TRANSCOM 2015, 22-24 June 2015
University of Žilina, Žilina, Slovak Republic

Digitization Workflow at University of Žilina

[*]Marcela Palkechová, [**]Veronika Murgašová
[*]University of Žilina, Faculty of Humanities, Department of Mediamatics and Cultural Heritage, Univerzitná 1, 01026 Žilina, Slovakia, marcela.palkechova@mediamatika.sk
[**]University of Žilina, Faculty of Humanities, Department of Mediamatics and Cultural Heritage, Univerzitná 1, 01026 Žilina, Slovakia, veronika.murgasova@mediamatika.sk

Abstract. This article aims to give latest information about digitization process at the University of Žilina in the context of Memory of Slovakia – The national center of excellence. We briefly describe hardware and software equipment and process of digitization. Paper describes digital objects created during the process of digitization of cultural heritage ongoing at the University of Žilina and their presentation.

Keywords: workflow, digitization, repository, digital objects.

1. Introduction

Often digitization is understood as just being the conversion of an analogue information object into a digital format. However, digitization is more than just the technical conversion from analogue to digital. Digitization is a process that involves various stages. It starts with determining digitisation goals, then proceeds with the selection and the preparation of documents, the definition of the quality parameters, the actual conversion from analogue to digital, the quality control of the digital files, the long-term storage, and ends with making the digital content accessible. Hardware and software used for digitization are critical to the success of the whole project [1].

1.1. About Memory of Slovakia – The National Centre of Excellence

Memory of Slovakia – The National Centre of Excellence in research, protection and accessibility of cultural and scientific heritage was founded in 2010. This project was developed in cooperation with the Slovak national library in Martin. One of the main objectives of the project was to build up The Centre of Excellence located in the University Library of University of Žilina. It serves as a training centre for students of library and information science.

1.2. Hardware

The centre is equipped with modern technology. There are three types of scanner robots.

- **Treventus ScanRobot 2.0 MDS -** is scanner for mass digitization. This automatic scanner is able to scan up 2500 pages per hour. It is used to scan of industrial book production [2].

- **Bookeye 3** - is ideal scanner for digitization of valuable scripts, folders, drawings, plans and especially historical books because this scanner has motorized book cradles, which are gentle on delicate materials and bookbinding [3].

- **XINO S700** - is the fastest automatic scanner, which has been purchased to the Centre of Excellence. The XINO S700 scanning system is equipped with a feeder for 500 sheets. It features batch-oriented processes, professional image editing and standardized user dialogs. The scanner is suitable for single sheets without bookbinding [4].

1.3. Software

During the digitization process it is necesary to use different types of softwares for different types of tasks.

- **ScanGate -** is software for image treatment. It has many functions, that are used for automatic correction of scanned pages like automatic border recognition, deskewing, cropping, resizing, binarization, brightness or contrast adaption, unsharp masking, page rotation and many others.

- **MarcEdit -** is a free library metadata software. It includes a built in Z39.50 which is national standard defining a protocol for computer-to-computer information retrieval. It allows user to query other library systems and download bibliographic records [5].

- **ABBYY Recognition Server –** is a server-based OCR software that allows to establish process of converting paper to searchable and reusable electronic documents. ABBYY Recognition Server takes care of the whole document capture routine, providing convenient tools for recognition, verification, attributing, full-text indexing, and document conversion. It converts scanned documents to searchable PDF and PDF/A standard for long-term preservation. Server has 190 supported recognition languages [6].

2. Digitization Workflow

The center's digitization workflow describes the standards, specifications, and processes involved in our digitization efforts. Developing a workflow reduced the time to scan and keeping the process consistent [7].

Stages of workflow:

Measurement

As a first step we create documents containing information about physical conditions of every object before scanning. This involves measurement of paper thickness, book width and height, classification of bookbinding type and type of paper. We define degree of yellowing and degree of damage bookbinding. This document serves for statistical purposes and it helps to researcher to prognose the lifetime of the document. It is a part of submission information package (SIP) which is deposited in archive.

Scanning

Choice of scan robot depends on type of documents and its bookbinding. Default resolution is set to 300 DPI. During throughout the process the operator performs continuous quality control.

Image treatment

The main steps that may need to be done with batch processing with ScanGate software:

- Rotate the image
- Crop the image
- Adjust tone and color
- Assign or convert color space for the image

- Set the format of the output file

Metadata

With the MarcEdit software we can download records via z39.50. Z39.50 service offers access to MARC21 bibliographic records from the Slovak national library, the British Library's full catalogue or to a selection of the Library of Congress cataloging records. Metadata are saved as a part of the SIP in XML format.

