Преглед НЦД 22 (2013), 140–147

Marija Šegan
Mathematical Institute of SASA
Natural Sciences and Technologies, University of Belgrade, Serbia
Nikola Petrović
Morena inženjering Niš, Serbia

ARCHIVE OF THE MATHEMATICAL INSTITUTE OF THE SERBIAN ACADEMY OF SCIENCES AND ARTS: DIGITIZATION OF THE REPORTS FROM THE SESSIONS OF THE SCIENTIFIC COUNCIL (1948 – 1964)

Abstract: The Archive of the Mathematical Institute of Serbian Academy of Sciences and Arts (MI SASA) exists from the date of the establishment of the Institute as its integral part, and includes more than a dozen of handwritten volumes (i.e. reports, diaries, and journals), that testifies to its financial, administrative, technical, cultural, scientific and other activities. Although the archive is formally open to the public, , access to information is in practice limited; so there is an initiative to use digitization, preserve the archival material, provide a presentation and open access to the users. Using the example of digitization of the oldest preserved report manuscript from the sessions of the Scientific Council of MI SASA (1948 – 1964), this paper deals with following questions: 1) the standards of digitization of the old manuscripts in the absence of a unified national strategy for digitization of the cultural heritage, and 2) the named entity recognition – automatic extraction and linking of a person, other named entities and keywords from the digitized manuscripts in order to facilitate its effective search. The results of this paper are presented in HTML presentation.

Keywords: MI SANU Archive, cultural heritage, digital preservation, OCR/ICR, NLP, NER

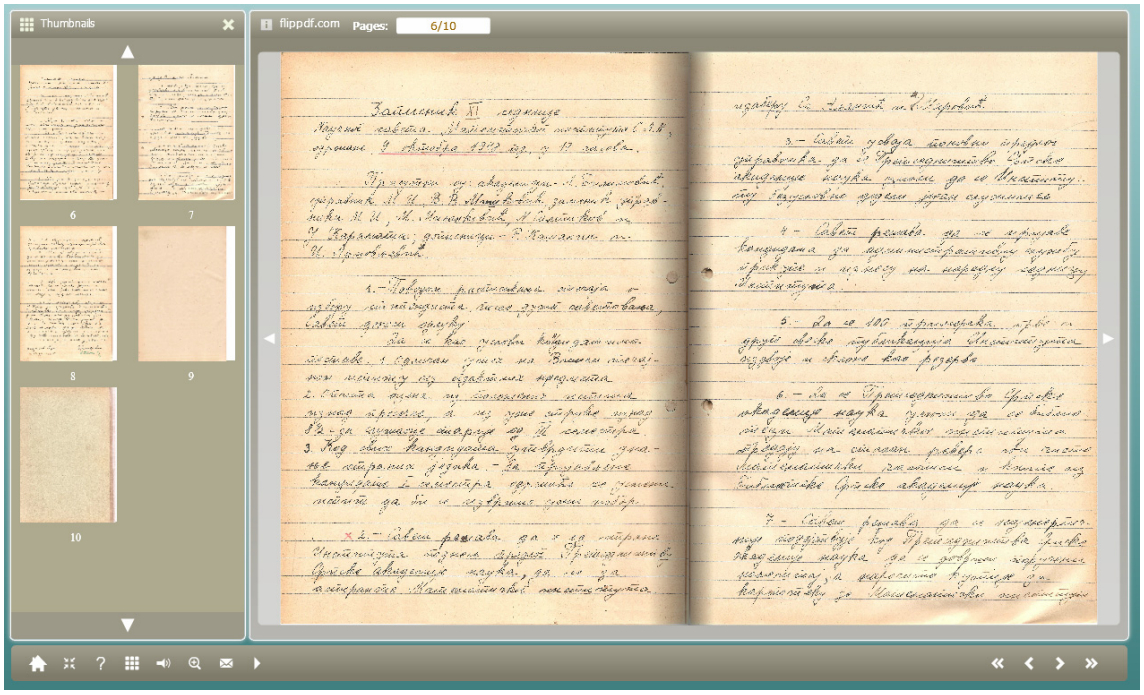
Introduction

The Archive of the Mathematical Institute of Serbian Academy of Sciences and Arts (MI SASA) was created gradually, during the years of functioning of the Institute. Since it is hard to define the exact year of its establishment; most probably it was established the same year as the Institute, thus in 1946. Today it contains about a hundred of volumes and folders of mostly administrative documentation, among others also, the special group of documents of the records of sessions of the Scientific Council of the Institute.

