Преглед НЦД 14 (2009), 29–37

Vladimir Risojević
(Faculty of Electrical Engineering, Banjaluka
Republic of Srpska, Bosnia and Herzegovina)

Dalibor Pančić, Bojana Milošević,
Ranko Risojević
(National and University Library of the Republic of Srpska,
Banjaluka, Republic of Srpska, Bosnia and Herzegovina)

DIGITIZATION PROJECTS AT THE NATIONAL AND UNIVERSITY LIBRARY OF THE REPUBLIC OF SRPSKA

Abstract. We present two digitization projects at the National and University Library of the Republic of Srpska. These are digitizations of magazines „Razvitak“ and „Školski vjesnik“. We describe the material, as well as procedures for scanning and creation of metadata. We briefly describe Greenstone, software for creation and distribution of digital collections. We also describe the design and implementation of web-based user interface for accessing these collections. The collections we created are searchable by titles, keywords and authors' names, and hierarchically browsable by volumes and issues or by the index of authors.

Key words. Digitization, cultural heritage, Greenstone

1. Introduction

For centuries the main purpose of libraries has been preservation of human knowledge. In the past this goal was identified with preservation of written, and later, printed material. The libraries of the past aimed at acquiring and cataloging each and every existing publication. This has been the main guiding principle behind the operation and growth of libraries and, as a result, libraries have grown exponentially. However, more recently we witness the growth of the number of publications, introduction of new media and lines of dissemination of information. In the same time, users' information needs also grow. Consequently, it became clear that by limiting their activities to traditional printed books, libraries would lose their central role in preservation and dissemination of knowledge. Moreover, even if they limited their activities to printed publications, libraries would not be able to allocate enough space for storing the vast body of publications.

In order to maintain their position in human society, libraries have to answer the challenges of the information revolution – increasing need for storing, organizing and accessing the information. It is not publications that are central to the today's library but information. Appropriately organized information constitutes knowledge. Therefore, in order to meet its old purpose, today's libraries have to meet the information needs of their users.

Technological developments opened new possibilities for meeting this purpose. For example, nowadays computer-based library catalogs are de facto standard. But,

lately it became possible to include digital publications to databases in addition to the metadata. As a result of this digital libraries have emerged. In their book "How to Build a Digital Library", Ian H. Witted and David Bainbridge define a digital library as "a focused collections of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance of the collection" [1]. In shifting the focus towards information it became less important how the information is embodied, and, therefore, the subject of librarians' interest became "digital objects".

In this paper we present two digitization projects at the National and University Library of the Republic of Srpska, namely digitizations of magazines "Razvitak" and "Školski vjesnik". These are the first two digital collections of the National and University Library and the first digitization projects at the Library. The projects started in September 2007 and, at the moment, the digitization of "Razvitak" is complete, and the digitization of "Školski vjesnik" is nearly finished. We added metadata according to the Dublin Core standard [2] to the collections of scanned images. Collections are managed using Greenstone software for building and distributing digital collections. Among various abilities of Greenstone are publishing collections on the web and access and searching of collections using a web browser. At the moment, only a part of these collections are published at the website of the Library[1].

This paper is organized as follows. In section 2 we describe the material for digitization. Greenstone is briefly described in section 3. In section 4 we discuss details of the implementation. This is followed with conclusion and references.

2. The material

Projects described in this paper are concerned with digitization of two magazines, "Razvitak" and "Školski vjesnik". "Razvitak" was started in 1910 by Petar Kočić [3]. It was published monthly in Cyrillic script. Six issues were published before June 1st, 1910 when Kočić was elected to the Bosnian parliament and stopped publishing "Razvitak". A group of intellectuals, continuing the tradition of Petar Kočić's "Razvitak", started a magazine with the same name with the purpose to study lifestyle of the people, to help cultural development and to cover events in the country and abroad. The first issue appeared on January 1st, 1935, and the last on April 1st, 1941 when World War II ended its fruitful life. Overall 76 issues were published.

Authors in "Razvitak" dealt with various issues of life, past and present of the people of Krajina. They tended to objectively research historical events, and to assess conditions of the time. The magazine was more inclined towards folklore than art, and it continually pointed that it was necessary to improve the lifestyle of the people. Consequently, this magazine is an important source of information for every researcher of the past of Krajina.

