Преглед НЦД 16 (2010), 35–42

**Branislav Tomić**

# OLD SLAVONIC MANUSCRIPT TEXTS
## - AN ONLINE DATABASE -

**Abstract**. This paper describes an online database application for Old Slavonic texts, designed for The National Library of Serbia. Its purpose is to present cultural heritage, consisting of medieval Old Slavonic manuscripts, online. It allows the user to perform searches of Old Slavonic text content and view images of original manuscript pages simultaneously, as well as translations and basic manuscript info.

It is a system that consists of two database applications, one for data preprocessing, and the other one, for online data display. The first application ("PrePis"), is a desktop application with user-friendly interface to be used by average user with ease. Its purpose is to enable easy gathering and preprocessing of manuscript page images and texts written in Old Slavonic. It was created in FileMaker software. After preprocessing, gathered data is uploaded to the web server that houses the second application and MySQL database. This second application, is a designed for web use. It communicates with online users and MySQL database, by accepting search requests from users, fetching the requested data from the database, generating html pages and sending them back to the online user. It was written in PHP programming language.

It can be accessed at: http://digital.nb.rs/rukse

**Keywords**: digitization, cultural heritage, database, Old Slavonic, medieval manuscript

As with any project, to be able to successfully complete it, its goals have to be defined first. These goals, in turn, dictate the workflow and the technologies that have to be used.

## 1. Project goals

**1.1. Presentation of cultural heritage online.** Medieval written cultural heritage in this part of Europe consists mostly of manuscripts written in Old Slavonic language, using Old Slavonic, or as it is also known, Old Cyrillic alphabet. At this point in time this heritage is described and presented in different books and journals, some written more than a century ago, usually in small runs, therefore making them very hard, and in some cases impossible, to find. So, I thought that it would be great if I could present them in a modern way. The best way for this is, of course, using the internet.

**1.2. Gathering of already published Old Slavonic manuscript texts into a single database.** To be able to present them, I had to locate them first. Well, I was in luck there, because I worked on this project for a library that had most of them. Others were to be found elsewhere. I also needed to find the images of original manuscripts in the highest quality possible, which was no easy task. After finding the books they had to be digitized and converted into text form. This was rather hard, considering it was mostly Old Slavonic. As a second step, these texts and images had to be loaded into a database, so that they can be used online effectively.

**1.3. Ability to search Old Slavonic content.** In order to be able to study this material online, we have to be able to search not only manuscript descriptions, but Old Slavonic content as well.

Old Slavonic texts were presented in books and journals in an inconsistent way, as every author had slightly different ideas on how to present them. In order to be able to search them,

they had to be converted, so that they are presented in a uniform manner, following the structure of the original manuscript. This would make it possible to quickly find words in one or all texts in the database, as well as their exact position within the original manuscript.

**1.4. Display of original pages alongside transcriptions.** Displaying original images alongside the transcribed text is essential. First of all, we need to have the ability to always compare the transcribed text against the original, just in case we have some doubts to the correctness of the transcription. To be able to check this thoroughly we need to be able to zoom in the detail of the page.

**1.5. Ability to view translation (Serbian or English) on demand.** The application interface is bilingual. Most of these texts have translations in Serbian, and some in English language. So, at any time, we wanted the viewer to have the capability to see either translation of the whole text scrolled down to the position of the page that she/he is looking at. This should be displayed via pop-up window that can be moved around the screen, to facilitate comparison to either transcribed text or the image of original page.

**1.6. Show basic info on manuscripts.** All the manuscripts have the same metadata attached to them, bibliographical information, or some interesting texts from the history of the manuscript. All of these should be visible in a similar fashion as the translation.

## 2. Technological obstacles

To be able to achieve all goals set, several technical problems had to be overcome.