The Scientific Council of the Mathematical Institute which is as old as the Institute itself, represents governing body whose members are distinguished scientific workers and collaborators of the Institute. Its task is to solve important questions in the field of scientific and research work of the Institute. All sessions and the decisions of the Council are documented in the aforementioned records that are preserved in the Archive. These records testify not only about the history of the Institute and its members, but also about social, economic, political and other aspects of the Yugoslavian and Serbian science in the period after World War II. Although its cultural, historical and scientific significance is undisputed, the records are not easily available to the public. There is also only one record for each volume (handwritten or typewritten). For this reason the methods of their possible conservation must be considered. The method of preservation and presentation that is proposed in this work is the method of *digital* preservation and presentation.

The pilot project dedicated to the digital preservation and presentation of the records of the Scientific Council of the Mathematical Institution started with the digitization of the oldest preserved record, which was a handwritten volume (Picture 1). The record covers the period from 1948 until 1954 that includes

64 sessions of the Scientific Council[1]. It is considerably badly preserved, and it differentiates the four handwritings in Cyrillic of the famous national scientists: Jovan Karamata, Milan Vrecko, Radivoj Kasanin and Tatomir Andelic. In the record, the sessions are recorded according to a pattern included: the ordinal number, the date of the session, the names of the participants, the topics that were discussed and the signatures of the recorders and the chairman. The topics were usually related to the scientific and research work of the Institute, like the research of the methods and the results of the scientific work, the suggestions for the publications, the organisation of the seminars, election of persons for the scientific titles, etc.



Picture 1: The flipbook of the scanned original volume of the Reports of the Scientific Council (1948 – 1954)

In this work we present the first results of the digitization that relate to the creation of the working scheme for the digital preservation and presentation, as well as the realisation of the technical part of the process.

1. Creation of the Digitization Scheme

In Serbia, although featured as the priority of the Ministry of Culture, there is no national strategy for digital preservation and presentation of the cultural and scientific heritage. There are some suggestions (like the initiative for the production of the *Rule Book for Digitization* with the goal of establishing the „normative framework for the activities in the field of digitization“), but they are not officially recognized yet. In such circumstances and without any defined strategy, being faced with the digitization of the handwritten volume, we relied on the existing suggestions[2] and earlier experience with the goal of creating our scheme (Picture 2).

1 It is assumed that there are the records made since the day of the establishment of the Institute, 1946. until 1948, but to the authors of this work its whereabouts are unknown

2 Z. Manžuch, I. Huvila, T. Aparac-Jelušić, „Digitization of Cultural Heritage" published in: *European Curriculum Reflections on Library and Information Science Education*, Copenhagen 2005; Stefana Janićijević, *Analiza politika, standarda i menadžmenta u digitalizaciji biblioteka*, Beograd 2008. <http://www.mi.sanu.ac.rs/~stefana/stef.pdf>

This scheme differentiates 4 work packages: 1) Strategic planning (includes the questions of what and why to digitize, how to digitize, for whom to digitize etc.), 2) Development and Implementation (includes the realisation and the use of the strategic plan with the help of technology), 3) Quality control (includes the process of checking and rechecking) and 4) Sustainability (includes the question of the long-term preservation). For now, the first two packages are theoretically determined, while the other two are in the phase of development (Attachment 1).

Picture 2: The scheme of the digitization process

2. The Technical Implementation

The technical part of the pilot project of the digital preservation and presentation of the record included the use of the technology for converting, storage, delivery, presentation and creating the meta-data (Table 1).

Table 1: The Technical Implementation

Technology of Conversion	
Scanning	Scanner: HP DeskJet F370, Software: FreeKapture;
Image processing	Software: GIMP 2.6, JPEGtoPDF
Optical Character Recognition	Automatic – Software: Tesseract – and Manual
Text Editor	Software: OpenOffice3.3
Technology of Storage	
Disk	Adata Superior SH93
Type of storage	Scanned image
Format of storage	TIFF, JPG, PDF
Resolution	600dpi, for online presentation 72dpi
Colours	Million of colours
Technology of Delivery and Presentation	
	Plain text, HTML, Flip Book (PDF to Flip Book, freeware)

Wanting to avoid using licensed software and making the process of digitization financially tolerable, the free software was used where possible. In respect of authors' and owners' rights, the digitized material was not presented on-line completely, but in the frame of a local html presentation (Picture 3). After obtaining the rights of the Mathematical Institute SASA, which is the owner of the digitized material, the plan is to make public presentations of the material in the appropriate on-line database, for instance:
http://www.mi.sanu.ac.rs/main_pages/about.htm.

