge for librarians. A lot of time was spent to find at least one of the answers to the questions that were imposed: whether to purchase some software? Which one? If it is not freely available (Open Source), how much would it cost the library in the long term? Will the library be in a position to ensure financial resources for a longer period? What about some free versions? Or, would it

be good to join the Digital National Library of Serbia? Or, perhaps create a new program? We had to answer all of these questions that multiplied at every stage of the process. Overview of the situation shows the following: colleagues from Belgrade use software obtained from abroad, with their own innovations; colleagues from The Republic of Srpska use freeware software (Open Source) and colleagues from Jagodina have developed their own software, and therefore, before concluding anything, professional consultations were necessary. In that way, it was easy to recognize the advantages, disadvantages and the solutions that are currently in use in other libraries.

From the technical and organizational point of view, the simplest solution is to display the digitized material within the Digital National Library of Serbia. The procedure is as follows: scanned images must be sent to the National Library of Serbia, in the required resolution, via DVD (Digital Versatile Disc – portable optical drive) or on some other medium. In this case, a library that provides data does not have to have software for the presentation of images on the Internet, as well as technical equipment – the server. It is also not necessary to provide long-term storage, or to train employees who would deal with the process of digitization. On the other hand, the library that sends the data has no direct impact on the development and destiny of the system that allows display of the digital material.

The situation is even more complicated in the case of purchased software. Even if the library would provide funds to purchase the software, maintenance of such software usually requires additional financial investments for its improvement. Libraries are reluctant to such a solution, taking into account the unpredictable funding in the future imposed by economic crisis.

Using freeware (open source) software is a very tempting solution. A lot of libraries in The Republic of Srpska digitized their materials

using such software. There are many free types of software that are upgraded almost every day, but the most complete, most advanced and most widely used is Greenstone¹⁰. This software is relatively easy to install, there is a tutorial on the Internet as well, while the period of updating can be set. Using this software, digitized material is presented in an acceptable manner, moving around the pages is easy and quick, pages can be magnified and displayed in the .pdf format, searching of publications is well designed. From the user's perspective this software is excellent. The only complaint might be placed in the domain of metadata creation, which requires specially trained personnel. So, from the point of view of librarians, there is a specificity of data entry – there are specific fields in which the metadata should be entered that are linked to the publication in a particular way. The data that have already been inscribed into the electronic catalogue are useless, they must be typed again.

In the Library of Nis, after a detailed review of the situation, in accordance with the current financial situation, available resources and personnel that could be employed, it was concluded that the best solution for us was to continue working on the existing software for library automation. Since all library activities in the Library of Nis are already automated with software NIBIS, the simplest solution was using the bibliographic data from the existing electronic catalog. Consequently, it is not necessary to purchase special software in order to create metadata that describes digital objects, but the most practical solution is to upgrade the NIBIS system. As it is the case with all other segments of this software, the work on digitizing materials is very simplified. In this software, the way in which data are entered into the system requires only basic computer knowledge, all the options are offered in the Serbian language, with the fields in which contents are typed in, while a large number of options are offered from the

drop-down menus or buttons. Therefore, the question of special training of employees who work at the operational part of the digitization process is easily solved. Even the data entry is simplified because it is reduced to connecting images mostly with the existing metadata from the electronic database. Briefly, the method of work is as follows: the publication whose bibliographic description is already in NIBIS should be chosen from the search section. Then, in a separate part, in the next window, the name of the file that contains the images is typed in and thus, the path to the image is given after which the type of the material should be chosen from the drop-down menu. For instance, an exact date of publication for a serial publication must be marked in a small calendar. Finally, the choice is confirmed by a mouse click on the given button. Another advantage is that it does not require any additional installation of a special software to support the work of the program, because it is all done through a web browser (Windows Explorer, Mozilla Firefox, Google Chrome ...), so the work is available from any computer that is in the LAN (Local Area Network) while, with the knowledge of a static IP address, it is possible to work with any computer that has an Internet connection, no matter where it is. The most important fact is that the existing mode for searching the electronic database in the library and information system NIBIS is used for searching digital collections – in the search window one should only mark the option to search digitized material. The software is installed on the server located in the Library building, which, from the point of maintenance, is the best option. At any moment, in accordance with the progress of information technology, engineers, in cooperation with librarians, can make changes within the system, have an influence on its development and upgrading, as well as the tempo of changes.

This improvement of NIBIS, in the form in which it is now, is not applicable to other

systems. Thanks to the part for digitization of library materials, NIBIS can be used for digitization purposes by all libraries with this library information system.

