DOI: 10.2352/issn.2168-3204.2016.1.0.87

Long-term Preservation and Archival File Formats: Concepts and solutions

Peter Fornaro and Lukas Rosenthaler, Digital Humanities Lab, University of Basel, Switzerland

Abstract

In this paper we focus on the long-term stability of existing image file formats and possible new standards for archival purposes. It will be shown what significance digital file formats in a digital preservation workflow have and how they affect the success of digital archiving. Besides basic format requirements, like following an open standard and being widely used, we will discuss technical details of existing format definitions and their ability to encapsulate metadata for successful preservation. Based on the well-known Tagged Image File Format (TIFF) we will show in detail what weaknesses exist, that endanger the correct future rendering of the content. As a result a recommendation for an image file format definition for archival needs is proposed, based on the already existing widespread and accepted standard TIFF. The proposed approach follows the concept of a limited subset of a given file format like it is done with the Portable Document Format (PDF) and its archival derivative PDF/A. The approach results in a recommended use of the TIFF baseline specification. We call the optimized application of the existing file format TI/A, Tagged Image for Archives.

Motivation

In the "analog" past, the longevity of documents, be it texts, images or moving images etc. has been defined basically by the properties of the medium or the stability of the support the information has been stored on. While text written on parchment has a longevity way beyond 1000 years, it is well known that for example photographic material starts to deteriorate after approximately 150 years for b/w or 50 years for chromogenic color material. An alleged solution is to copy the original, be it an image or text, periodically to a new carrier to increase its lifetime. Unfortunately every analog copy process introduces new loss because the transfer to a new medium cannot be done without physical limitations, like reduction in sharpness or color fidelity. Eventually, after some generations, the original information gets lost because every transfer process introduces new errors. In recent times the digital age promises a theoretically unlimited longevity of digital information: A properly done digital copy process can be achieved with zero loss. Based on that, the digital age promises for the first time the unlimited preservation of information. Of course the limited lifecycle of hard- and software demands periodic migration, the major cost driver in archiving digital assets. However, today it is inevitable to digitize analogue images because the originals are endangered by the described physical decay. Additionally access to analogue objects is in most cases limited, because required dark, cold and dry storage is contrary to easy access. Such storage conditions are not required with a digital object and various on-going open access initiatives [1] demonstrate the importance of accessibility in the digital domain. Besides other motivations, those aspects are important considerations in the decision to digitize cultural assets. Because of the continuous process of scanning, digital images are an important part of our cultural heritage already and they account for a constitutive part of our contemporary multimedia output in social, scientific and economic ambits [2]. If a transition to the digital domain is carried out it is obvious that the maximum possible quality shall be captured and stored in the most sustainable way, in a digital image file.

We consider long-term preservation of digital images in the context of GLAM institutions (galleries, libraries, archives and museums) as the task to be able to render the image data in a correct way well beyond the timeframe of 10 years, possibly even an infinite period of time.

Problem

Despite the fact that digital data could be essentially preserved forever, there are some major hurdles to overcome. On one hand, all digital data carriers have a very limited lifespan due to technical obsolescence and/or unstable materials – they are on a physical level analog. This problem can be solved by a timely periodical migration onto new data carriers [3]. This process is known as bit-stream preservation. On the other hand, the file formats, which are used to represent the digital data (e. g. digital images), may also become obsolete over time. Any file format basically defines the logical structure and the meaning of the bits within the bit-stream and is thus essential for correct interpretation and proper rendering of the coded data. Both, the data carrier migration and the format migration are addressed in the OAIS reference model. Unfortunately, a file format or parts of its logical structure and definition can become obsolete. As a result the information renders useless, even though the bit-stream is still properly readable. A format migration is more complex than creating a plain copy of a bit-stream, by copying it to a new data carrier. The existing file format – the logic structure and its content – must be read, translated and written to a new data container. In such a process e.g. important metadata can be easily lost, due to improper transformation into new code, which leads to files that cannot be rendered correctly. Besides that every copy process can reduce image quality or introduce artifacts. Therefore it is necessary to use a stable and proper file format for long-term preservation of digital data [4]. The stability of a format is determined by the following criteria:

- The format has to be well documented
- The format should be in wide used
- The format should not contain proprietary or patented elements (algorithms etc.) and it should be an open format
- From a technical point of view the format should be as simple as possible
- No inter-linkage with external data is allowed (e.g. fonts or any other resource must be integrated)

- In addition to that it must be given, that the file-format is capable to store the relevant information without significant compromises. In the case of image files this could for example be the quantization depth for correct tonal reproduction or an appropriate color space

For long-term image preservation, usually the digital image with the highest resolution and largest tonal representation forms the base. At the end of the process of defining the archival master, the image data must be represented in a format that conforms to the requirements of long term preservation as described above. The goal is that a future format migration can be delayed as long as possible. Such a sustainable format is even more important if a very stable and technology independent data carrier, like Peviar/Monolith[5] or any other migration free data carrier is used.