We digitized all 76 issues (8 volumes) of "Razvitak". This collection contains 2,712 pages. Master files for this collection take about 300MB.

"Školski vjesnik" first appeared in January, 1894, and it was a professional magazine of the State government of Bosnia and Herzegovina, as was printed on it [4]. From the very first issues „Školski vjesnik" dealt with pedagogical past in Bosnia and Herzegovina, and history of schools in Croatia, Vojvodina, Serbia, Montenegro, and Dalmatia. „Školski vjesnik" also covered the most important pedagogical events in

[1] http://www.nubrs.rs.ba/dl/

Europe and worldwide. It is very valuable for the researchers of the history of education in Bosnia and Herzegovina because it covered all personal changes in schools (appointments, promotions, transfers, and awards of the teaching staff).

We have so far digitized 11 volumes of "Školski vjesnik", from 1894. to 1904. which amounts to over 11,000 pages. Master files occupy about 1.80GB.

Both magazine were printed in black and white. There are no illustrations in "Razvitak", and all illustrations in "Školski vjesnik" are in black and white.

3. Greenstone – software for managing digital collections

Greenstone is a software for building, managing and distributing digital library collections. It is developed at the University of Waikato as a part of the New Zealand Digital Library Project[2]. Nowadays Greenstone is developed in cooperation with UNESCO and Human Info NGO from Belgium. It is issued under the GNU General Public License, and it can be freely used. Moreover, since it is an open source software, users are free to modify and adjust it to fit their needs. At the moment, Greenstone is used in a number of digitization projects, for example:

- UNESCO MOST digital library, http://digital-library.unesco.org/shs/most/gsdl/cgi-bin/library?c=most&a=p&p=about
- Chopin's early works, http://chopin.lib.uchicago.edu/
- Books from the past, http://www.booksfromthepast.org/
- Washington Research Library Consortium Special Collections, http://www.aladin.wrlc.org/dl/
- Papers Past, http://paperspast.natlib.govt.nz/cgi-bin/paperspast

Some of the main features of Greenstone are [5]:

- It works on Windows, Unix/Linux and Mac OS X platforms,
- interoperability through Open Archives Protocol for Metadata Harvesting,
- collections are accessed using a web-based interface,
- graphical user interface for collection building,
- support for various metadata sets: Dublin Core, MARC, CDS/ISIS, etc. and it is possible to define new metadata sets,
- support for various document types: PDF, PostScript, MS Word, RTF, HTML, Excel, etc. images in nearly all known formats: TIFF, JPEG, GIF, PNG, etc, as well as audio files in formats: MP3, Ogg Vorbis, MIDI, etc.
- extensible through a mechanism of *plug-ins.*
- support for multilanguage user interfaces,
- support for Unicode and multilanguage collections,
- collections can be published on the web or CD-ROM.

4. Implementation details

The material was scanned at the National and University Library of the Republic of Srpska using a single Epson GT15000 scanner. Scanning was done at 600ppi, and images were captured as black and white. Thus obtained master images are archived as TIFF images with CCITT Fax 4 compression. Master images occupy about 150-250KB,

[2] http://www.nzdl.org

each. Scanned images are degraded by noise which is due to the bad condition of the material, for example, discolored paper due to age or dirt or speckle noise. It is also impossible to achieve perfect alignment of pages during the process of scanning. As a result some images contain pages with skewed text. The processing of pages in order to remove the degradations is done using filters available in ABBY FineReader, as well as filters we implemented in MATLAB for this purpose. After filtering the images are still in TIFF format and high resolution. In order to publish them on the web it is necessary to convert them into a format which is more often used in the web environment. In this project we decided to convert images into GIF format and downsample them to the resolution of 72dpi which is enough for viewing images on the screen. We created two groups of pictures, the first one with small size images which are used for browsing the material, and the second one with large size images which are used for detailed display. Large size images occupy about 150-250KB. Although with lower resolution than master images they obviously still occupy the same amount of memory. The reason for this is that the LZW algorithm for compression, used for GIF images, is not suitable for black and white pictures as the CCITT Fax 4 algorithm used with TIFF format.

