1.09 Objavljen strokovni prispevek na konferenci
1.09 Published Professional Conference Contribution

Adnan TINJIĆ[*]

TOWARDS SYSTEMATIC DIGITIZATION: LESSONS AND SPECIFICS OF DIGITIZATION IN THE ARCHIVES OF THE TUZLA CANTON

Abstract:

The paper deals with the implementation of the digitization process that would include necessary steps from scanning archival material, creating metadata to creation of document formats for long term preservation of digital material. It includes information on some good practices and devotion to digitization in the Archives of Tuzla Canton, while using open source software, namely Xena and archivematica (for managing digital material in a repository), and also tries to point out issues and problems this Archives is facing during its attempts to improve digitization efforts.

Keywords:

digitization, Archives of Tuzla Canton, digital material, Xena (software), archivematica

Izvleček:

K sistematični digitalizaciji: lekcije in posebnosti digitalizacije v Arhivu Tuzelskega kantona

Prispevek se ukvarja z implementacijo procesa digitalizacije, ki bi vključeval vse potrebne korake, od skeniranja arhivskega gradiva, ustvarjanja metapodatkov do kreacije dokumentnih formatov za dolgoročno hrambo digitalnega gradiva. Vključuje informacije o dobrih praksah in digitalizacijo gradiva v Arhivu Tuzelskega kantona, ki v ta namen uporablja odprtokodno programsko opremo, tj. Xeno in archivematico (za upravljanje z digitalnim gradivom v repozitoriju). V prispevku so izpostavljene izzivi in problemi, s katerimi se arhiv sooča pri svojih poskusih izboljšati proces digitalizacije.

Ključne besede:

digitalizacija, Arhiv Tuzelskega kantona, digitalno gradivo, Xena (programska oprema), archivematica

1 INTRODUCTION

To digitize means to convert (as data or an image) to digital form.[1] In modern practice, the digitized data is in the form of binary numbers, which facilitate computer processing and other operations. However, "digitizing" should be understood not just as the act of scanning an analog document into digital form, but as a series of activities that result in a digital copy being made available to end users via the Internet or other means for a sustained length of time. The activities include:

- document identification, selection and preparation,

[*] *Adnan Tinjić, archivist, Archives of the Tuzla Canton, Franje Ledera 1, 75000 Tuzla, Bosnia and Herzegovina, contact: tinjicadnan@gmail.com; arhiv.tk@bih.net.ba.*

[1] *Digitize. (n.d.). Retrieved 16.1.2017, from https://www.merriam-webster.com/dictionary/digitize.*

- basic descriptive and technical metadata collection sufficient to allow retrieval and management of digital copies and to provide basic contextual information for the user,
- safety of the material being digitized,
- digital conversion,
- quality control of digital copies and metadata,
- providing public access to the material and
- establishing IT infrastructure to ensure that it can sustain long term growth, storage, and preservation of digital copies and metadata

Digitization is at the forefront of modern archival practice as one of the easiest ways to ensure needed protection, preservation and increased availability of records, i.e. archival material. This is the reason why it must not be neglected even by smaller, regional archives. These institutions must follow modern trends despite difficulties, by using any technical means and wide arrange of open source and related software solutions, thus reducing the costs and outsourcing of digitization process. The Archives of Tuzla Canton is one such archives, where new modes of practicing digitization must be found in order to overcome thelack of funding and personnel.

Digitization is done for several reasons:

- It makes archival material more available – we are able to send it via email or publish it online, thus enabling users and researchers to take advantage of modern technologies.
- It protects archival material – users and researchers can be provided with digitized copies exclusively, while paper originals can be kept safe and preserved. Publicly-available digital copies reduce handling and potential damage to valuable and often-fragile original items, which increases their longevity and historical value.
- Better control – digitized material can provide us with feedback and control in a more precise manner. It is easy to gather data regarding which fonds are the most used and for how long, etc.
- Faster and better quality of service – it is much easier and faster to search digital material, thus improving the service provided to researchers and users (Smajlović, 2012, str. 203).

