Digital long-term preservation in Albania, determining the right file formats

Ardiana Topi
University of Tirana, ardianatopi@yahoo.com

Aleksandër Xhuvani
Polytechnic University of Tirana

Digital long-term preservation in Albania, determining the right file formats

Ardiana Topi[1, 2], Aleksandër Xhuvani[3]

[1]University of Tirana, Faculty of Natural Sciences, Department of Informatics, Blvd. Zogu 1, Tirana, Albania
[2]National Archives of Albania, Street Jordan Misja, 8303, Tirana, Albania
[3]Polytechnic University of Tirana, Faculty of Information Technology, Department of Computer Engineering, Nënë Tereza Square, Tirana, Albania
ardianatopi@yahoo.com[1]

Abstract. The National Archive of Albania (NAA) is the main institution in Albania, according to legislation, responsible to create the rules for document life management produced by any organization, and preserve them. Almost all archival documents stored in archival institutions, must be saved for unlimited period of time, or forever. Preservation of digital documents is different from preservation of paper documents. It is a very complex procedure directly connected with the inevitable phenomenon of hardware and software obsolesces. A successful mission of the NAA must have a well-defined strategy regarding the Digital Long-term Preservation. This paper aims to present the possible file formats, able to use during the digitization process of paper based documents and photography.
Technical characteristics of same useful file formats are analyzed. The study tells that sustainability and quality factors are very important in digital file format selection for long term preservation.

Keywords: file format, image quality, digital collection

1. Introduction

When we speak for the digital preservation, the Archival Institutions must ensure not only the preservation of their archival files but ensure that these files will be authentic and readable for long time. This task is not simple in the digital world. Long-term Digital Preservation refers to the series of managed activities necessary to ensure a continuous access to digital materials or at least to the information contained in them for as long as necessary, in the most of cases indefinitely [1]. The machine and software dependency of digital collections, the rapid changes in technology, the fragility of the media, the ease manipulation makes the digital long term preservation a very complex task. File formats are very important in the ability for reusing and accessing data into the future that's because we must ensure to produce to high quality digital files. Those files must be produced as the best copy by a digitizing process, where "best" is defined with meeting to the objectives of a particular project or program [2].
The file formats in this study are chosen based on useful file formats in the Institution of General Directory of Albania [3]. The digital file formats vary from simple to complex format structures. As the complexity of the individual formats increases, the requirements for the structure and completeness of the documents must be adjusted accordingly and realized consistently [4].
Our study aims to give recommendations regarding the file formats that can be used for digital long term preservation and suggest the evaluation methods that can help the establishment of national level standards and guideline in digital preservation field.

2. The issue of digital preservation

Nowadays, there is a rapidly increasing volume of documents and information created in digital form. In every office we are going to create the documents digitally. In the same time there are lots of public institutions that have started digitized projects that are going to produce large amounts of digital content. But how does the creation of a digital document works? To create a digital document we need hardware devices (a computer and/or scanner) and an application that define the structure and encoding of that document. Reuse or editing the same digital document, needs an application that can distinguish the way in which this document is encoded and based in that information the application will be able to render it to the monitor. In the most of cases each application generates its own file format. Actually, a large number of file formats are used worldwide. The issue stands, whether we are able to read those documents after a long time without missing information and how to ensure that? The file format that is defined as "a standard way that information is encoded for storage in a computer file" [5].

Since we can use different file format for the same types of documents the issue raised is which is the best file format for long term preservation and how can we evaluate it?

To answer this question, initially we identified the types of documents preserved in archival institutions[1]. The documents found in different forms are grouped in four different categories: paper based, photography, audio and video tape files. To convert those types on digital version different types of file format need to be used. Each of those formats is completely different in structure, in encode/decode algorithms and offer different advantages or disadvantages.

In this paper are presented digital file formats that can be used for preservation of digitally converted paper based documents and photography. The study was based on research in scientific literature review, International Standards and Best Practices applied and results of our project. The tests are accomplished on the factors that have consequences in image quality and storage capacity. Finally we have collected information about sustainability factors [6] for every file format studied.