**2.1. Software that supports Unicode fonts in all phases of workflow.** In different phases of workflow it was necessary to use software either custom designed or off the shelf. Since in this project, we are mostly dealing with Old Slavonic text, which relies heavily on Unicode standard, all software had to be able to work with this standard. This was not a large problem, since most software is Unicode compliant. Only the drivers that handle transfer of data from one database to another were a little problematic, since most of them do not work correctly with Unicode.

**2.2. Find/design a font that can display Old Slavonic texts from different sources in a uniform way.** It was decided that **the** font that was to be used, was going to be "Monah5". This font contains over 4000 Old Slavonic characters and glyphs. We actually use only a small subset of these, but in case a different need arises, this font can handle it.

**2.3. Method of display of Old Slavonic texts on user's computer when user does not have the necessary font installed.** Once the font was chosen, I had to find a way to display it in the user's browser. This is not a standard font, so users most likely do not have it in their systems. This problem was solved by writing a routine in PHP programming language that converts Old Slavonic text to 'png' graphic files, before it is sent to the user. Since the user receives 'png' files instead of a text, there are no display problems. Of course, because we are dealing with pictures instead of a text, a broadband internet connection is necessary. It was tested at 128 mbit connection speed, and performed satisfactorily.

**2.4. Communication between user and server using Old Slavonic character set, when user does not have the necessary font installed.** This is a similar problem to the one just mentioned. The user does not have the necessary font. So, how is he going to type the search request? Since most characters used in this database already exist in the Cyrillic code page of

the Unicode standard, for those we could just use the normal Cyrillic. For characters not available in a standard font, the number codes are used. Only 15 Old Slavonic letters and all numerals are not in compliance with the Unicode standard. When typing a search request we are usually typing only a part of the word, therefore it is not a problem to type a character or two using number codes. The codes table is displayed on the screen, next to the search box. These number codes are converted to their respective characters, when they reach the server, before the search is performed.

**2.5. Easy processing of input data (pictures and texts) before uploading to server.** In order to be able to work efficiently in preparing the data to be uploaded to the server, I had to design a solution that would facilitate this task. This solution can be used by anyone, even if this person is not very savvy with computers. I will describe this application in greater detail later in the text.

**2.6. Fast download and display of large pictures with ability to zoom in.** One of the design requirements was that we should be able to view images of manuscript pages, alongside Old Slavonic text. Sometimes, these images can be quite large, depending on the manuscript page size. For instance some of the documents in the database can be scrolls. These are typically very long, up to several meters. This makes their images very large and totally unsuitable for download over internet. This problem was solved by using technology called "Zoomify". With this technology an image is broken into tiles, and only tiles that are to be actually viewed onscreen, are sent to the user. This speeds the display of images considerably, and even enables zooming.

## 3. “PrePis”, an application for data preparation

In order to be able to successfully populate an online database, first we have to gather and process data. To simplify this task I deigned a desktop application, called PrePis.

This too is a database application, this time created in FileMaker. It achieves several things:

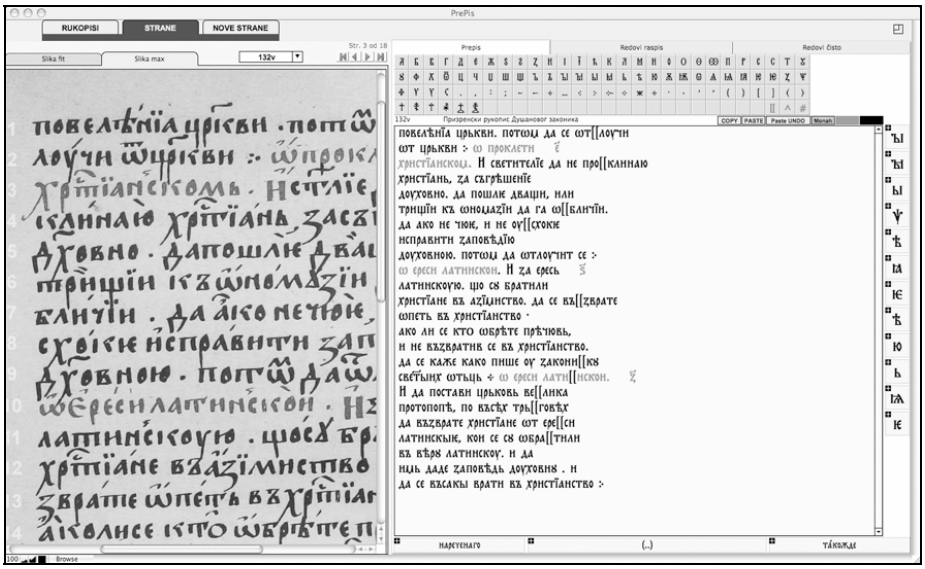
First of all, it allows gathering and management of manuscript page images. Before these images can be loaded, they have to be processed in an image editing software such as Adobe Photoshop. We need to create images in jpg format Recommended resolution is 150 dpi at 100% of manuscript page size. This enables us to zoom to approximately 200% in a web browser. If we need higher magnification, we can use higher resolution.