Picture 3: The HTML presentation of the results of the project

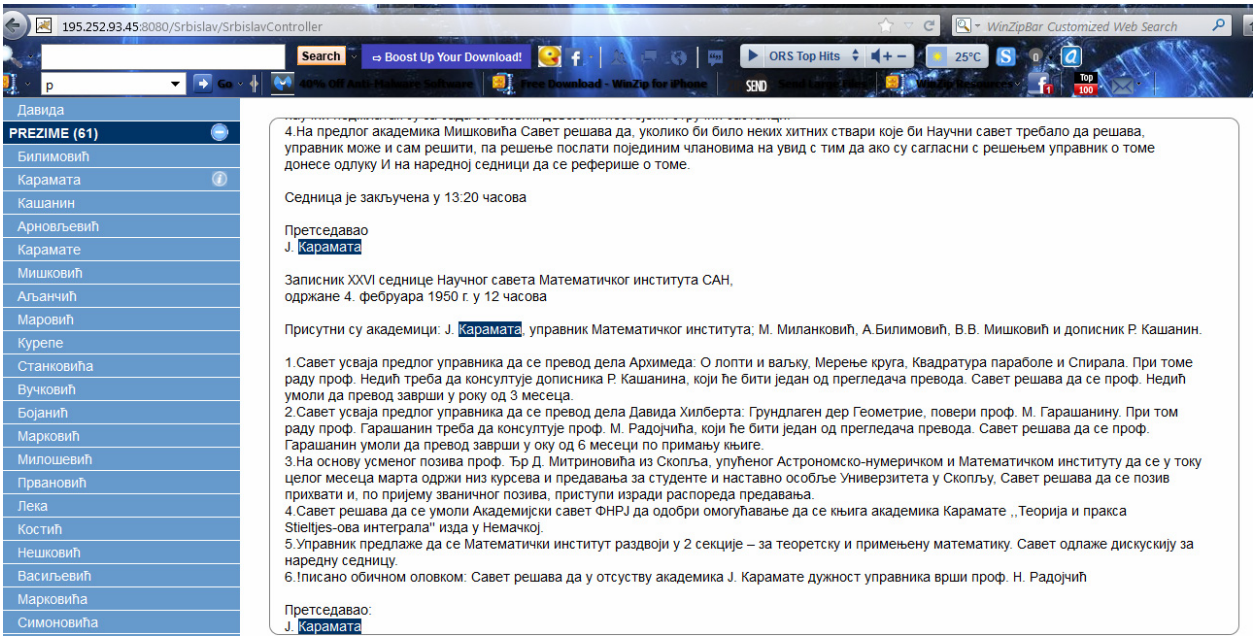
3. The Named Entity Recognition (NER)

While digitization is a necessary first step to fully expose the knowledge of an archive, one should be able to easily browse, search and query digitized documents from archive. This requires metadata extraction and indexing, a process that is often too time-consuming to be performed manually. Computer science, more precisely natural language processing, can help with automating this task, at least partially.

As a pilot project, personal names from our digitized archive were obtained. This task is more demanding in Balto-Slavic languages, such as Serbian, because personal names (and all nouns) change their form (decline) to reflect grammatical function in a sentence. Thus we have used two approaches for personal names extraction: dictionary based and machine learning. Dictionaries were upgraded by rules of declension. For machine learning approach, we have used a relatively small training corpus of hand-annotated data and MaxEnt algorithm. The programs that were used for the realization of the extractors of the personal names are:

1. Open-source developmental environment „Eclipse“ (Eclipse Public Licence)
2. Apache UIMA Framework – open-source UIMA implementation (UIMA – Unstructured Information Management Architecture) (Apache licence)
3. Apache OpenNLP library for machine learning (Apache licence)

We achieved F1 (precision/recall measure) score of 69% and the results of the work of extractor can be seen on the web site: http://195.252.93.45:8080/Srbislav/ (Picture 4).



Picture 4: The Interface of site Srbislav and recognition of the last name „Карамата"

Conclusion

The paper is dedicated to the digitization of the oldest report of the Scientific Council of the Mathematical Institute, which represents a significant history resource. The paper deals with two components of digitization: the working scheme and the technical infrastructure. The results are, in the lack of national strategy, the creation of a useful scheme for digitization of the cultural and scientific heritage, and its application in the preservation and presentation of the digital data. In addition to standard technical steps of the digital preservation and presentation, such as conversion, dissemination, and organisation... we successfully applied our software for the name entity recognition and made it available to the public. In the future we will be working on further theoretical foundation of the working scheme, as well as further exploration of the name entity recognition.

