Преглед НЦД 5 (2004), 42–48

Žarko Mijajlović (Matematički fakultet, Beograd)
Zoran Ognjanović (Matematički institut SANU, Beograd)
Nada Đorđević (Matematički fakultet, Beograd)
Tijana Zečević (Matematički fakultet, Beograd)

VIRTUAL LIBRARY – DATA BASE OF TEXTUAL DATA

Abstract. Objectives of the project which concerns electronic archiving of old manuscripts and books and their publishing in electronic phototype form are discussed in this paper.

Key words. Digitization, scientific heritage, cultural heritage, virtual library

1. Introduction

The overall aim for the digital library is to create a comprehensive and semantically interconnected collection of retro-digitized books and other digital documents. Objectives of the project include electronic archiving of old manuscripts and books, their publishing in electronic phototype form and finally presentation to the general audience. Our proposal is mathematically inclined, more specifically, the project is concerned with digitization of old books and other manuscripts satisfying the following criteria:

1. Books and other manuscripts contained in Serbian public and semi-open libraries. By a semi-open library we mean any departmental library and private collections of books.
2. Books and manuscripts selected for digitization should be related with mathematical sciences: mathematics, mechanics, astronomy, physics and mathematical geography.
3. Books considered for digitization had to be published before certain date in the past. We have chosen for this date the beginning of the Second World War (1941).
4. Preference will be given to Serbian authors, or to written works related in some way to Balkan area.

There are relatively large collections of rare books in our libraries. For example, the library of the Mathematical Faculty in Belgrade has about 500 mathematical books published before and during the 19. Century. On the other hand, just a few of them can be found in joint catalogues of largest Serbian libraries network (this includes the National Library and all university libraries, about 16 all together). For some of these

books it is known that only a few printed form copies are left. We think that it is very important to preserve them in some form, not only as a cultural or scientific heritage important for our local community, but also as a part of the World heritage. Therefore, we decided to preserve and present them to the wide audience in electronic - digitized form.

This project gathers about ten collaborators and volunteers including students, mainly from the Faculty of Mathematics and the Mathematical Institute in Belgrade. The group has for this purpose on its disposal very modest equipment, and aims of the project might seem as a hardly reachable aspire. However, we think that someone has to start this task. And we started it.

2. Technological aspects of the project

Digitization is heavily if not completely based on information technologies. But information technologies are one of the fastest changing parts of the contemporary world. For example, since the appearance of disk and tape based digital storage in the 1960s, we have witnessing evolution and proliferation of more than 200 different storage formats. Therefore libraries and information archives face a continuing challenge in maintaining files on currently supported storage hardware and media in currently supported file formats, but operating systems as well [1].

We had in mind all these issues when we placed in focus the following aspects of the project:

1. Rules for choosing preferential books and manuscripts to be digitized.
2. The mathematics of information storage (database architecture, data manipulation, data forms, etc.).
3. Information retrieval.
4. Metadata and rules for handling metadata.
5. Distribution of informations to the wide audience using Internet.
6. The ability to search electronic files efficiently and retrieve information quickly.
7. The ability to reuse information in other documents and other formats.
8. Fast retrieving of digitized material through electronic transactions and data downloading.
9. The proper choice of data format and resolution.

There are many proposed standards for the database architecture, distributions of information and information retrieval. Common practice in Database Projecting is usage of so called Data Base Management Systems (DBMS) and especially Relational DBMS (e.g. well known is – ORDBMS for example, Oracle RDBMS). These systems include variety of tools which allow users to create, update, and extract information from their databases. However, some procedures for data description that are independent from hardware and software-platforms were proposed more or less recently.

We also note that metadata for digitized documents may differ significantly from data for printed books obeying some of the existent standards (e.g. "Unimarc" [5]). For example, metadata for a digitized document should include the following items, ordinary not found in the mentioned standards: name of the editor (person and/or committee) who made decision for digitization, then resolution, file type, revision number and date, etc. Some of the new initiatives, like TEI and MASTER [3, 8], can

help us to obtain very expressive description methods. Also, the Dublin core metadata initiative [6, 7] concerns developing metadata standards, community- or disciplinary-specific metadata sets and frameworks for the interoperation of metadata sets.

3. Some technical details of digitization

In our approach we are digitizing documents basically in four steps:

1. Scanning document pages
2. Processing scanned pages
3. Storing processed pages in appropriate file format
4. Collecting processed pages into an electronic document.

The future use of the digitized document determines how these steps will be performed.

3.1. Scanning. In most cases we used 300 dpi resolution and gray-scale mode for the best quality/size trade off. We found that acceptable resolution range is 200-600dpi, particularly if one wants to perform OCR on the scanned document. Even if the mentioned facts might sound trivial they may be useful, as the following example show. One very recognized institution in Germany performed digitization in so called screen resolution (96 dpi). After some time they discovered that so scanned printed works are of small value. So when they turned to higher resolution, they decided to repeat the scanning of all documents.