From the original TIFF files we also created PDF documents for every issue of both magazines. In this way it is possible for a user to download, browse and print complete issues on a local computer. For building PDF documents from TIFF files we used GNU libtiff tools[3].

As mentioned before, we adopted Greenstone for managing digital collections. This system supports building hierarchical collections of scanned documents, which is exactly the case in our project. The collection building process is controlled by means of *.item* files which are used for specifying the structure of individual documents. The format of these files is basically XML. Each *.item* file corresponds to one scanned document and it contains one PagedDocument element. In our case one scanned document is one issue of the magazine. Each PagedDocument element contains one or more PageGroup elements corresponding to the groups of pages which make logical units of the document. In our case PageGroup elements contain individual articles or other units, such as tables of contents or appendices, for example. Individual pages are specified using Page tags. Attribute *imgfile* of the Page tag points to the file which contains the scanned image. A part of an *.item* file is shown in Figure 1.

```
<PagedDocument>
      <Metadata name="Series">Школски вјесник
1894</Metadata>
      <Metadata name="Volume">1</Metadata>
      <Metadata name="Number">01</Metadata>
      <Metadata
name="pdffile">SV_SL_1894/SV_SL_1894_01.pdf</Metadata>
      <PageGroup>
            <Metadata name="dc.Title">Читаоцима</Metadata>
            <Metadata name="dc.Title">Čitaocima</Metadata>
            <Metadata name="dc.Title">Citaocima</Metadata>
            <Page pagenum="1"
imgfile="SV_SL_1894/SV_SL_1894_01/0000a.gif"/>
            <Page pagenum="2"
imgfile="SV_SL_1894/SV_SL_1894_01/0000b.gif"/>
      </PageGroup>
```

[3] http://www.libtiff.org

```
      <PageGroup>
            <Metadata name="dc.Title">Школске прилике у
Босни и Херцеговини од окупације до данас</Metadata>
            <Metadata name="dc.Title">Školske prilike u
Bosni i Hercegovini od okupacije do danas</Metadata>
            <Metadata name="dc.Title">Skolske prilike u
Bosni i Hercegovini od okupacije do danas</Metadata>
            <Metadata name="dc.Creator">ДЛУСТУШ
Љубоје</Metadata>
            <Metadata name="dc.Contributor">DLUSTUŠ
LJuboje</Metadata>
            <Metadata name="dc.Contributor">DLUSTUS
LJuboje</Metadata>
            <Page pagenum="1"
imgfile="SV_SL_1894/SV_SL_1894_01/0001.gif"/>
            <Page pagenum="2"
imgfile="SV_SL_1894/SV_SL_1894_01/0002.gif"/>
            <Page pagenum="3"
imgfile="SV_SL_1894/SV_SL_1894_01/0003.gif"/>
            <Page pagenum="4"
imgfile="SV_SL_1894/SV_SL_1894_01/0004.gif"/>
      </PageGroup>
</PagedDocument>
```

Figure 1. Example of the structure of *.item* file.

Besides specifying document structure, *.item* files are used for assigning metadata to the scanned documents and to their parts. We assigned Dublin Core metadata to each article. The metadata were automatically generated from the bibliographical database of the National and University Library of the Republic of Srpska. All articles from the digitized magazines are cataloged and entered into the database. Bibliographical database of the National and University Library is based on Winisis system[4] and in order to create *.item* files it was necessary to perform a data conversion. Records in the database are encoded using modified Windows-1250 character encoding used in Winisis installations in Serbia and Republic of Srpska so, during the conversion, it was necessary to change the character encoding. We wrote a Perl script using Biblio::Isis module[5] to access Winisis database, change the character encoding to UTF-8 and write *.item* files.