These images are loaded into PrePis software, whole folder of images at once. We need to put all page images of a manuscript into a single folder, named as page in the manuscript numbers (15r, 15v etc.), so that they sort in the correct order. As they are loaded, a record in a database is created for each image. The images can also be loaded one at the time or replaced, but batch loading is the simplest way to go.

After the images are loaded and records, one per page, created, we can input the text for each page. This text can be typed directly or pasted form another file. In most cases, we are dealing with already transcribed material present in books. OCR software can be used to obtain this text. We can of course type it directly into the field provided in the database. To facilitate this task, manuscript page image, as well as on-screen keyboard are conveniently placed on the screen. We can use a standard keyboard, set to Monah5 font (we get all characters) or standard keyboard, set to Cyrillic (we get all Cyrillic characters, the missing ones can be typed from on-screen keyboard) or we can use on-screen keyboard exclusively.

If the text was obtained via OCR, the correction should be done in PrePis. There are a

few rules that have to be followed when typing. The text has to be structured as it is structured in the original, line by line. Words that are broken between lines of text are moved to the upper line, with line end marked within a word. This enables software to recreate the original look of the page for display, while at the same time defines every word as a whole word. This is used when performing searches or generating vocabularies.




1. Input screen of application PrePis, where text is edited.

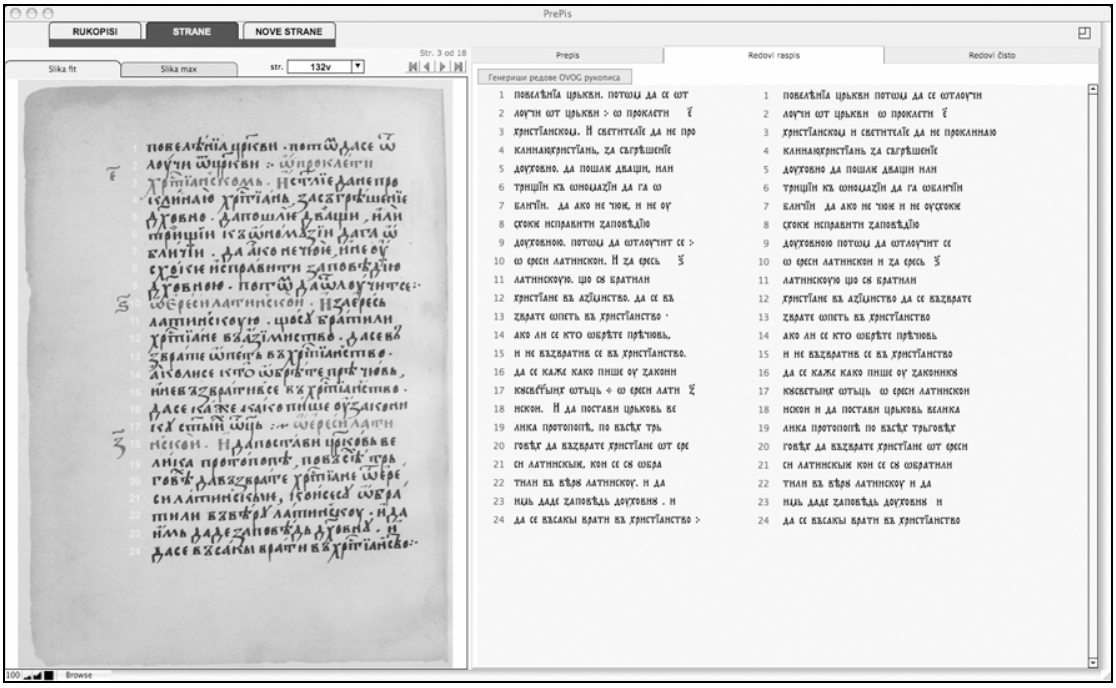
Once a manuscript text is input and corrected, we can process it, by activating a script for generating lines. This script splits page text into separate lines. Thus, each line becomes a record in the lines table of the database. The lines table has two columns for two different versions of the same line. One column contains lines of the text exactly as they are presented in the manuscript, with words broken between lines, capital letters, punctuation, accents, etc. The other column contains text lines simplified in the following way: The entire text is converted into lower case, all punctuation is removed, each word is surrounded by two spaces, any accented letters are converted to their base letter any alternate glyphs are converted to their base character. This simplified text is used for performing searches in the online database. It is never shown to the user. The user sees only the column with the text which mimics the original page look.

