Преглед НЦД 17 (2010), 45–49

Marija Bogdanović
National Library of Serbia
Nenad Jeremić
Blink N Agency, Belgrade

DATABASE OF THE SERBIAN RETROSPECTIVE BIBLIOGRAPHY OF BOOKS: 1868–1944

Abstract: The project Serbian retrospective bibliography of books for the period from 1868 to 1944 has been run by the National Library of Serbia from 1954 until today. Holdings of over 100,000 bibliographic items published in 20 volumes contain information about each book that was published in that period. The idea of digitalization of materials i.e. conversion to electronic form and creating a program that will allow to search material by larger number of parameters, originated from the Bibliography department of the National Library of Serbia, which is the main carrier of this project since 2002. The whole idea was to create a database serving primarily as a program to enter and edit the author and subject heading of the printed edition of a bibliography, and also to make an export of data to a user request, as well as to resolve links from the bibliography units to the scanned images of the printed Bibliography on the server. Software solution includes the MySQL database and PHP scripting language to communicate with the database. Linking scanned printed material and the possibility of online searches are the basic qualities of a software solution that is available online [1].

Introduction

Project development for the Serbian retrospective bibliography of books [2] for the period from 1868 to 1944 is very long and difficult, but the importance of identifying and presenting literary production of the Serbian people is incomparable with all the barriers that were faced and still faces its construction. The first steps in the preparation of Bibliography were made in 1954 on behalf of the National Library of Serbia. Holdings of more than 100,000 bibliographic items were collected during research process in all the major libraries of the former state, but also in major foreign libraries possessing materials related to the Serbian people. To this date as initially planned 20 volumes of published materials have been finished. Also, 21st volume which will contain the supplement of the voluminous material collected over the years successively is in the process of preparation. Selection of materials is carried out according to the principle of territorial, national and linguistic criteria according to bibliographic description rules of Stojan Novaković. The units are ordered alphabetically by headings written in Latin, while the bibliographic descriptions are written in language of publications. For the complete bibliographic materials, the author and subject index were completed, and it is planned to create the following cumulative indexes: author, subject, printer, publisher, title, subtitles, pseudonyms and codes and chronological one.

Chief editor of Serbian Bibliography: Books 1868–1944 was msc Miodrag Živanov. The first president of the Editorial Board was academician Radovan Samardžić, and member's and reviewer's board consisted of great number of eminent scientists from many of our cultural institutions. The current president is academician Miroslav Pantić.

As the project was bringing to the end, the circumstances were changing as well as the user's need for this kind of materials. The information that is readily available has become imperative for the modern user. In the era of digitalization when the information has to be quick, clear and detailed, with the ability for searching and linking the information that goes beyond the printed form, it appears that this type of bibliographic material should be made more accessible because of its great value.

Beginnings in the digitalization

The idea of digitalization of materials i.e. conversion to electronic form and creating a program that will enable (re)searching of material over a larger number of parameters, originated in the Bibliography department of the National Library of Serbia, which is the main carrier of this project since 2002. In that year 15 volumes of bibliography were published, and one of the further goals was to create a software solution for facilities of job of making the author and subject indexes for remaining volumes of the Bibliography. Soon it was clear that the cumulative production of indexes will be a long and difficult process during which much information would remain inaccessible for scientists involved in studying these types of materials.

So, the Bibliography department was requested for assistance from experts employed in the National Library Computer Center about making the bibliographic software for the Serbian bibliography which had to be grounded on:

- programming system for conversion and bibliographical entry in the selected bibliography format (which format to use pre-press)
- programming system for the indexes formation
- programming system for the publication of bibliographies in electronic form (CD and database searchable over internet).

Such bibliographic format was supposed to meet all the rules of the existing bibliographic description and serves as a matrix for entry of new data in the volumes not printed yet and then in making all the indexes. This form of bibliographic database converted from printed material containing all the elements of bibliographic processing, in addition to programs that allow the creation of a register of holdings searchable online was too much demanding for the resources of the National Library.

The next step of the Bibliography department was made in 2006, when 17 volumes of printed material had been already finished, asking for something more modest. At this point, 16 volumes of Bibliography were scanned in the frame of the National Library digitalization project. The files were set up (1 file = 1 page) with capabilities for browsing and retrieval of content by the number of volumes and the first letter of last name of the author in the author heading. At the meeting of the Working Group for software development and creating of cumulative indexes for the Serbian bibliography it was estimated that the design of a new database (especially for the retrospective bibliography) would be an unreasonable project and that any future project should focus on making the cumulative indexes for authors and subject headings and the link between bibliographic items in these registers with the scanned pages in which were the units. It was decided that the programmer should create a database that will be using primarily for writing and editing the author and subject indexes. That program also would allow export of data on demand, as well as to resolve connectivity issues with the unit in the registers of scanned images on the server. It was concluded that the user interface for viewing the entries with links to records of scanned images was sufficient.