Creating Submission Information Package

The result of digitization is packaged in several "delivery packages"- Submission Information Package (SIP). Each SIP will correspond to a particular digitized document. Each SIP will contain all the files that make up a digital image of a document. The preservation master (MST) is the highest-quality digital surrogate of the physical document in TIFF format. As it should accurately represent the original document, this digital copy should not be altered for aesthetic reasons. Web-access copies of documents in JPEG format (TRT) are created for digital exhibits [8].

Ingesting to the archive

Ingest refers to the processes of preparing data and digital objects for adding to a digital archive and of adding them to the digital archive. In this step master file is stored in archive and presentation copy is sent to ABBYY server for optical character recognition.

OCR

Optical character recognition is conversion of images into machine-encoded text. ABBYY recognition server is used for OCR. This type of software can automatically analyze images of printed texts and turn it into a form that a computer can process more easily. Recognition of historical fonts poses big problem. The recognitions of Latin-script is still not without mistakes, but results are very satisfying [9].

Presentation

MediaINFO is a web presentation solution for books, newspapers, manuscripts, maps and other scanned material. It enables visitors to browse, search and use content interactively. Whole interaction is done through Adobe Flash, which is most widespread viewing platform with 99% penetration in browsers. Domain address is http://mi.ceps.mediamatika.sk/

Main application window allows browsing through main categories and detailed browsing through visual tree of subcategories. User can simply (de)select any of the nodes to turn on/off display and searching through specific publications.

Mediainfo has several features of end-user web application:

- full-text search with support for BOOLEAN operators
- filtering and searching through various metadata fields
- the results can be viewed in number of ways (Zoom view, Book view)
- creating personal notes and shared it with other people
- special hyperlinks can be created to link directly to content without required login
- customization (basic changes to backgrounds, transparency level, language)[10]

Fez – another presentation software - is an open source and is actually used for Fedora based digital repository and workflow management system. Domain address is http://fez.ceps.mediamatika.sk/. Structure in a Fedora repository consists of communities (departments of all faculties), every community has collections (researchers, teachers…) and every collection has records (publications). Records can belong to both collections and communities. Its advantage is that it allows presenting the multimedia files.