Why would we need to do this? First of all, there is no collation table available for Monah5 character set, because it is partially outside the Unicode standard. Therefore, case insensitive search is not possible. So, we use only lower case letters. Several glyphs can mean the same character, so we convert them into their base character, and search for that. Accents are removed for the same reason. Spaces are added at the beginning and end of the line so that we can search for word beginnings and endings. Full-text search will not work with this font, so we have to use this workaround.

After the text in form of lines is generated, it is transferred to MySQL database, located on the web server. In this case, the server is located in National Library of Serbia, so ODBC connection, was used.

Images have to be converted into a format suitable for Zoomify technology. This generates a lot of small files. It can be done with Photoshop or proprietary software provided by Zoomify. These images can be transferred to the server, using FTP connection, or copied

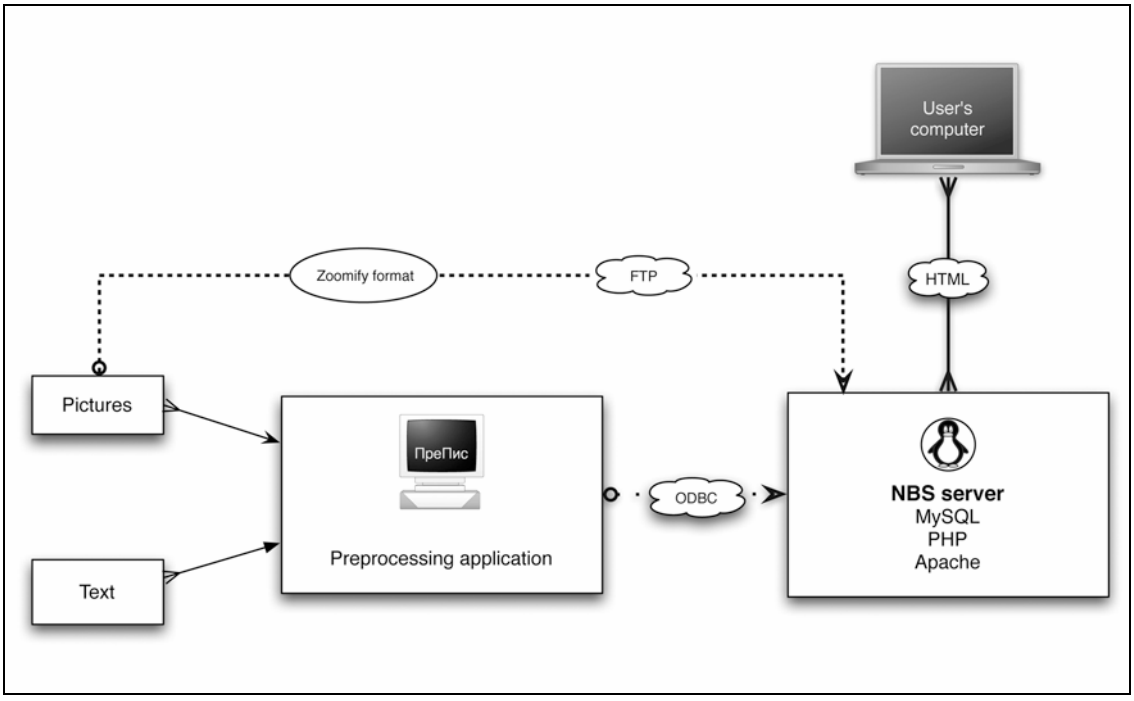
from external disk, if one has physical access to the server.