Software Solution

Due to the fact that complete OCRing for the author index for all volumes had been already done further steps were taken when programmer Nenad Jeremić was tasked to develop the software. The task that was supposed to be carried is as follows:

1. Connecting the author index files in one with the OCR-ed material,

2. Linking the numbers of Bibliography units from the index to the digitalized material, i.e. images that contain the bibliographic units. Input file would be one that contains the pairs: the number of bibliographic unit on the first page and name / address of a scanned image of the page.

3. Facilitate entry of new and existing entries in the editorial database.

4. Enabling the printed output of author index from the database for each volume.

5. Adjusting the database for the web with the ability to view (browse only) page via links from the register.

Software solution includes the MySQL database and PHP scripting language to communicate with the database.

OCR scanned material

Initial test with a few pages of author and subject index, showed a large amount of errors due to poor conversion material using OCR (bad printing; older versions of the OCR program). According to the programmer's proposal the main input of database was done in this manner after which was possible to create and export data to Excel file ruled as a main format for editing in the Bibliography department. Firstly, is was done with author index, so that each volume was inserted into the base after the OCRing, then exported to Excel, where the bibliographers worked on correction and partial editorial work, and then returned to the database. The whole idea was based on the ground that the subsequent corrections should be done directly in the database if needed. But the programmer met various problems with OCR, thus making the whole process slowed down. Since the OCR program generates the same characters in different ways, the same pattern was transferred in the same way into the database for searching. Therefore, completely new database has been created, in which the whole account of numeric codes of all possible variants of the letters that appear in the author index were given as searchable.

These codes were used to encode certain textual fields in the database. It is possible that different units in which they were given various terms and names remain in their home fields and in this way and present to the user, but for the purpose of search and sort using their code. Although during this process the amount of data in the database was significantly increased, but the speed and all other performance results and their treatment were given as much share as well.

Thus, problem of mixed characters due to the aforementioned malfunctions of OCR program was successfully solved creating the possibility of proper alphabetization of the search results and the entire database, too.

Editing the database

After months of shared programmer's and bibliographers' work resulting in completing the software for the author index and their accumulation, the entire database was installed on the web server. At this point, except the author's index of 8th volume (due to the fact that there was found too much errors in the design of the index demanding revised editorial work), the database of author index has been completed in full format. In order to speed up the process and reduce errors, it was proposed to the programmer to solve the automated redaction of indexes in the database based on the possibility for final editorial work directly in the database.

After all at present state two ways of editing were allowed through the software. One way based on exporting or importing the Excel file from or in database programmed in special Pearl script where one part of database entries was replaced by the Excel data. This method is most suitable in cases when we need to replace a larger amount of data in the database. Another way is online editing allowing work directly in the database that it is suitable mostly for minor changes of data. The testing of both ways for editing has been done before using.

Editorial work for all author and subject indexes in the Bibliography department is in progress. After its competition, the conditions for printing the cumulative register will be created, what actually was a starting idea of this project.

Searching the database

After completed work on the 20th volume of author index and its accumulation in the existing database, the instruction's for searching database was written and the online database was linked from the website of the National library. Online database for users as well as the basic information about project of Bibliography is available on the Internet [1], and the search are running through the search engines defined in accordance with the requirements of the database creation.

The database search engine provides a number of possibilities and presents the main tool for working with the database to search as well as to editing. Results are available directly in the Internet browser, but also in the more convenient text format in which the data are grouped by name and number of entries. Data entered in this way in the database can be searched on the several parameters: the last name, the first name and the number of the bibliographic unit for author index; also, for subject index on: the subject heading, sub-heading and used for headings. Number of volume gives us the full listing of author or subject index of that volume and can be used along with other parameters. Print results gives us the author or subject headings, the numbers of volumes through which they are linked to the scanned pages of the printed Bibliography which contains the number of the bibliographic unit. The digitalized scanned material can be browsed with forward or backward arrows.

Extended search allows automatically the small Latin letters c, s, d, z, to be recognized as č, ć, š, dž. Exceptions are the last names and names with a hyphen, where it is necessary to enter the author's name as it appears in the registry. The searching is not case-sensitive. The peculiarity of the author index present codes and pseudonyms that need to be entered in the existing form. The option "*contains*" could be checked for each parameter providing the broadest search options. By entering only part of the words in search box, we will get all the entries including given part of the entered word.

Conclusion

With setting the database on the website of the National Library through education of the NBS employees it has been shown that the interest is huge including extremely positive feedback. In that manner the usefulness of the database was further justified. General characteristics of the database for users are simple and easy searching and viewing of the data, with ability to print and group the information. As for the editing of the database, it has been proved to be accelerated because the data from all volumes could be grouped and sorted, with ability of an online editing allowing easy and fast access to data. The process of such a large-scale digitalization of printed material with creating a database of the registers that are using OCR transferred to the base and then revised, with linking the scanned pages and databases across a number of bibliographic resources with ability of researching, makes this project the more important in the field of digitalization of our cultural heritage.

References

[1] Online database for users of Bibliography, http://www.nb.rs/pages/article.php?id=1381

[2] Serbian bibliography: books: 1868–1944. (in Serbian) / [chief editor Miodrag Živanov]. - Beograd: Narodna biblioteka Srbije, 1989–2008