3. The implemented methodology

Three main issues were studied: the storage capacity, the quality and sustainability factors of file formats. Those issues were studied based on random sampling of documents from NAA's collections and were grouped in three different categories: machine write documents, manuscripts and photography sets.

We used a workstation connected with a flatbed scanner. The software used for capture and edit was the most difficult task, due to the lack of definition of the "image quality". The image itself can be subjective (measured by human eyes) or objective (measured by calculate digital image properties) [11]. In order to achieve an acceptable image quality, it is necessary examination of the whole image processing chain especially the issues concerning the calibration of scanning and rendering equipment (monito and printer). The quality is affected by many factors such are source material, devices used to capture, the subsequent processing, and compression method A quality rating system must be linked with defects appears on digitization and processing of digital information based on evaluation on-screen comparing with originals. The technical characteristics that influence on the image quality are: Resolution, Color space, Bid depth.

The resolution was selected based on Quality Index formula (QI = 0.013 dpi x h) introduced by Kenney and Chapman (1995). This resolution benchmarking method has its roots in micrographics, where standards for predicting image quality are based on the quality index too. QI provides a tool for relating system resolution and legibility, where h is the weight of the smallest letter in document (usually e). This method employs three levels: high level QI = 8; medium level QI =5; and marginal

[1] The study is done on documents collections of ANA and here we are speaking for traditional documents not digital born documents. It will be object of another study.

level QI = 3.6. For non-text material we adapted this formula by replacing the high of smallest character in text with the smallest complete part that is considered essential to an understanding of the entire page, for example width of the lines used to render and it can be different for a photography or for a papyrus manuscript.
The storage capacity was evaluated based on the information collected after finishing the digitized process for each case. The data are organized and illustrated in following tables. The sustainability factors are evaluated based on information available for each in terms of different convention scoring. One example is demonstrated in the paper.

4. Technical aspects of digital collections

The paper based documents in archives are found in different forms such as: textual documents including manuscripts, artworks and photographic formats and are found in both black and white; and color mode. For this category of documents the conversion from analog to digital form is accomplished by using scanners or digital camera. We have chosen the scanners since the light conditions are very crucial for digital camera and have direct influence on image quality. The devices in scanning system (monitors and scanners) are calibrated using ISO charters, in order to take the best image quality, regarding the tone and color reproduction, system spatial uniformity, spatial frequency response, noises, etc. [7].
The digital images obtained by scanning process must be vector or raster. Vector images are made up of basic geometric shapes such as points, lines and curves. The relationship between shapes is expressed as a math equation, which allows the image to scale up or down in size without losing its quality. Raster images are made up of a set grid of dots called pixels where each pixel is assigned a color value. Unlike a vector image, raster images are resolution dependent.
In the most of the cases for preserving their documental heritage the archival institution use the raster images over the vector images. For both types of digital images exists a various number of file formats. The most useful file formats used worldwide for long-term digital preservation as master copy are TIFF format, JPEG 2000[3], and for access copy JPEG, GIF. The quality of digital raster images is depending on some technical characteristics such are: Spatial Resolution; Signal Resolution or Bit- depth; Color Mode, and Compression method, presented below.

4.1. Digital file formats

4.1.1. Tagged Image File Format (TIFF)

The TIFF format was originally developed by the Aldus Corporation, and since 1996 the TIFF specification was maintained by Adobe Systems Incorporated. The TIFF file is a tag-based file format for storing and interchanging raster images. It starts with an 8-byte image file header (IFH) that point to the image file directory (IFD) with the associated bitmap [8]. The IFD contains information about the image in addition to pointers to the actual image data. The TIFF tags, which are contained in the header and in the IFDs, contain basic information, the manner in which the image data are organized and whether a compression scheme is used, for example. The TIFF 6.0 version offers users the option to use their additional tags. TIFF supports color depths from 1-24 bit and a wide range of compression types (RLE, LZW, CCITT Group 3 and Group 4, and JPEG), as well as uncompressed data. TIFF also incorporates the most comprehensive metadata support of any raster format, allowing the addition of a wide variety of technical and resource discovery information to be included. The TIFF specifications are freely available for use.