2. Screen with display text column and search text column

If we discover errors in texts already uploaded to the sever, we can go back to the manuscript text in “PrePis”, correct it, generate lines and upload it to the server again.

At his point we have everything needed to display the manuscript online.




3. Workflow schema

## 4. An online database

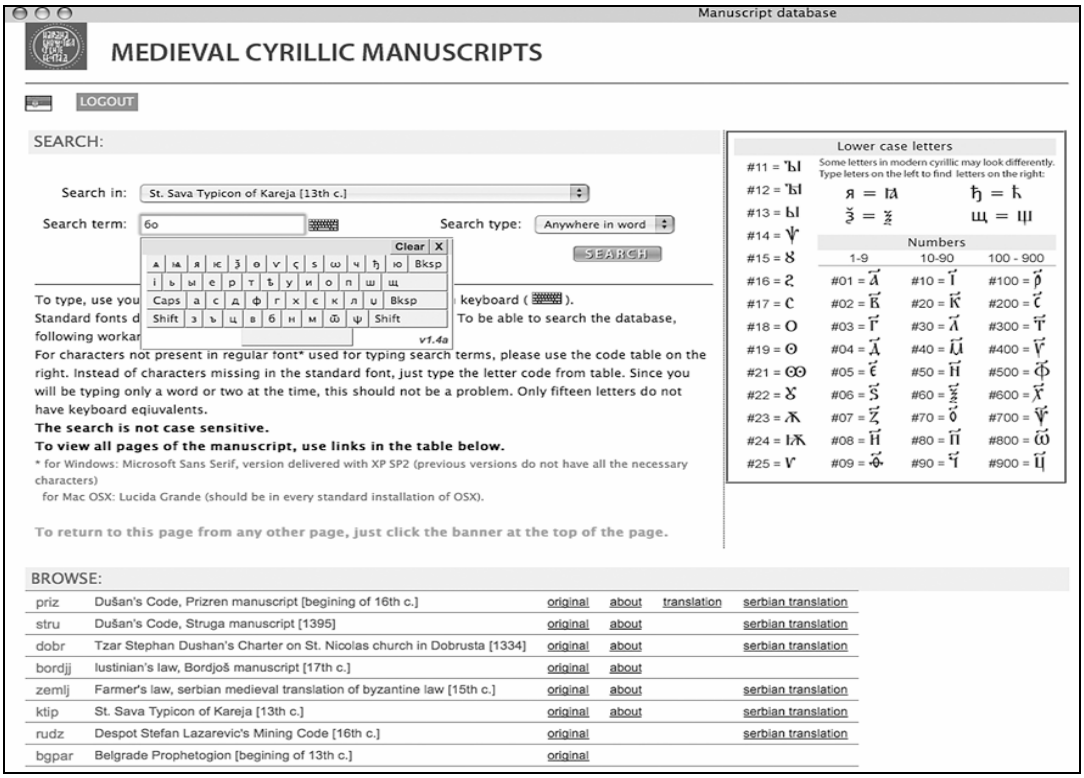
Now that we uploaded some data to the server into the MySQL database, we need an application on the server to communicate with both the user and the MySQL database. This application was written in the PHP programming language. It generates dynamic HTML pages, at user's request.