Until recently, most collections in the Archives of Tuzla Canton were digitized by contractors who specialize in various types of originals: unbound paper, bound paper, searchable texts, microfilmed documents, etc. These digitization projects were often hit and miss regarding the result – first attempts with contractors were unsupervised and we were largely unaware of our own needs in the Archives, therefore, scanned data, while having decent quality in IT sense, was lacking in structure and metadata: names of files were not consistent with originals, folder structure did not follow series and

subfonds and there was less metadata[2] provided. This digitization procedure was improved. In the last couple of years, digitization process in the Archives of Tuzla Canton, either done by employees of the Archives or via contractors, is following certain simple, necessary steps:

1. Selection and preparation of archival material and equipment
2. Scanning (creating digital copies) of archival material
3. Digital preservation

2 SELECTION AND PREPARATION

Selection of archival material that needs to be digitized is usually done in advance, especially if scanning is being done by a contractor. Currently, the Archives of Tuzla Canton aims to fully digitize 34 archival fonds and collections that have the status of national monuments, i.e. important movable cultural heritage.[3] These fonds and collections are a priority for every digitization project in the Archives. They range from Ottoman era documents and manuscripts to records related to the Second World War. Several of these fonds and collections are already digitized, and getting the rest of them to digital form is planned in near future.

Preparation of a fonds for digitizing means that the fonds or collection will not be available for use during the period it is being scanned. Prior to scanning, we check the state and condition of material, type of paper (sensitive or too thin and already damaged paper usually must be scanned via overhead scanners), size, colour, folding state etc. Some documents, especially from the period after 1940's are bound together via paperclips or staples. These are usually completely rusted and must be carefully removed. This is of course the case with any other binding materials. We often try to remove them ourselves, rather than let contractors do it, at least for the most sensitive documents.

When outside contractors are hired to do digitizing (and microfilming), they provide their own equipment. For in-house digitizing, the Archives uses two flatbed scanners (one A4 and second one up to A3 format), one overhead scanner and one sheet-feed scanner. Sometimes, a digital camera is used for this purpose as well.

Size and type of documents determines what type of scanner will be used – most of the sensitive material is scanned with overhead scanners. As expected, the idea is always to minimize chances of damaging documents in any way.

With these relatively modest means and equipment, we have managed to digitize tens of thousands of records. During the previous year, our employees (mostly

---

[2] *Metadata can be defined as an information which describes significant aspects of a resource. Metadata is required to successfully manage and preserve digital materials over time and it will assist in ensuring that essential contextual, historical and technical information is preserved along with the digital object.*

[3] *The Comitee for National Monuments of Bosnia and Herzegovina declared 34 fonds and collections kept in the Archives of Tuzla Canton as national heritage monuments in 2009.*

volunteers and new archivists) scanned 8 collections and 3 fonds. Some digitization is still in progress – the Archives has an agreement with the municipal government and religious communities in which they are giving us parish registers for scanning. The Archives keeps digital copy (or even originals as is the case of the Islamic religious community of Tuzla) of these books and gives a complimentary digital copy to original owners when the original is returned to them. This project resulted in dozens of parish registers being digitized and preserved, at least in digital form, at the Archives of Tuzla Canton (and more being done each month). Most of scanned parish registers so far date from 1920's, and the originals are already in a rather poor condition. Digitizing them is probably the only way to ensure their preservation.

3 SCANNING OF ARCHIVAL MATERIAL

Whether the scanning is done internally or by a contractor, we have learned that certain standards and requirements must be applied. Especially if there is a contractor, demanding a certain level and quality of service as part of the contract proved to be very useful. For this purpose, we have turned to experiences of other archives, most of them abroad. According to these experiences and our situation, we have defined the following specifications:

- Textual Documents (and in rare cases maps and plans, etc.)

Media formats to be provided for scanning include original records, photocopies, photographic copy negatives or copy transparencies, or microfilm. The scanning resolution for the master files of 300 dpi[4] for smaller documents was selected to be compatible with OCR[5] software. The lower scanning resolution of 200 dpi for larger documents was selected to be of reproduction quality and to save file storage space.