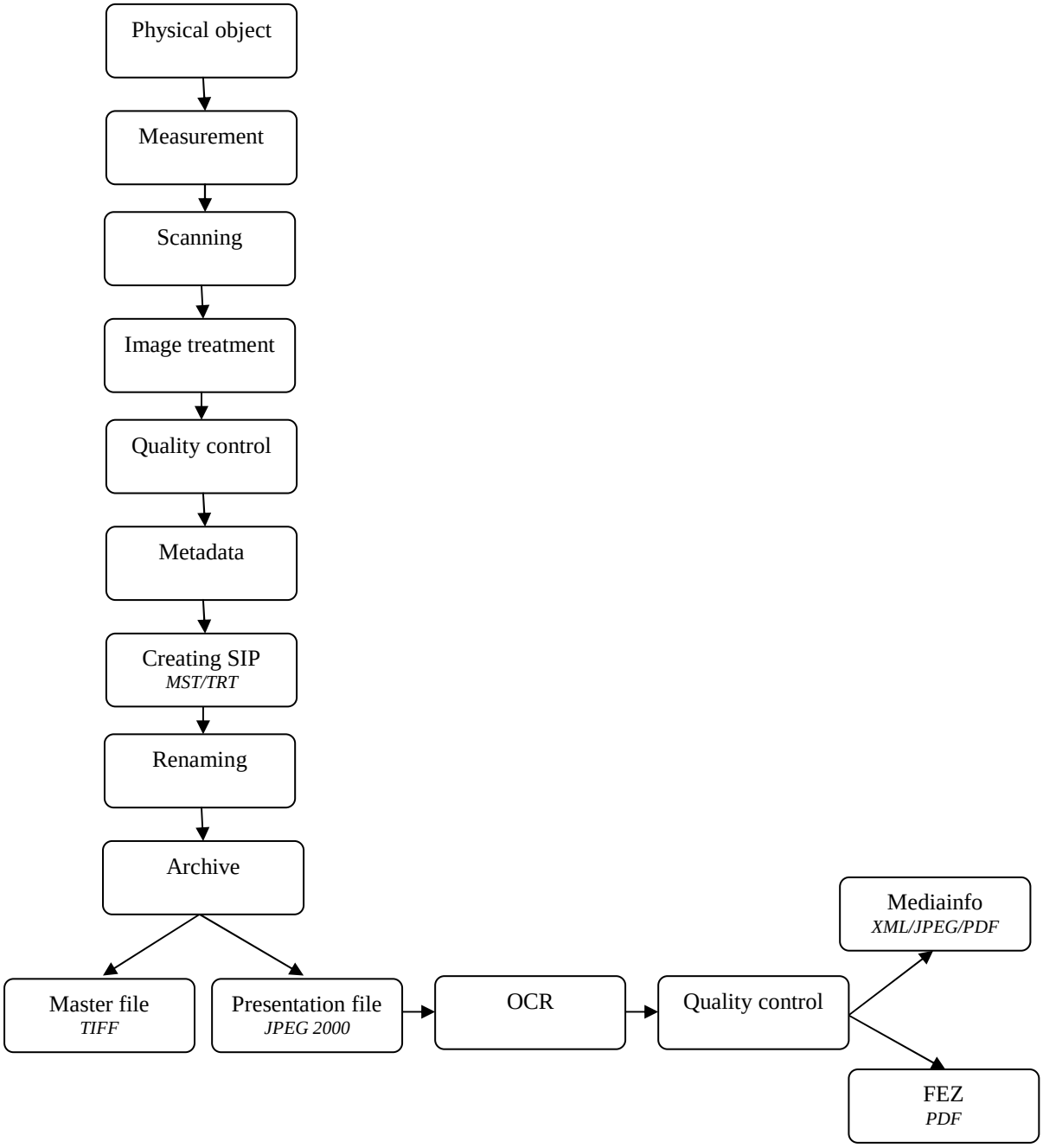

Fig. 1 Digitization workflow

3. Research in The National Centre of excellence - Memory of Slovakia

Created and modified digital objects form the output of digitization. These digital objects are used for research activities in various field of study, especially by PhD. students.

Currently, The National Centre of Excellence - Memory of Slovakia is working with the research sample, which, was created by scanning documents from the library of The Department of Mediamatics and Cultural Heritage. These objects will serve the students for various purposes and will be accessed by Mediainfo.

Tab. 1 Digital objects in archive

Objects	390
Pages	103 099

The second research sample is the book collection from Tranoscius library. Tranoscius is a historical library with rare old books. Books from this collection need very special approach and careful handling. There are 110 objects to primary research of historical books. The processes applied to the sample are: research of optical character recognition for books printed before year 1830 or impact of the bookbinding on the speed and effectivity of a digitization process and the quality of the resulting digital image.

4. Conclusion

Digitization is a complex process that affects the number of factors. The aim of the research center is to optimize the digitization process, improve efficiency and speed of the process. Through optimized processes within the archive are fully digitized 103,000 pages which will be shortly available to students, teachers of Mediamatics and Culture Heritage and researchers in the field. Our intention is to create an academic repository as a common space for research staff and academic employees in which they can share their work and to increase the understanding of their research. Creating online library for students and academic repository for the staff increases the level of university.

Centre of Excellence serves not only to simple digitization. The informations about physical condition of the units are recorded. The outputs will be used as a methodology for other digitization projects in the future.

References

[1] *MINERVA, Technical Guidelines for Digital Cultural Content Creation Programmes (Version 2.0)* [online]. 2008 [cit. 2015-03-29]. Dostupné na: http://www.minervaeurope.org/publications/MINERVA%20TG%202.0.pdf

[2] *Mass digitization system. Treventus 2* [online]. 2009 [cit. 2014-12-08]. Dostupné na: http://www.treventus.com/downloads/ScanRobot_2.0_MDS_BookScanner_brochure_web.pdf

[3] *Bookeye 3. Operation Manual* [online]. 2014 [cit. 2015-01-05]. Dostupné na: http://www.imageaccess.com/PDFs/BE3-R1_OperationManual.pdf

[4] *XINO S700 High-performance scanning system* [online]. 2013 [cit. 2015-01-25]. Dostupné na: http://www.microform.de/pics/referenzen/107_XINOS700_Folder_E.pdf

[5] *MarcEdit Development* [online]. 2013 [cit. 2015-03-28]. Dostupné na: http://marcedit.reeset.net/features

[6] *ABBYY Recognition Server* [online]. 2015 [cit. 2015-03-23]. Dostupné na: http://www.abbyy.com/recognition_server/

[7] BANACH, Meghan et al.*Guidelines for Digitization* [online]. 2011 [cit. 2015-01-29]. Dostupné na:http://www.library.umass.edu/assets/aboutus/attachments/UMass-Amherst-Libraries-Best-Practice-Guidelines-for-Digitization-20110523-templated.pdf

[8] *Metadata for digitized newspapers in Project SAP Specification documents (Version 2.4)* [online]. National library of Sweden, 2014 [cit. 2015-02-11].Dostupné na: http://www.kb.se/namespace/digark/deliveryspecification/agreement/sap/sap_specifications_eng_v2.pdf

[9] WOODFORD, Chris. *Optical character recognition (OCR*) [online]. 2014 [cit. 2015-01-05]. Dostupné na: http://www.explainthatstuff.com/how-ocr-works.html

[10] *MediaINFO System presentation.Give Life to Digitized Books.* Switzerland: Geneza, 2012.