For example, the user chooses an interface language (English or Serbian), the application responds by sending pages in that particular language, or the user enters a search request and clicks the 'Search' button, an application responds by performing a search in MySQL database, gets the results and creates an html page to display them an sends it back to the user. This is basically how any online application works.

What can it do?

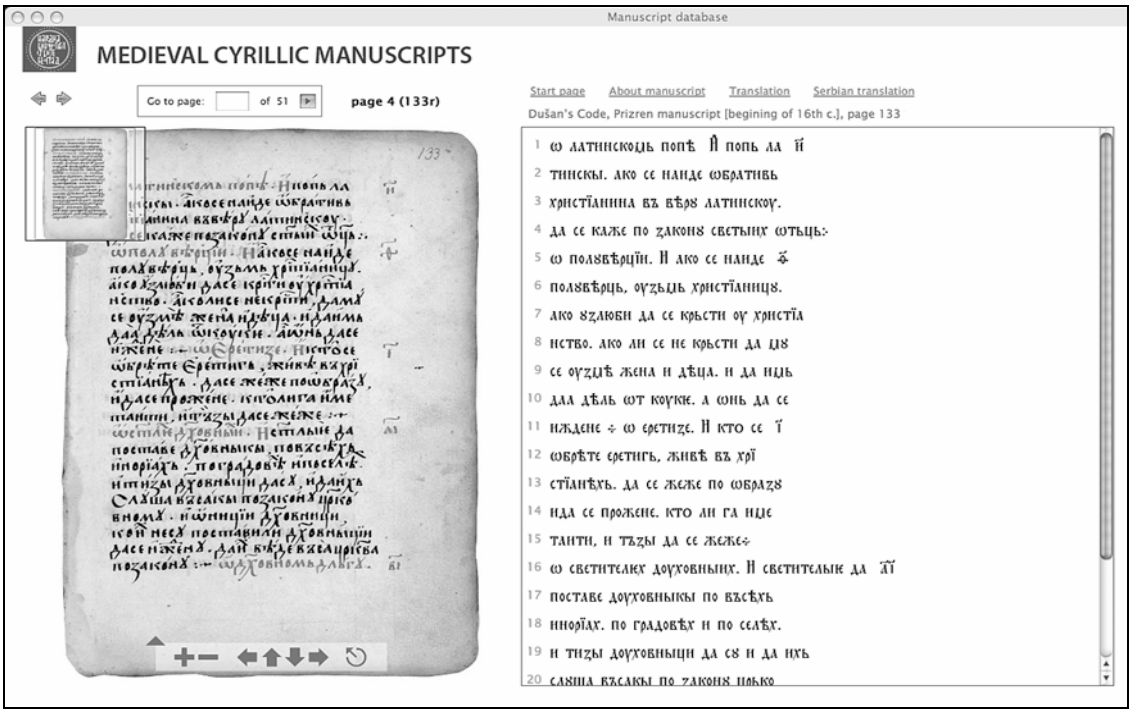
• First of all, it enables you to perform searches of any one or all Old Slavonic texts that are in the database. Available options are to search beginning of the word, end of the word, whole word or anywhere in the word. To facilitate input of Old Slavonic characters an on-screen keyboard is provided. It contains all Cyrillic letters that are in the Unicode standard. The Old Slavonic letters missing from this set are typed as number codes from the table that is provided on the screen.