Finally, it is important to mention the following: whichever system a library chooses, the platform on which digital collections are stored should allow conversion of metadata, so that all collections can be complete even in a more advanced system that will be implemented in the future, and digital collections should be visible through large aggregators of metadata at a global and European level. Depending on the internal policies of libraries, it would be useful to make a decision that all digital collections should be associated with the European cultural and scientific resources available for all in Europeana¹¹. The conclusion is that all digital collections need to be implemented in one system in which European cultural heritage is available through the multilingual service.

6. The scanning process

In the case of scanning performed by the scanner with a cradle, the procedure is as follows: after placing the publication in the cradle, a limiter regulates the position of the publication in order to avoid displacement of the publication when flipping pages, since every movement of the publication means re-setting software and additional work on the computer that is physically connected to the scanner. After the publication is properly fitted and pressed with a transparent panel on the top, the cameras should be set. The cameras should be set up physically first, then, depending on the format of the scanned publications, the optimal position on the lever should be found as well as the angle of the camera shooting. The wheel of cameras should be set to Manual, so that any adjustments can be performed separately. After that, the software settings of the cameras should be performed, after which the shooting of images

can start. First, a test shot should be done, the one that is not permanently saved on a computer, but just shown on the screen to see what digital copies look like. At this point, when a sample image is shot, final preparations should be done: the size of the image is to be determined by moving the limiters using the clicks of a mouse, which, after shooting, is automatically moved to the designated folder where the images would be stored; next, the condition of the batteries should be determined, and so on. After this step, either the manual shooting could be chosen (each camera makes one shot by pressing the button on the keyboard) or automatic shooting (camera takes the picture immediately after turning the pages of the publication, at the moment when the transparent panel touches the publication. This option is selected in the case when there will be no further moving of the publication, by the end of the work). It is possible to scan 700 pages per hour, according to the user manual of this scanner. In practice, we found that this number is much smaller, as the number of the scanned pages depends mainly on the condition, type, binding, size and other characteristics of the publication, and, also, on the competence of the person who is in charge of the scanning. Anyway, scanning with this scanner is much faster than scanning with a classic desktop scanner.

The process of scanning on the flatbed scanner with fixed-papers is simpler. Software settings are minimal and they can be done in the window that is offered during the scanning process. Options are selected once, and only when the appearance of the required image changes, there is the necessity for option change. Image cropping should be done before the scanning.

7. Formats and Image Processing

After scanning, the images get their digital format and they are stored on the hard disk of the computer. However, these images are not ready for the presentation on the Internet yet.

Depending on their resolution and on their extension, those images usually occupy large memory space (they take up to 25 MB). They need to be modified to about five or six times smaller size, so that Internet users may look through the images without too much waiting until they load. Depending on the software that would be used but also on the size of the space for storing images on a hard disk, the decision about file extensions should be made in advance. Usually the images are in .TIFF (Tagged Image File Format) file extension (which allows the creation of a high-quality archival file) or. PDF (Portable Document Format), but these images are not small enough. The most acceptable solution is .JPEG (Joint Photographic Experts Group) extension, because the texts of the digital copies with this extension are well readable and the picture does not take up much space on the hard disk, while the user can open them quickly because the images are reduced to the size range of 300 - 600KB. Depending on the plans of a library, it is necessary to save digital copies with the original extension because only the images with high resolution can be used in printing processes, or eventually in making phototype edition. Besides making digital images in different extensions, all images should be additionally reviewed and, if necessary, further processed in some of the programs for image processing. In practice, the processing usually involves additional lighting, contrast adjustment and sharpness of some digital copies, so that these digital copies could be presented on the Internet in the best possible way. If the institution uses some of the OCR software, this is the moment to convert images to text.

8. Long-term storage of digital copies

Technical equipment of the best quality is not a guarantee that digital collections will be permanently saved if they are stored in just one medium. Professional approach to a digitization

process includes a pre-prepared solution for long-term storage of digital copies. There are several options, so the library decides upon one or more possibilities. The scanned images are automatically sent to the selected location on the computer, using the software that comes with the scanner. If a library has a Server computer, it usually has an additional hard disk within its case. Performed with a minimal software setup, all files will also be copied to the second hard drive. However, one copy of the scanned images within the computer itself is not enough. It is necessary to make at least another copy.

Digital copies can be burned on a DVD disc, which is a way to secure another copy. But, some studies have indicated that the content that is burned on the DVD disc has a "shelf life" of about five years, so, after that time, no one can reliably confirm how long it takes before the data is completely lost. For this reason, storing digital copies on DVD disks is not the best solution. Although this is the most economical option, storing digital copies on DVDs can be justified as a short-term solution for a small number of data that will be transferred to a safer place as soon as the conditions are met.