Approach

Within this proposal we clearly limit the scope to images as they typically arise in the context of GLAM[1] institutions and in the field of the preservation of cultural heritage. Typically, these are:

- Digital reproductions of manuscripts, prints, etc.
- Digital reproduction of paintings
- Digital reproductions of photographs
- Photographic documentations of 3D objects such as sculpture or architecture.

One of the most widespread formats used to represent high quality image data is the TIF format. TIFF is a well known, established, flexible, adaptable file format [6] for handling images and data within a single file, by including various header tags for size, channel definition, image-data arrangement, applied image compression and others that define the technical configuration of the image. The flexibility of TIFF allows for many different variants and can also include metadata, which follows other format definitions such as IPTC-data, EXIF-data or ICC-data for color transformation [7]. It also allows proprietary elements and supports many different compression schemes, like LZW or even JPEG. There are some options within the TIFF standard that are rarely used and not supported by most applications. Furthermore some applications remove metadata that is not used by the application without notice of the user.

In most reproduction workflows in museums, libraries and archives TIFF is the final format for digital masters. At a first glance it has optimal features for digital preservation and it supports highest quality demands, like high tonal resolution. The basic technology of the format is rather simple and robust, it is widespread and it has the aura of being “professional”. Last but not least there is a lack of proper and acceptable alternatives.

However, at a closer look it can be seen that TIFF is quite complex and parts of the original definition have even become obsolete today, while new not formally standardized additions have been made within the format over the last couple of years. It would be easily possible to create a TIFF file that conforms to the TIFF Revision 6.0 specifications but would be virtually useless because no existing software is able to open and render it[2]. Since a digital archive should consist of files that can be used in the long term, a simplistic approach is necessary. Therefor a TIFF suitable for sustainable archiving should use only the minimal set of tags that is necessary to allow a correct future rendering of the data and to represent the essential descriptive metadata. We therefore propose a subset of the functionality of TIFF that is fully compatible with the de-facto TIFF standard itself but marks some tags as **mandatory**, some as **optional** and some as **forbidden** in order to guarantee the correct rendering in the future. In addition to the core functionalities, it is crucial to define a minimal set of metadata for archival applications, following standards like Dublin Core or METS. Such an approach is very similar to the well-known PDF/A [8] and its relative the PDF/A format for archival purposes. PDF is a very versatile and capable format but its flexibility and the large number of possible features turns it into a format that is hard to manage in an archival environment. As a consequence Adobe has developed PDF/A. PDF/A is a subset of PDF that also features mandatory, optional and forbidden components, to make it compatible with the needs of an archive. In analogy to the PDF/A format we propose to call the recommendation for the use of TIFF in archival environments *TI/A* or *Tagged Image for Archives*.

Why such a new recommendation for a rather old format that might be obsolete and replaced by a more modern format. The major motivation is not the future of image files; it is the waste of large amounts of data that are already existing and stored in archives. It is important to have a clear list of facts that says if such existing TIFFs must be migrated or not. It is a possibility to prove the quality of digital archival assets that goes beyond the test of data integrity and technical correctness [9]. To migrate large existing digital assets consisting of TIFF files would mean a significant financial effort that cannot be spend by most intuitions.

In cooperation with the University of Girona in Spain and EASY INNOVA, a technology and innovation center at Girona, we have started the process of specifying TI/A in co-operation with several memory institutions in Switzerland and Europe, namely the Center for Coordination for Digital Preservation and Archiving in Switzerland (kost-ceco) and the Swiss National Museum (SNM). □

Of course the concept of using a subset of the functionality of TIFF can be applied to any other format common for archiving digital image data or even video or motion picture.

The definition of such a recommendation for the use of TIFF in archives needs several steps:

- It is important to overview the ways how memory institutions store their data in order to find commonly used features. This is important because the requirements from a technological and a practical point of view do not match necessarily. E.g. the storage of only bi-tonal images might be unacceptable from a technological perspective but very common in archives.