Search results are displayed in a standard fashion, in a paginated table, showing lines from manuscript(s) that contain the word that matches the search request. Each of these lines is labeled with the manuscript that it comes from, as well as with its position within the manuscript. By clicking on a particular line, that we want to see in the context of the page, an image of the page is displayed, alongside the transcribed text of the whole page. The line in question is highlighted. To be able to study the page image closely, we can zoom-in the image. After doing it, we can return to results page, by clicking the back button, and choose another line to inspect.




4. An online database main page, with on-screen keyboard shown.

• If we do not wish to perform searches, but want simply to browse manuscript texts from the database, we can do so by clicking the appropriate links from the table on the lower part of the start page. When the link is clicked, the first page of the chosen manuscript text is shown. We can then navigate trough the text in a standard fashion, by clicking back or forward arrows or alternatively typing a page number that we want to go to.




5. Page display when browsing the manuscript or viewing a page as a search result.

• Whenever we are on any page of any manuscript from the database we can view its translation (English or Serbian, whichever are available) by clicking the appropriate link. The translation of the whole text is displayed in a pop-up window, and automatically scrolled down to the position that corresponds to the page that is on screen. Since the whole translation is displayed in this pop-up window, we can scroll up or down to see the observed page in the context of the whole document. The use of a pop-up window is handy, since we can move this window around, and are able to position the translation alongside either the image of the page, or its transcription.

• In the same fashion as the translation, we can see the basic information about the manuscript itself. This info window usually contains the bibliographical information, paleographic description of the manuscript, as well as an article about the manuscript, whichever are available.

• The translation and info can be also viewed independently by clicking the appropriate link on the start page.

Hopefully this database will be useful and interesting to both researchers and casual visitors. It can be accessed at: http://digital.nb.rs/rukse

**Branislav Tomić**

## STAROĆIRILIČNI SREDNJOVEKOVNI TEKSTOVI - INTERNET BAZA PODATAKA -

U ovom radu dat je prikaz internet baze podataka za skladištenje staroslovenskih, odosno staroćiriličnih tekstova. Ova baza podataka izrađena je radi postavljanja na sajt Narodne Biblioteke Srbije sa ciljem da se prikaže naše srednjevekovno pisano kulturno nasleđe.

Njene karakteristike su:

- Mugućnost pretraživanje staroćiriličnog sadržaja pojedinačnih rukopisa ili svih rukopisa istovremeno.
- Paraleni prikaz slike originalne strane iz rukopisa i raščitanog teksta.
- Mogućnost pristupa prevodu prikazane strane rukopisa u svakom trenutku.
- Mogućnost pristupa inernet strani sa informacijama o rukopisu, kao što su na primer, bibliografski i paleografski opisi i drugi tekstovi čija je tema prikazni rukopis.
- Dvojezični korisnički interfejs.

Inrernet baza je deo sistema koji se sastoji od dva programa:

– Prvi program, pod nazivom "PrePis" izrađen je da posluži prikupljanju fotografija strana rukopisa kao i tekstova pisanih staroćiriličnim pismom.

Nakon obrade u ovom programu, podaci se prosleđuju severu Narodne Biblioteke Sarbije i skladište u bazi podataka (MySQL).

– Drugi program je internet aplikacija, koje se takođe nalazi na serveru NBS. Ovaj program, pisan u programskom jeziku PHP, ima ulogu komunikacije između korisnika i MySQL baze podataka. Naime, on prima upite korisnika prosleđuje ih bazi podataka, nakon izvršenog pretraživanja prihvata rezultate, formira dinamičke html strane i prosleđuje ih korisniku.

Sve opisane transakcije koriste staroćirilično pismo. Da bi se ovi programi uspešno izvršavali primnjena je tehnologija za prikaz straoćiriličnog pisma na računaru korisnika, bez preduslova da mora posedovati staroslovenski font.

Bazi je moguće pristupiti na sledećoj adresi: http://digital.nb.rs/rukse

**Ključne reči**: digitalizacija, kulturno naseđe, database, Staroslovenski jezik, Staroćirilični srednjovekovni rukopisi

tomicb@ikomline.net