In order to enable searching the collection independently of the script used for entering the query, we decided to repeat each Dublin Core metadata element three times using: Cyrillic, Latin and ASCII character sets. In this way all three versions of metadata elements are included into index. However, this approach resulted in some problems during the creation of the index of authors, where all three versions of a name were included. We solved this problem in the following way. Only the first appearance of the name (in the original language) is entered as dc.Creator metadata, and other versions of the name are entered as dc.Contributor. Besides these metadata to each issue of the magazine we assigned global metadata: *Series*, *Volume* and *Number* which contain the title of the magazine, volume and issue number. Finally, a global attribute *pdffile* whose value is the path to the PDF version of the magazine is also added to metadata. It was

[4] http://www.unesco.org/webworld/isis
[5] http://www.rot13.org/~dpavlin/biblio_isis.html

necessary to explicitely specify this path because Greenstone at the moment (version 2.80) does not have the ability to maintain alternative versions of scanned documents.
After the collection was built, the next step was the development of the user interface for browsing and searching the collections. User interfaces to Greenstone digital collections are web based and in the version of Greenstone for Windows family of operating systems there is a small embedded web server, and all versions of Greenstone support integration with well known web servers such as Apache or Internet Information Server. By using web interface it is possible to access Greenstone collections from a wide variety of web enabled devices and platforms.