A good solution is to place a copy of the scanned material on the hard disk of another computer, but this computer should not be in the same room where the computer connected to the scanner is, for practical reasons (flood, fire, electrical troubles ...). Taking into consideration the fact that the size of the database will constantly increase, it is obvious that one hard disk will not be enough for storing the entire content, and that there will be a necessity, from time to time, for adding hard disks to the computer that is connected to a scanner and to the computer for data backup.

There is one more solution for data storage, and that is hiring of space on the hard disk of a distant FTP server. These server computers do not even have to be in the same city or in the

same country, and the data, stored on them, are relatively safe.

Digital copies can be stored on external hard disks, too. It is necessary to pay attention to the size of the external hard disk (to be purchased with the best capacity). The external hard disk should not be pre-divided into partitions (the partitions are never completely filled, so a lot of unused space is left).

9. Staff training

Every library should take special care about training the employees to ensure that all digital processes will be carried out in a way that is pre-planned. Staff engaged in the digitization of library materials have to acquire the skills and knowledge required for scanning, image editing and translating images into text if the libraries make such a decision. It is very important to provide users with constant access to digitized material, as well as the long-term protection and preservation of the scanned material. The training and the supervision of the staff is performed by the system engineer of the library or the technical employee who is sufficiently trained for this job. Usually, the training is performed next to the scanner, with practical examples, in a relatively short period of time.

10. Experience

The Stevan Sremac Public Library began the scanning with the newspapers "Narodne novine". The scanners that are used for scanning are: one with the cradle ATIZ BookDrive Pro¹², as well as the flat-bottomed scanner Epson GT 20 000¹³. Up to now all specimens between 1949 and 1977 were scanned. Two experts work on the digitization process every day. Each scanned page is verified before the final storage of digital copies. Some digital copies need to be reworked through software before storing in the corresponding file. The time required for making digital copies varies depending on several

factors: the condition of newspapers (some pages are either worn out or fractured, or separated), some copies must go through an additional software treatment (some pages are bound crookedly so it is necessary to place the volume in the cradle in a way that is different from what is suggested, then sometimes the images do not have the desired sharpness, some are too light or too dark, etc.). Backups to another hard disk for digitized library material are regularly performed in the library which is just a short term solution, because only one copy, which is located in the same office as the original, does not mean that the data are protected. The final decision about the long-term storage of digital copies has not been made yet. All digital copies are visible on the Internet, through the official website of the Library. They are searchable by all parameters entered in the bibliographic description of publications, but they are not searchable by text, because the conversion of images to text has not been done yet.

References and sources:

- 1 Википедија. http://sr.wikipedia.org/wiki/Autorsko_pravo
- 2 Atiz. <http://pro.atiz.com/> (приступљено 20. 07. 2012.)
- 3 Epson. http://www.epson.com/cm_upload/0/000/142/225/GT20000_ProdSpecs.pdf (приступљено 10. 07. 2012.)
- 4 Adobe Photoshop. <http://www.adobe.com/rs/products/photoshop.html> (приступљено 27. 05. 2013.)
- 5 Artweaver. <http://www.artweaver.de/> (приступљено 27. 05. 2013.)
- 6 Nuance Communications. <http://www.nuance.com/for-business/by-product/omnipage/> (приступљено 27. 04. 2013.)
- 7 CVISION Technologies. <http://www.cvisiontech.com/?lang=eng> (приступљено 27. 04. 2013.)
- 8 I.R.I.S. Product & Technologies. <http://www.irislink.com/c2-2115-189/Readiris-14--OCR-Software--Scan--Convert---Manage-your-Documents-.aspx> (приступљено 10. 04. 2013.)
- 9 ABBYY. http://www.abbyy.com/personal_homeoffice/ (приступљено 10. 04. 2013.)
- 10 Greenstone tutorial exercises. http://wiki.greenstone.org/gsdoc/tutorial/gs2-current/en/all_tutorials.html (приступљено 23. 03. 2013.)
- 11 Europeana. <http://www.europeana.eu/> (приступљено 27. 03. 2013.)
- 12 Atiz. <http://pro.atiz.com/> (приступљено 27. 05. 2013.)
- 13 Epson. <http://www.epson.com/cgi-bin/Store/jsp/Product.do?sku=B11B195011> (приступљено 27. 05. 2013.))

Received: 5th March 2013

Accepted: 21st May 2013