- Photographs

Media formats to be provided for scanning include black and white and colour photographic prints, negatives or transparencies. Minimum value is 600 dpi, increasing resolution in intervals of 25 dpi as necessary to achieve a minimum of 6,000 pixels along the long axis (Puglia, Roginski, 1998).

---

[4] *The definition of a printed image will be given in DPI (dots per inch) or PPI (pixel per inch). This point number means that a printer can print so many points per inch (=2.54cm). The higher the value, the finer the print. A decent print definition is 300dpi, which meets most requirements.*

[5] *Optical character recognition (also optical character reader, OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document or from subtitle text superimposed on an image (for example from a television broadcast). It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line.*

Scanning itself is more or less a straightforward process. Due care must be given not to damage or change the state of original documents in any way. Also, formats in which we are scanning are either *tiff* and *jpeg* formats (most often used combination when scanning is done by contractors) or *pdf*, *jpeg, tiff* or *png* format (scanning done by our staff).

Of course, the most important issue is getting useful digitized material from the process and ensuring digital preservation of newly scanned data.

As for the metadata, different projects and collections may warrant more in-depth metadata capture than others. The functional purpose of metadata often determines the number of metadata that is needed. Identification and retrieval of digital images may be accomplished on a very small number of metadata, but in this case, more is always better. Metadata input is still largely a manual process and will require human intervention at many points in the object's lifecycle to assess the quality and relevance of metadata associated with it. This does not work well with contractors doing the digitizing. Supervision and additional "training" of contracted employees are modus operandi when metadata is concerned. We aim to include minimal descriptive elements (identifiers, captions or titles and creators). Some administrative and technical metadata is usually added as part of digital preservation methods and software.

4 DIGITAL PRESERVATION METHODS AND TOOLS

Digital preservation is a formal endeavour to ensure that digital information of continuing value remains accessible and usable. It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies and actions to ensure access to content, regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time (Digital Preservation Coalition, 2008).

To ensure digitization efforts have their long-term effects, the Archive of Tuzla Canton is using 2 software packages: Xena and archivematica.

Xena (shorter for "XML Electronic Normalising for Archives") is open-source software for use in digital preservation. It is a Java application developed by the National Archives of Australia, available free of charge under the GNU General Public License[6].

Xena software aids digital preservation by performing two important tasks:

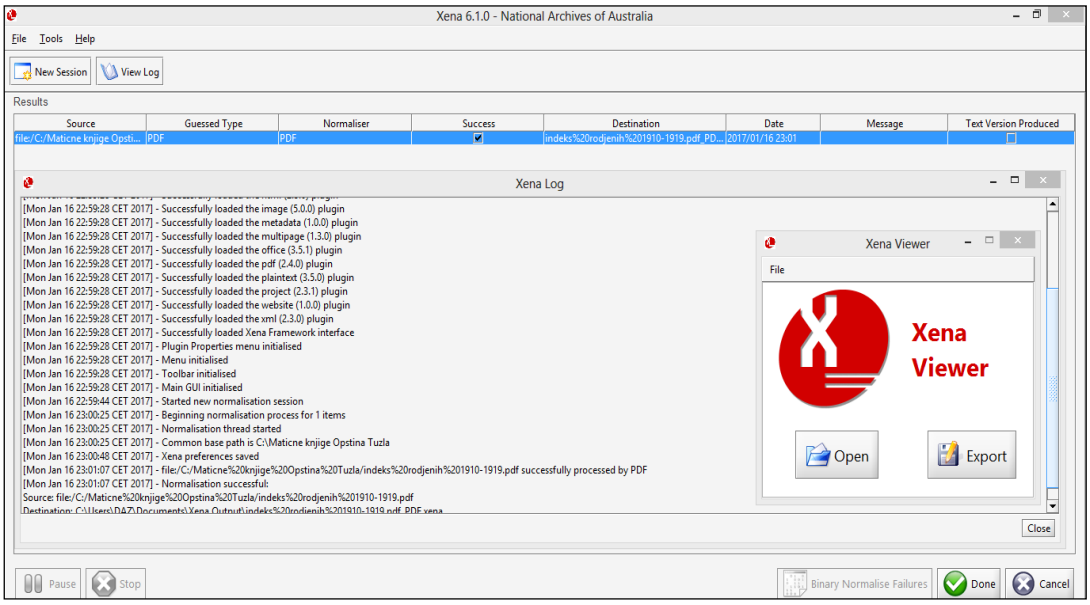
- detecting the file formats of digital objects
- converting digital objects into open formats for preservation.