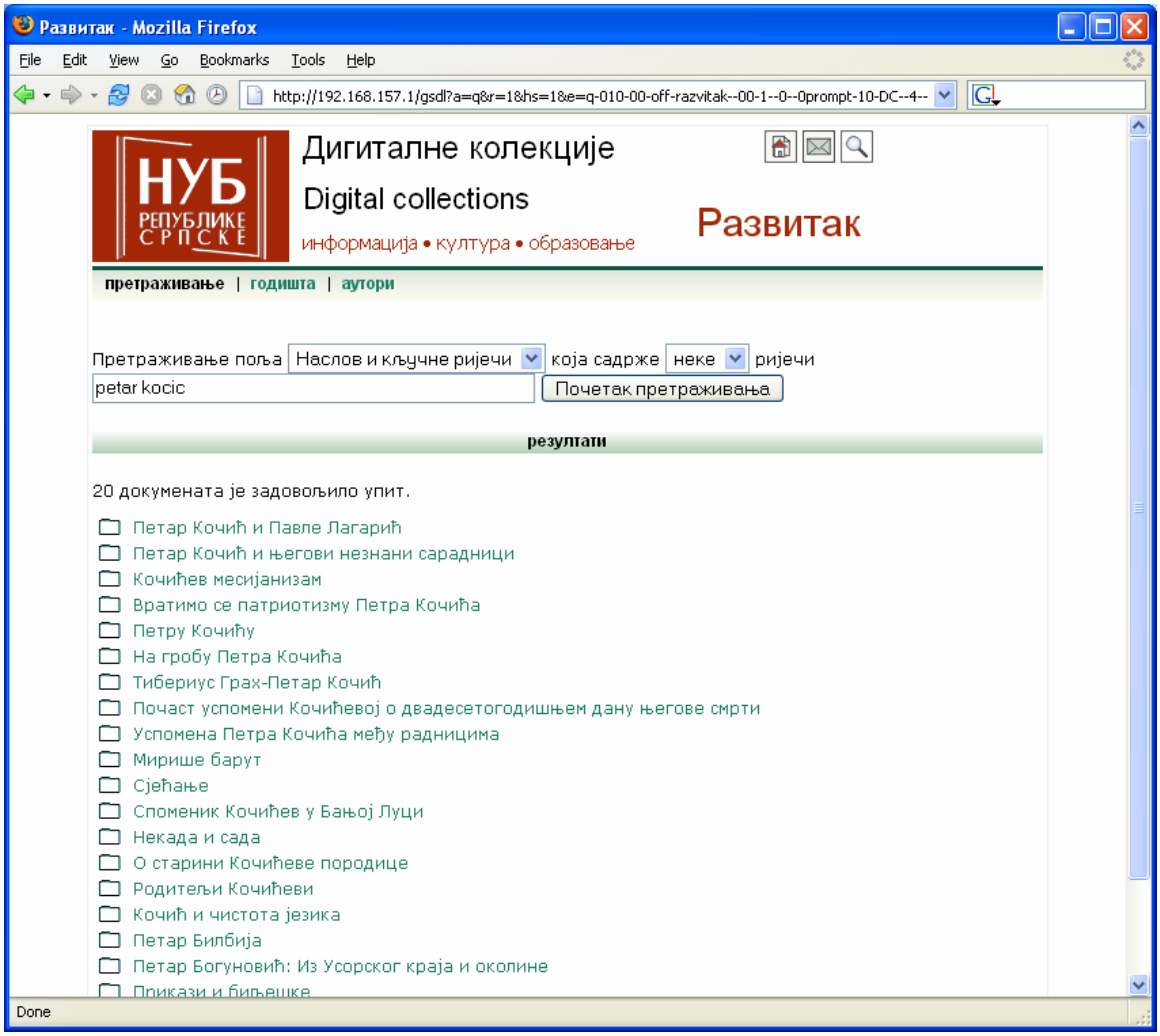

Figure 2. Example of searching the collection using title and keywords

Interface can be adjusted to the needs of users using both the macro language of Greenstone and standard web technologies such as Cascading Style Sheets (CSS) and Javascript. It is possible to change the look and feel of the whole digital library as well as individual collections. We designed and implemented the user interface for our digital collections to blend with the design and layout of the website of the National and University Library.

Collections can be accessed in two ways. First, it is possible to search the collection using supplied metadata – titles and keywords, as well as authors' names.

Queries can be formulated using Cyrillic, Latin and ASCII encoding, and the results will be the same. For example queries „Петар Кочић“, „Petar Kočić“ i „Petar Kocic“ are all equivalent and the results are the same. In Figure 2. is given an example of searching one of our digital collections.

The second way of accessing the collection is browsing of the hierarchical organization of magazines by volumes and issue numbers. Each issue is also hierarchicaly organized and consists of a number of articles. When an article is previewed, there are links to individual pages on the left, and scanned page on the right, Figure 3. Scanned pages can be magnified for easier reading, Figure 4. It is also possible to download the complete issue as a single PDF file.