4.1.2. JPEG File Formats

The JPEG standard (ISO/IEC 10918) was created in 1992 (latest version, 1994) as the result of a process that started in 1986 [9]. It is composed of four separate parts and an amalgam of coding

modes. JPEG itself is not a file format, but represent an image compression algorithm. The File Interchange Format (JFIF) has become a de facto standard; and is commonly referred to as the JPEG file format. The JPEG algorithm is supported by a number of other raster image formats, including TIFF. JFIF and SPIFF are 24-bit color formats and use lossy compression. JPEG is particularly suited to the storage of complex color images, such as photographs.

4.1.3. JPEG 2000 File Formats

JPEG 2000 is an improved version of lossy JPEG algorithm, developed by the ISO JPEG group in 2000. The JPEG 2000 has the option for lossless compression using wavelet compression to achieve higher compression rates with a lower corresponding reduction in image quality. The JPEG 2000 can compress a file to a desired size specifying the size in bytes, or specifying a compression ratio. It supports color depths up to 24-bit (true color) and is best suited to complex color images, such as photography. JPEG 2000 in the year 2007 is divided into twelve standards that are all more or less derivations of or supplements to the first standard. A JPEG 2000 file consists of a succession of boxes. Each box can contain other boxes and have a variable length determined by the first four bytes and a type that is determined by the second sequence of the four bytes. Every JPEG 2000 file begins with a JPEG 2000 signature box, followed by a file type box where is determined the type and the version. This is followed by the header box, which contains the boxes with the resolution, bit depth and color specifications.

4.2. File format result assessments

The selected test images are scanned with the same resolution in different file format. To decide whether which resolution can provide the right image quality we referred to equation of Quality Index (QI). In order to get a quality index QI=8 to a document where the weight of smallest character is 1 mm we tested the resolution on: dpi = 2*8/0.339*0.1 = 400 pixel per inch. Observed by human eyes it does not distinguish any change to three different types of file format used. In the figure is illustrated a part of the same document scanned with resolution 300 dpi, RGB 8-bit for channel and are saved in three different file formats at actual size 1" = 1".

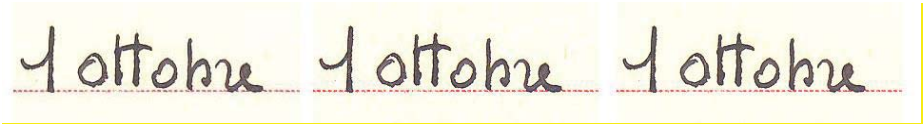

Figure 1: Examples of fragments (TIFF uncompressed; TIFF compressed and JPEG 2000 lossless)

The LZW compression method is lossless that's because we have not found any degradation of the image quality. The image clarity of both JPEG 2000 lossless and 50 quality was high and there was not any visible defects when evaluate their histogram.

For the same images were calculated their storage capacity. As a result we have found that that TIFF LZW in lossless mode may save about 49% of storage capacity compared to an uncompressed TIFF. The images scanned with JPEG 2000 (jpf) in lossless mode has not offer significant benefit compare with TIFF LZW. The documents saved in JPEG 2000 (jpf) format with 50% quality they occupy about 80% less storage capacity then JPEG 2000 in lossless mode. In the figure 2 are demonstrated the results of 5 tested images, scanned with 300dpi resolution and saved in five different file formats.

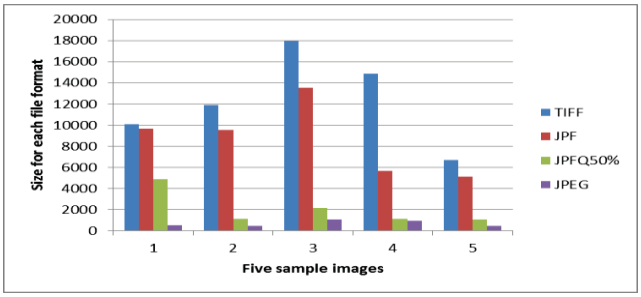

Figure 2: Storage capacity of documents saved in different file formats.

Regarding to the sustainability factors, we employed two different methods for evaluation process proposed by Rog and van Wijk [11] and the second proposed by Arms and Fleischhauer [12]. Both methods analyze almost the same sustainability factors but use different ways for scoring. In the first method the decision way on weighing factor values of different criteria and characteristics was unclear and influent directly to the assessment values we decided not to use it.