Platforms supported by Xena are Microsoft Windows, Linux and Mac OS X. The software uses a series of plugins to identify file formats and convert them to an appropriate openly specified format. It can create plain text versions of file formats such as TIFF, Word and PDF. The Xena interface or Xena Viewer can be used to view or export

[6] *The GNU General Public License (GNU GPL or GPL) is a widely used free software license, which guarantees end users the freedom to run, study, share and modify the software.*

a Xena file (extension .xena) in its target file format. These files contain the normalised file as well as any extra information relevant to the normalisation process (Tinjić, 2012, str. 323).

Essentially, this enables us to preserve all kinds of formats as one, xena format, and to convert all licence format to their open-source counterparts. For example, we can keep images and documents in converted XML format along with original formats, as well as xena format. Instead of making sure that we own multiple software (and licences) to open these files in (near) future – we only have to rely on one, Xena software package.



Picture 1: Normalisation (conversion) session in Xena software

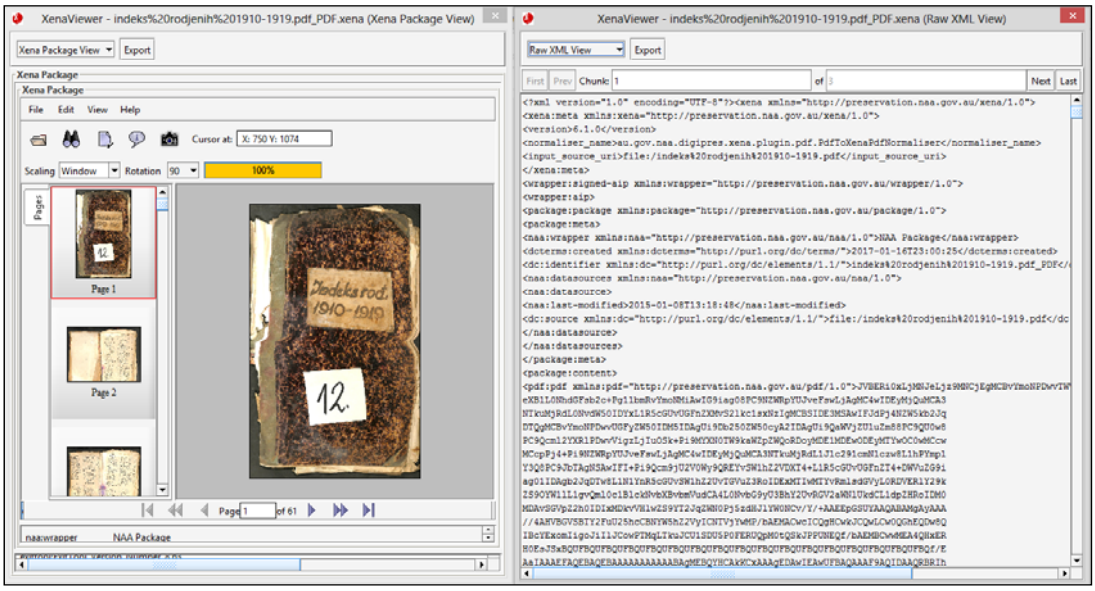
Since its open-source, it can even be upgraded (if needed) with some programming knowledge in Java. Making sure we can always access digitized material no matter what operating system or version of a software we are running, either for text processing, image viewing, etc. makes digitization truly purposeful.

Since late 2015. we have started using a more complete, feature-rich software for digital preservation, which is used in parallel to Xena, but it can be considered as significant upgrade.