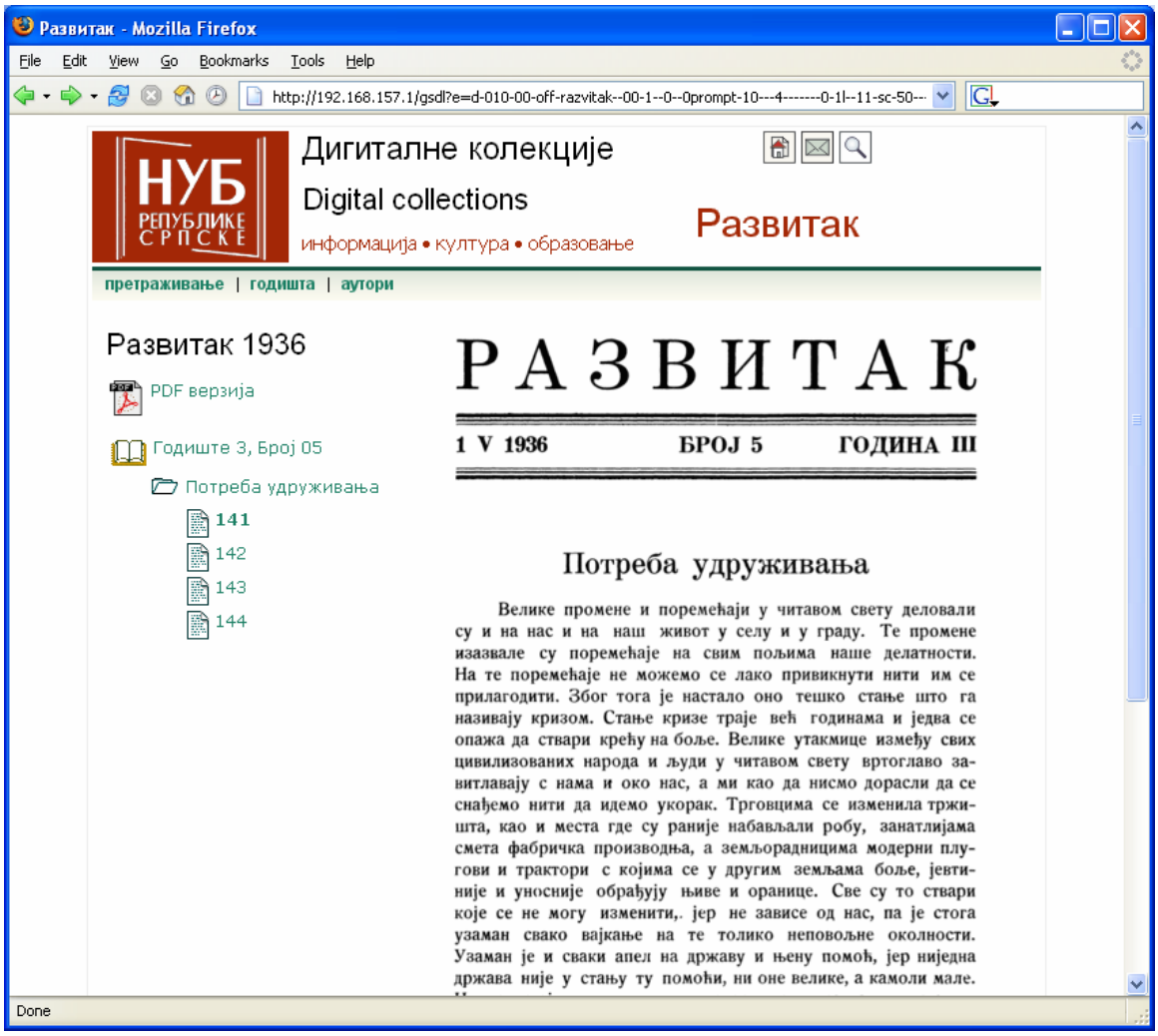

Figure 3. Part of the user interface for page preview

Finally, the third way to browse the collections is by using the alphabetical index of authors. The names in this index are taken at face value from the catalog and some are included in Latin and some in the Cyrillic version.