Attachment 1 – Working phases in the digitization of the cultural and scientific heritage

WP 1: The strategic planning

- Development of an idea/theme of digitization

Includes the fundamental analysis of the current political, economic, cultural, sociological and technological aspects with the goal of answering the current needs of the society (i.e. digitization: as a support of the scientific research and education, in the service of cultural tourism, strengthening and spreading of the national consciousness, as a mean of preserving of the cultural heritage…)

- The selection of the material for digitization

The selection of the material must be in accordance to the purpose, the selected theme of the digitization and it depends on its critical analysis (capacity for the estimating the relevance of the source, the capacity for estimating one's own point of view…). Considering that the selection is often the result of a subjective point of view (prejudice, partiality), most of the stakeholders rely on the certain criteria which will contribute to its objectivity, i.e. taking into consideration: the targeted population of the users, needs and the expectations of the society, benefits, digitized rights…

- The production of the long-term plan of the working activities

Includes the identification of the time needed for the realization of the project, as well as adequate resources (financial, personnel, technical) to start appropriate actions. The procedure needs to take into account the potential risks, as well as the legal environment (i.e. to solve the issues of the copyrights). The plan has to ensure the standards of the quality and the control procedure, reviews the issues of sustainability and controlling of the digitized collections in a long-term perspective.

W.P. 2: The Development and the Implementation

- Creation

Includes the production of digitized copies, analogue objects, as well as the modification of digitized surrogates with the purpose of dissemination. Includes technologies for converting materials from their analogue to a digital form, as well as for converting digitized material from one form to another. During the process, the originality, authenticity, and integrity of the new digitized artefacts in the relation to its originals must be considered.

- Dissemination

Includes the production of mechanisms (delivery technology) by which users can access the digitized material in the collection. The technology of delivery is not only related to the real delivery, but also to the complete organization of the informational infrastructure (ICT components: hardware, software, networks, databases, service, management...).

- Organisation

Includes the production of mechanisms (technology of research) that upgrade the research and provide help in finding. It also includes knowledge of the semantics: search based on an algorithm of natural language

- Storing and preservation

Includes adequate methods and technology for storing and preservation of digitized documents. The current issue with storing, from the technological point of view, is the non-existence of the concrete information of the expiration date and sustainability of the current digitized media for storing. Also, there's the issue of the

further insurance of the availability of the digitized documents that were stored during time in different (today usually outdated) formats. Possible strategies are: refreshing the data or migration of the data or replication of the data or emulation...

- Presentation

Includes the development of the user's interface, the presentation of the metadata, etc...

- Metadata

The cultural heritage, like the history of sources and the artefacts are only partly self-explanatory. The digital cultural heritage, therefore, must be supported by the appropriate metadata, so that firstly, they could be identified and secondly, useful. The issues that we face during the selection of the scheme of the metadata are numerous: the issues of the textual description of the non-textual entities, subjectivity in the evaluation of the cultural goods, the multitude of interpretations, paradigmatic changes in the related disciplines, the specificity of the language and the culture of the user group, the cultural diversity, etc.
The metadata of the digitized document imply a clear statement of the state and the usability of the digitized document itself. Metadata also represent and independent and a unique document that contains information on the author, authenticity, identification… There are many types of metadata, included in different phases of the process of digitization of the cultural goods, like in the phase of the creation of the digitized document, metadata includes the information on the author of the original object, as well as about the author of the digitized replica of the document, the type of the digitized document, the surrounding in which the digitization was made; in the phase of dissemination, the metadata are enriched with the information on the possibilities of the access to the digitized document, author's and owner's rights, license; in the phase of the organization of the digitized document, the metadata include the information on the publication, participants…

W.P. III: Quality Control

- Includes the process of checking and the quality control, legibility and accuracy of the: content, user access, used methodology and technology, delivery and new forms of preservation of the digitized and digital cultural goods

W.P. IV: Sustainability

- Includes the defining of the process of sustainability of the project, i.e. the ability of a stakeholder to provide a long-term service that is to provide an adequate budget, human recourses and other, for the long-term maintenance of the project, equipment etc.

msegan@mi.sanu.ac.rs
nikola.morena@gmail.com