This software was developed by company Artefactual Systems, also known as creators of ICA AtoM (Access to Memory) software solution. Archivematica is a free and open-source digital preservation system that is designed to maintain standards-based, long-term access to collections of digital objects. All of the software, documentation and development infrastructure are available free of charge and released under AGPL and Creative Commons licenses.[7]

[7] *Archivematica web page, Retrieved 16.1.2017, from https://www.archivematica.org/en/.*

Archivematica uses a micro-services[8] design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. As for the standards, it uses METS, PREMIS (events, agents, rights and restrictions), Dublin Core, the Library of Congress BagIt specification and other best practice standards and practices to provide trustworthy, authentic, reliable, and interoperable archival packages (AIPs) for storage in repository.



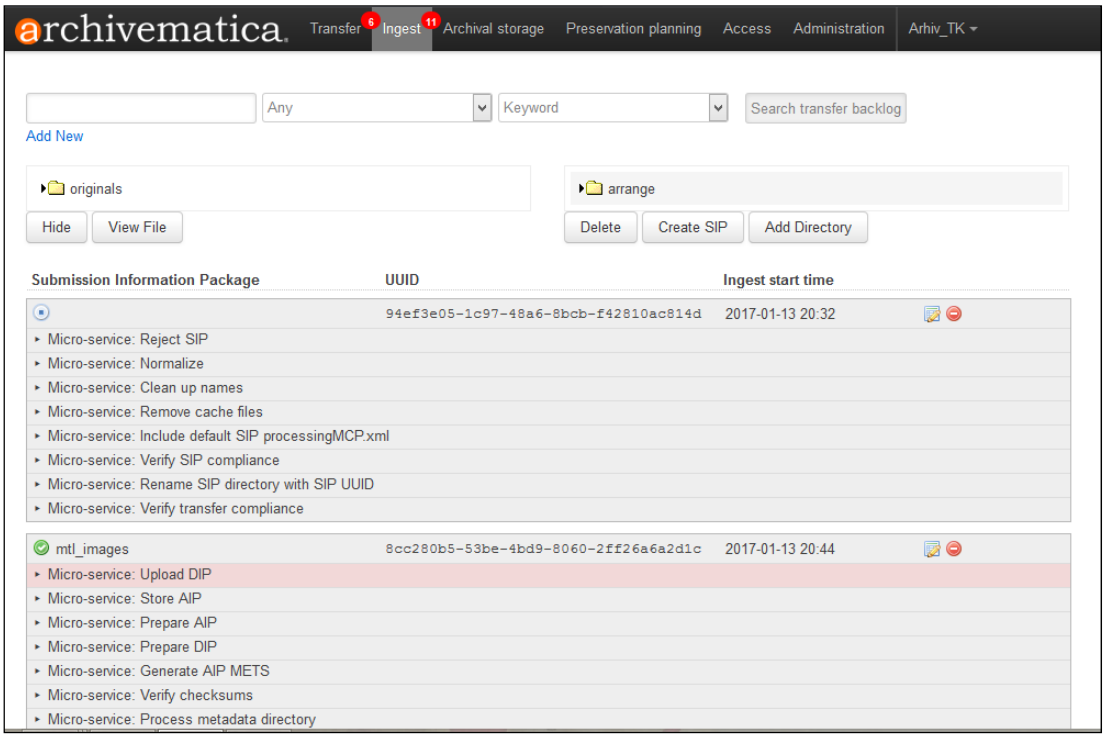
Picture 2: Image of a parish register in original digitized (pdf) format (left) and same document converted in XML via Xena software (right)

From the end-user perspective, all of Archivematica's functions take place within a web-based dashboard, which can be accessed from anywhere by logging in through a web browser.