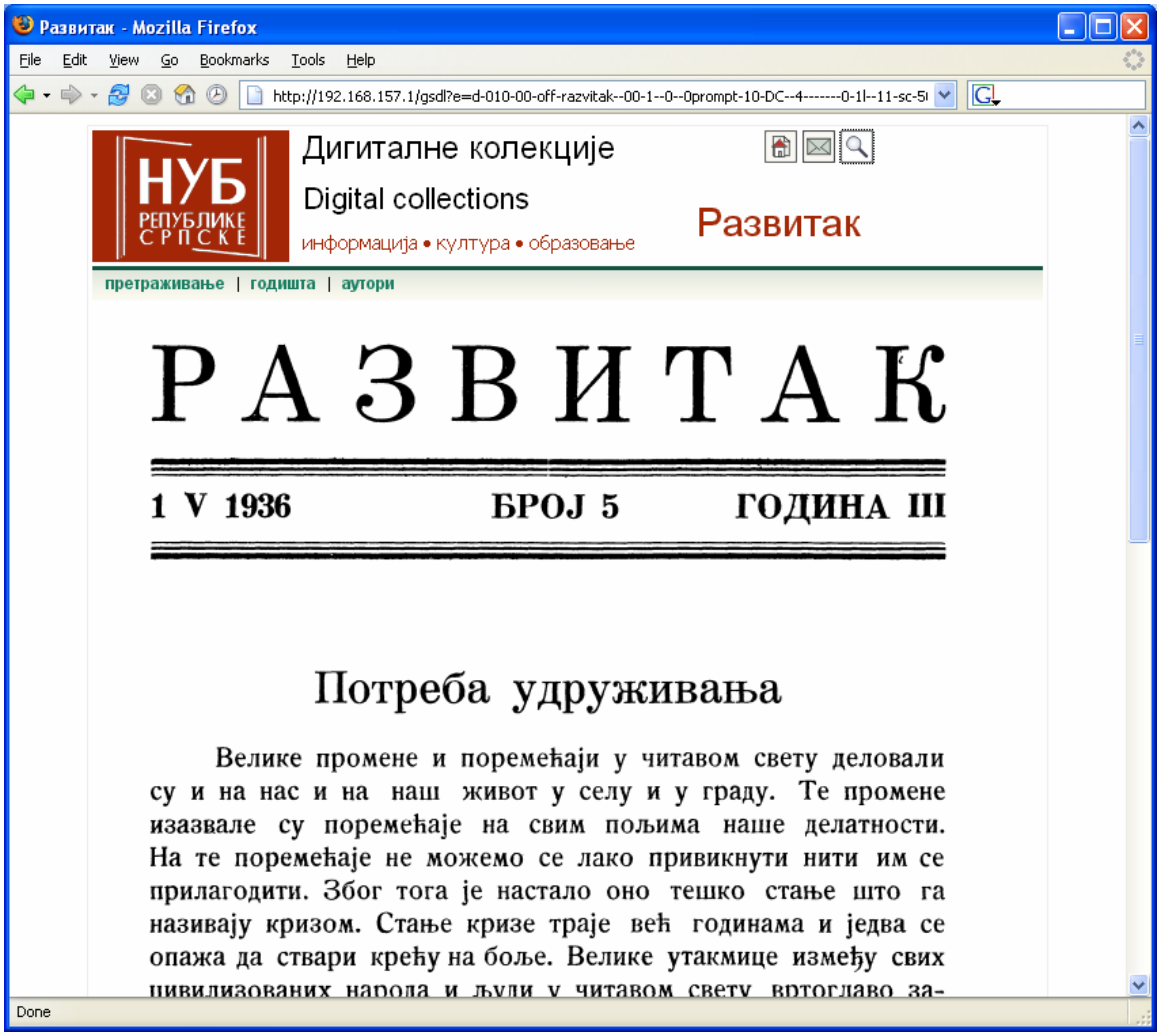

Figure 4. Magnified page view

5. Conclusion and future work

In this paper the first two digital collections of the National and University Library of the Republic of Srpska are presented. These collections contain two magazines published in late 19th and early 20th centuries. For building and managing collections we adopted the freely available Greenstone Digial Library software which is supported by UNESCO. Collections consist of scanned magazine pages with the addition of metadata in Dublin Core format. Collections completed so far are published at the Library webiste http://www.nubrs.rs.ba/dl

In the future work we intend to modify the Greenstone system in order to enable internal change of the character encoding and eliminate the need for multiple repeating of Dublin Core metadata and thus departing from the standard. We will also begin OCR of scanned pages which will make it possible to index and search full text of the magazines. Finally, we will continue to publish new digital collections on the web.

Acknowledgments. This work was supported by UNESCO through project 375415 03BiH: Digitization of books and periodicals from 1878 to 1941, and by Academic and research network of the Republic of Srpska.

References

[1] Ian H. Witten, David Bainbridge, *How to build a digital library,* Morgan Kaufmann Publishers, 2003.
[2] The Dublin Core Metadata Initiative, http://dublincore.org/ (Visited: March 6th, 2008).
[3] Ranko Risojević, Note about magazine „Razvitak“, http://www.nubrs.rs.ba/dl/ (Visited: March 6th, 2008) (in Serbian).
[4] Todor Kruševac, *Magazines in Bosnia and Herzegovina in the 19th century*, Veselin Masleša, Sarajevo, 1978. (in Serbian),
[5] Greenstone Digital Library Software Factsheet, http://www.greenstone.org/factsheet. (Visited: March 6th, 2008).

Владимир Рисојевић
(Електротехнички факултет, Бањалука,
Република Српска, Босна и Херцеговина)

Далибор Панчић, Бојана Милошевић
Ранко Рисојевић
(Народна и универзитетска библиотека Републике Српске,
Бањалука, Република Српска, Босна и Херцеговина)

ПРОЈЕКТИ ДИГИТАЛИЗАЦИЈЕ У НАРОДНОЈ И УНИВЕРЗИТЕТСКОЈ БИБЛИОТЕЦИ РЕПУБЛИКЕ СРПСКЕ

Сажетак. Приказана су два пројекта дигитализације у Народној и универзитетској библиотеци Републике Српске. Ради се о дигитализацијама часописа “Развитак” и “Школски вјесник”. Описан је материјал као и процедуре скенирања и креирања метаподатака. Укратко је описан Greenstone, софтвер за креирање и дистрибуцију дигиталних колекција. Такође су описани и дизајн и имплементација интерфејса заснованог на вебу за приступ колекцијама. Колекције које смо креирали се могу претраживати по насловима, кључни ријечима те именима аутора, а могу се и хијерархијски прегледати по годиштима и бројевима или индексу аутора.

Кључне ријечи: Дигитализација, културно наслијеђе, Greenstone.

vlado@etfbl.net