3.2. Processing scanned images. Due to imperfections of the scanner, dust and document state, some improvements should be made to scanned images: impurities cleaning, contrast improving, right positioning, setting the same size for all pages, etc. We use Adobe Photoshop and various embedded filters in this software in order to increase the quality of the scanned pages.

3.3. Storing scanned images in an appropriate file format. We are always archiving our master copies of scanned images in tiff format because it is the widley used and supported standard. But the main advantage of this format is the high images quality. Using tiff is not a question, only some options were taken in mind when we faced photo images because tiff photo images are pretty large. In that case we are using tiff with LZW compression that is more effective for gray-scale images than colour (as we do our scanning in gray-scale mode), but it is lossless – which means there is no quality loss due to a compression. Also, tiff can be easily converted into other file formats and the choice depends on what we are intending to do further. Usually it is pdf because we are mostly archiving the old books.

3.4. Collecting processed pages into an electronic document. Scanned pages are automatically stored into an electronic document by appropriate software. We use Adobe Capture software and PDF (Portable Document Format) file format. Very often and whenever is appropriate, we are performing OCR. In this case the final digital document consists of two layers. The first layer consists of the exact image of the original document, while the second layer consists of the recognized texts. The second layer enables one to search through the document for certain contents. In some cases

further ramifications and structuring are done on documents: introducing hyperlinks on the content and the index of the original, embedding annotations, comments etc.

4. Similar projects in the World

First of all we are going to mention one historicaly significant person – Vannevar Bush (1890–1974) [10] from the Massachusetts Institute of Technology (MIT), who was the first one that introduced the idea of electronic publishing in his seminal 1945 article *As We May Think* [11]. His conception of the Memex, was the idea of an easily accessible, individually configurable storehouse of knowledge. Douglas Engelbart and Ted Nelson, men who developed the hypertext mechanism, were directly inspired by his work, particularly by his article. We found that was an interesting fact unknown to the general public so it took place in our article.

At the moment there are many projects in the world that are concerned in digitization of different materials. We were influenced by their ideas, but we are going to mention maybe the first one (Project Gutenberg) and an initiative in cooperating different digitizing projects – Digital Mathematics Library (DML) that we found particularly interesting.

Project Gutenberg [4] was started by Michael Hart in 1971. and for nowadays it is the Internet's oldest producer of free electronic books. Those books are distributed mostly in the **plain text** format, nevertheless there is an insistence on plain text. Main submitters are volunteers from all over the world and the present collection is more than 13 000 books in number. The only limiting factor in redistribution of all Project Gutenberg eBooks is national copyright laws and therefore it is possible that some eBooks which are public domain in the U.S. are still under copyright protection in other countries. There are three portions of the Project Gutenberg Library, basically be described as:

- light literature, such as Alice in Wonderland, and Aesop's Fables,
- heavy literature, such as the Bible and Shakespeare, and
- references, such as almanacs, and a set of encyclopaedia, dictionaries, etc.

Also, there are many projects related in digitization of mathematics materials and those are mostly connected with different national initiatives (like this project, that is correlated with National Center for Digitization in Serbia). The great problem is lack of standards in that area and therefore there is an effort for ***Digital Mathematics Library*** [2, 9] which is the project coordinated by Cornell University Library and funded by the U.S. National Science Foundation (NSF) toward the establishment of a comprehensive, international, distributed collection of digital information and published knowledge in mathematics.

The following words of John Ewing, executive director of American Mathematical Society probably are best describing the true aim of this project, but some dilemmas as well:

> *Mathematicians have talked quietly for some time about the need to digitize the past mathematical literature. During 2001, the conversation became more intense as several new digitizing projects were announced. Should we coordinate those projects? Could we integrate the recent literature that is*

already in digital form? How we could digitize far greater amounts of older material? The goal to create a virtual library containing much of the past literature - a library that could eventually grow into "a World Mathematics Library".

At first, DML was a one year planning project (2002-2003), but the NSF has extended the grant period for the original DML to October 31, 2004. In the future the project is going to be under the leadership of International Mathematics Union (IMU) and it is going to be a support for additional planning of continued interaction among different digitization projects. The DML is intending to be a World Digital Mathematics Library (WDML).

5. NCD Virtual Library

Serbian national initiative NCD (National Center for Digitization) gathers about ten different national institutions (museums, libraries, institutes…) which all has the same goal of digitizing national and cultural heritage. This project is under the leadership of NCD and it is directed by Žarko Mijajlović. Other principal investigators are Zoran Ognjanović, Dragan Blagojević and Vesna Vučković.