Main dashboard contains several tabs: Transfer (adding files), Ingest (processing files/normalisation process), Archival storage, Preservation planning, Access and Administration. If the transfer is accepted, the tool performs an initial analysis – calculating checksums, scanning for viruses, extracting metadata – and then offers to create a Submission Information Package (SIP). Metadata can then be added to the SIP before it is ingested. At ingest, the curator can choose various routes such as Preservation (where the digital objects are normalised to archival formats and transformed into an Archival Information Package, or AIP), Access (where the digital

[8] *Microservices is a specialization of an implementation approach for service-oriented architectures (SOA) used to build flexible, independently deployable software systems. Services in a microservice architecture are processes that communicate with each other over a network in order to fulfill a goal. Archivematica micro-services are granular system tasks which operate on a conceptual entity that is equivalent to an OAIS information package: Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP). The physical structure of an information package will include files, checksums, logs, submission documentation, XML metadata, etc.*

objects are normalised to dissemination formats and transformed into a Dissemination Information Package, or DIP) or repackaging without normalisation. The system is fairly easy to use, as it is based on OAIS Reference Model[9]. In the case of the Archive, software is used on a Linux based web server, but it can be used with the virtual machine image as well. By bundling together numerous open-source tools in an easy-to-use interface and providing a selection of ready-made processing scripts, Archivematica allows archivists with even modest technical knowledge to process electronic records according to best practices, thereby greatly increasing the effectiveness of digital preservation (Houston, 2014).



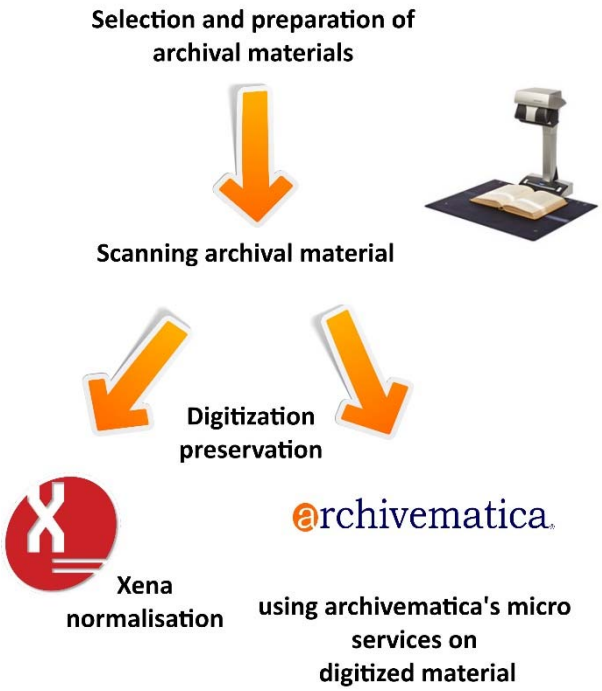
Picture 3: archivematica dashboard with listed micro-services

Finally, after processing digitized material with our software, copies are made for safekeeping. If there is no contractor involved, scanning is done only on dedicated PCs (personal computers) chosen only for this purpose, and material is processed and moved (as additional copy) to external hard drives and DVD/CDs.

Users and researchers are provided with the copy kept on external hard drives, DVDs or CDs. There are always at least two copies of every digitized fonds or collection. For the most important archival material (and eventually for all digital material), we make three: one of them is stored in a fire-proof safe (external hard drive is usually the carrier in this case), one on the computer's hard drive and one on an external hard drive (or CD/DVD, depending on the size). With xena and archivematica, we do not have to

[9] *OAIS (Open Archival Information Systems Reference Model — ISO 14721:2003) provides a generic conceptual framework for building a complete archival repository, and identifies the responsibilities and interactions of Producers, Consumers and Managers of both paper and digital records. More details can be found at: https://public.ccsds.org/pubs/650x0m2.pdf.*

worry about additional software we might need, but migrations are needed for long-term safety of digital material.



Picture 3: Three-steps approach to digitization in the Archives of Tuzla Canton

While we must aim to make duplicate copies and digitize more and more archival material, we also have to balance this with financial costs – storage space comes with a price. Migration can be costly, so it must be planned ahead, and also have priorities. Digital carriers of choice in the Archives of Tuzla Canton are mostly hard drives and external hard drives due to their capacity and relative longevity. But they can, and will, break down. While our battle with format types and obsolete software might be temporarily won by using open source software, the battle against wear and tear of digital media carriers has only just begun.