Table 1: A collection of the sustainability factors for both three file formats: TIFF, JPEG 2000 and JPEG.

Sustainability Factors	**Format: TIFF**		**Format: JPEG 2000(jpf)**		**Format: JPEG[2]**
	Uncompressed	**Lossless**	**Lossless**	**50%**	**JPEG**
Disclosure	Good	Good	Good	Good	Good
Adaption	Wide Adoption. Negligible support in browsers.	Wide Adoption. Negligible support in browsers.	Moderate Adoption. Negligible support in browsers.	Moderate Adoption. Negligible support in browsers.	Wide Adoption
Self-documentation	Acceptable	Acceptable	Good	Good	Acceptable
Impact of Patents	No Impact	Low Impact	Little or No Impact	Little or No Impact	No Impact
Technical Protection Mechanisms	No Impact	No Impact	No Impact	No Impact	No Impact
Availability of tools	Wide Availability	Wide Availability	Limited to Moderate Availability	Limited to Moderate Availability	Wide Availability
Dependencies (hardware& Software)	No dependencies	No dependencies	No dependencies	No dependencies	No dependencies

We adapted the second method but eliminating the quality and functionality factors. The results on the table highlight that both file formats can be used as digital long term preservation modes but more acceptable is the TIFF format due to the limited or moderate availability of tools. A number of important European Institutions, like National Archives of United Kingdom from 2013 have decided to use the JPEG 2000 format, as their preferred file format for long term preservation.

[2] This will be used only as access copy

Conclusions

The results reveal the situation that for deciding the file formats that will be used for long term preservation is necessary to consider same physical aspects of documents (size, original material, physical conditions) and reason why we need to preserve them. Both file formats TIFF LWZ and JPEG 2000 Lossless met all technical characteristics of clarity and quality and can be chosen as file formats for long term preservation.

References

1. Digital Preservation Coalition (2008) "Introduction: Definitions and Concepts". *Digital Preservation Handbook*. York, UK. Last retrieved 29 August 2015.
2. Puglia, S., Reed, J., Rhodes, E., U.S. National Archives and Records Administration (2010) "Technical Guidelines for Digitizing Archival Records for Electronic Access: Creation of Produciton Master Files – Raster Images".
http://www.archives.gov/preservation/technical/guidelines.pdf. Last retrieved 05.09.2015.
3. Topi, A., Xhuvani, A. (2015). E-archiving Architecture and Actual Challenges in Albania, International Conference on Innovation Technologies IN-TECH 2015, September 2015, Croatia, 391-394.
4. Levigo solutions GmbH (2011), "Reproducibility of Archived Documents" http://www.pdfa.org/organization/levigo-solutions-gmbh; Last retrieved 04 October 2015.
5. https://en.wikipedia.org/wiki/File_format. Last retrieved 2 September 2015.
6. Arms, A., Fleischhauer, C., (2005): "Digital Formats Factors for Sustainability, Functionality and quality". IS&T Archiving 2005 Conference, Washington, D.C.
7. Murphy P. E., Rochester Institute of Technology (2002): "A Testing Procedure to Characterize Color and Spatial Quality of Digital Cameras Used to Image Cultural Heritage"; http://www.art-si.org/PDFs/Metric/EPMurphyThesis05.pdf. Last retrieved 16 September 2015.
8. Part 1 from the TIFF 6.0 specs: http://partners.adobe.com/public/developer/en/tiff/. Last retrieved 22 September 2015.
9. http://www.jpeg.org; Last retrieved 18 September 2015.
10. Kenney, Anne R.; Chapman, Stephen (1995): Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality.
http://www.clir.org/pubs/reports/pub53/pub53.pdf. Last retrieved 14.09.2015.
11. Franziska, S., Frey, J., & Reilly M. (2006). Digital Imaging for Photographic Collections.
12. Van Wijk, C. & Rog, J. (2007). "Evaluating File Formats for Long-Term Preservation." Presentation at International Conference on Digital Preservation, Beijing, China, Oct 11–12. http://ipres.las.ac.cn/pdf/Caroline-iPRES2007-11-12oct_CW.pdf (accessed September 21, 2015).