Our intention is to project a database which is going to include books, dissertations, articles etc, and to create a web interface for administering and searching the database. Searching criteria could be: the author, the title, the time period the material was printed in, the topic the material is dealing with and keywords. The problem we are facing with is that there are no standards in descriptions of those books like ISBN and ISSN numbers so classifying them is rather hard.

We have chosen SQL server 2000 for a database. JAVA programming language is used in developing a web application for administering and searching the database, especially JAVA advanced features like java beans and strucs, which enable a high performance web application. Also, Tomcat as an application server and Apache as a web server are used.

We have digitized, by now, more than 100 books and manuscripts in mentioned fields:

1. One of the finest digital collections is the electronic edition of collected works of the prominent Serbian mathematician, Bogdan Gavrilović (1863-1947). This collection includes about 50 items: books and articles (about 2000 printed pages).
2. Mathematical works of Ruđer Bošković (1711-1787).
3. 12 doctoral dissertations of Serbian mathematicians, including all written before the First World War (8).
4. Two books of the famous Serbian scientist (mechanics, mathematician and astronomer) Milutin Milanković, including his celebrated book *Kanon Der Erdbestrahlung* (The Canon of Earth Irradiation).
5. All books of the 19. Century mathematician and astronomer Milan Andonović (on the Probability theory and Astronomy).
6. Books of other Serbian mathematicians from the 19. Century and the beginning of 20. Century: Kosta Stojanović, Dimitrije Nešić and Mijalko Ćirić.

Electronic editions of these retro-digitized works were published since 1995 on several compact disks. In the occasion of this conference they appear for the first time on the Internet in quite simple form (http://alas.matf.bg.ac.yu/biblioteka/home.jsp). It will be available soon in its' intended form at http://virlib.matf.bg.ac.yu.

Acknowledgements

We would like to mention that we had rather helpful advices of our prominent historians of mathematics: Simon Dragović, Dragan Trifunović and Rade Dacić.

The following addendums are published only in electronic forms. They annex the article at http://www.ncd.matf.bg.ac.yu

Addendum 1. *Old Serbian books in logic*
Addendum 2. *Elderly Mathematical books of the Library of the Mathematical Faculty*
Addendum 3. *The collection of digitized books in mathematical sciences*, edited by Žarko Mijajlović
Addendum 4. *Books and periodicals in elementary mathematics in Serbian, 1800–1920,* bibliography compiled by Dragan Trifunović

Bibiliography

[1] Building a national strategy for digital preservation: issues in digital media archiving, a collection of papers, Council on Library and Information Resources, Washington D.C. and Library of Congress, 2002. http://www.digitalpreservation.gov/ [Date of last access: 2004-10-14]

[2] Digital Mathematics Library, http://www.library.cornell.edu/dmlib/ [Date of last access: 2004-10-14]

[3] M. J. Driscoll, The Text Encoding Initiative, to appear in the Review of the National Center for Digitization.

[4] Project Gutenberg, http://promo.net/pg/ [Date of last access: 2004-10-14]

[5] MARC standards, http://www.loc.gov/marc/ [Date of last access: 2004-10-14]

[6] M. Milenković, Dublin Core Metadata Initiative (DCMI), Review of the National Center for Digitization 2, 70 – 79, 2003.
http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/2/d011/document [Last access: 2004-10-14]

[7] The Dublin Core Metadata Initiative, http://dublincore.org/ [Date of last access: 2004-10-14]

[8] Text Encoding Initiative, http://www.tei-c.org/ [Date of last access: 2004-10-14]

[9] Twenty centuries of mathematics: digitizing and disseminating the past mathematical literature, John Ewing, preprint, 2003.

[10] Vannevar Bush, http://www.iath.virginia.edu/elab/hfl0034.htm
http://en.wikipedia.org/wiki/Vannevar_Bush [Date of last access: 2004-10-14]

[11] Vannevar Bush, As We May Think, The Atlantic Monthly, July 1945.
http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush0.shtml [Date of last access: 2004-10-14]

[12] *Lives and works of Serbian scientists*, Series of eight books, ed. Miloje Sarić, Serbian Academy of Science and Art, publ. In the period 1996-2003.

VIRTUALNA BIBLIOTEKA – BAZA TEKSTUALNIH PODATAKA

Sažetak. U radu su predstavljene osnove projekta elektronskog arhiviranja starih knjiga i rukopisa i njihovog predstavljanja u elektronskoj fototipskoj formi.

Ključne reči. Digitalizacija, naučno nasleđe, kulturno nasleđe, vurtualna biblioteka

Žarko Mijajlović zarkom@eunet.yu
Zoran Ognjanović zorano@mi.sanu.ac.yu, http://www.mi.sanu.ac.yu/~zorano
Nada Đorđević nadicadj@matf.bg.ac.yu
Tijana Zečević gemma@matf.bg.ac.yu