SOURCES AND LITERATURE

- Archivematica software documentation. Retrieved 17.1.2017, from https://www.archivematica.org/en/docs/archivematica-1.6/.
- Digital Preservation Coalition (2008). "Introduction: Definitions and Concepts". Digital Preservation Handbook. York, UK: Digital Preservation Coalition. Retrieved 17.1.2017, from http://www.dpconline.org/docman/digital-preservation-handbook/299-digital-preservation-handbook/file.
- Domazet, S. (2015). Archival response to the challenge of long term preservation: Archives of Bosnia and Herzegovina in the era of digitization. *Atlanti, 25 (1),* pp. 71-81. Trieste: International Institute for Archival Science of Trieste and Maribor, State Archive of Trieste.
- Fröhlich, S., Schöggl-Ernst, E. (2015). Digital long-term preservation in Austria. *Atlanti, 25 (1),* pp. 265-275. Trieste: International Institute for Archival Science of Trieste and Maribor, State Archive of Trieste.

- Guidelines for digitization projects for collections and holdings in the publicdomain, particularly those held by libraries and archives (2002). The Hague: IFLA. Retrieved 18.1.2017, from http://www.ifla.org/files/assets/preservation-and-conservation/publications/digitization-projects-guidelines.pdf.
- Houston, B. (2014). *Archivematica Review*. Milwaukee, University of Wisconsin: The American Archivist Reviews. Retrieved 17.1.2017, from https://reviews.americanarchivist.org/2016/07/02/archivematica/.
- McKay, S. (2003). Digitization in an Archival Environment*. Electronic Journal of Academic and Special Librarianship v.4 no.1 (Winter 2003*). Retrieved 18.1.2017, from http://southernlibrarianship.icaap.org/content/v04n01/Mckay_s01.htm.
- Puglia, S, Roginski, B. (1998). *NARA Guidelines for Digitizing Archival Materials for Electronic Access*.College Park, Maryland: National Archives and Records Administration.
- *Reference model for an open archival information system - OAIS.* (2012). Washington, DC, USA: Consultative Committee for Space Data Systems. Retrieved 16.1.2017, from https://public.ccsds.org/pubs/650x0m2.pdf.
- Smajlović, L. (2012). Importance of information systems for the development of archival science. *Arhivska praksa 15, pp. 305-310.* Tuzla: Arhiv Tuzlanskog kantona.
- Škoro Babić, A. (2015). Management, appraisal and long term preservation of e-records: role of archivists and IT professionals. *Atlanti, 25 (1),* pp. 217-225. Trieste: International Institute for Archival Science of Trieste and Maribor, State Archive of Trieste.
- Tinjić, A. (2012). Open source software for archives – experiences of the Archives of Tuzla Canton. *Arhivska praksa 15,* pp. 311-322. Tuzla: Arhiv Tuzlanskog kantona.

POVZETEK

K SISTEMATIČNI DIGITALIZACIJI: LEKCIJE IN POSEBNOSTI DIGITALIZACIJE V ARHIVU TUZELSKEGA KANTONA

Majhni regionalni arhivi se težko spopadajo z modernimi praksami, digitalizacija pa je ena takšnih, ki je nujna. Arhiv Tuzelskega kantona, omejen s financiranjem in osebjem, se s tem sooča ob vsakem projektu, ki se ga loti. Prvi poskusi so bili enostavni, vendar ne dovolj dobri, da bi zagotovili dolgoročno uporabno digitalno gradivo. To prakso smo uspeli izboljšati z implementacijo izkušenj kolegov iz tujine in tako ustvarili enostaven, vendar učinkovit proces, ki zagotavlja, da je naše novoustvarjeno digitalno gradivo v skladu z minimalnimi zahtevami in ga lahko dolgoročno hranimo, prav tako pa je uporabno tako za raziskovalce kot za arhiviste. To je le korak proti boljši in popolni digitalni hrambi na državnem nivoju, ki bo kmalu postala realnost tudi v Bosni in Hercegovini, kot je to že v večini evropskih držav.