[1] GLAM: Galleries, Libraries, Archives and Museums

[2] It would be possible but quite time consuming and difficult to develop a new renderer that would open *any* conforming TIFF file.

DOI: 10.2352/issn.2168-3204.2016.1.0.87

- A community of experts must discuss the importance and the value of possible features that are recommended for using TIFFs in archives.
- Such a recommendation must be standardized and promulgated to the community of museums and archives to have a broad impact.
- Ideally different software tools are developed to check the conformity of existing TIFF files in archives [10][11]

Results

Since a digital archive achieves to preserve its files in order to use them as long as possible, a simplistic approach is necessary, or to paraphrase Albert Einstein: "*Everything should be made as simple as possible, but not simpler*" That is, a TIFF file suitable for long term archiving should use only the minimal set of tags that is necessary to allow a correct future rendering of the data and to represent the essential descriptive metadata.

In a first step, based on feedback from experienced users and numerous tests, we have drafted a specification of a possible subset of mandatory, optional or forbidden features (tags) of the existing TIFF. This can be seen as a starting point for a recommendation that must be merged with important results from surveys that analyze existing assets of TIFFs in museums and archives. This knowledge is very important to address all existing data that must be checked for its archival quality. A second step involves an intense use of the community webpage (http://ti-a.org) [12] to have an intensive interaction with experts and users that generates a broad input from people working with digital images. The exchange of needs, requirements, dos and don't's will lead us to a final draft specification of an ideal archival file format for high quality image data that is based on the original TIFF standard and well supported by an international network of experts. Those two sources of knowledge help us to find an ideal set of features that address already existing TIFF assets of archives as well as technical issues. Following the original standard definition of TIFF allows us to define a format recommendation that is fully compatible with existing decoders. This approach helps to avoid that properly done "out on the market" software has to be modified or enhanced by any means.

The TIFF reference manual distinguishes between the *Baseline TIFF* and *TIFF-Extensions*. The "Baseline TIFF is the core of TIFF, the essentials that all mainstream TIFF developers should support in their products", as the TIFF documentation states. This means that the TIFF-Extensions are not expected to be supported by all developers. Thus, it must be discussed if TIFF-Extensions should not be used for long term archival.

Metadata and the position of the image data within a file are encoded through the use of 16-bit id's, the so-called *tags*. These tags are pooled into image file directories (IFD). A TIFF-file may contain multiple IFD's where each describes a variant of the image (e.g. a thumbnail image in B/W and a full resolution color image may be combined in one TIFF file). A tag in an IFD occupies 12 bytes according to the TIFF reference documentation. Within an IFD, the tags have to be sorted in ascending order to the tag id. Thus, a TIFF file may have a rather complex structure.

It is important to note that the same image may be represented by several TIFF files that differ on a binary level. But all these variants may be completely valid regarding the TIFF reference documentation and render to the identical (analog) image. In this sense these TIFFs have to be considered identical, even if they differ massively on a binary level.

The TIFF standard allows explicitly the use of private tags, that is, tags that are defined and used only within a special community or from a specific company. Tags numbered 32768 or higher are reserved for that purpose. Private tags might not be forbidden but they are just ignored.

Based on the feedback of numerous experts and the results of surveys we will define a list of **mandatory**, **optional** and **forbidden tags.** Such a tag list will look as follows (only a very limited number of tags are shown as example).

Tag ID	Name	Value	Recommendation
315	Artist	Artist or DC Creator	optional
265	Cell Length	Used for dithering	forbidden
320	ColorMap	Colormap for palette color images	forbidden Palette color images not allowed
33432	Copyright	Copyright notice	optional
306	DateTime	Date and Time of digital image creation	optional
388	ExtraSample	Additional components for image like masking etc.	forbidden

Table 1: An example of a list of tags with recommendations for their use.

Up to date it is not possible to define a finished list of tags and their relevance. What can be said is:

- TIFF files can contain image data with JPEG compression. Due to the fact that JPEG can lead to artifacts, it is recommended not to use JPEG within TIFF. Besides the problem of visible compression issues it is not recommended to use JPEG in TIFF due to the misleading situation of a file format containing a compression scheme (tag 259) that is well known as a file format by itself.
- A non-standardized quantization depth (tag 258) is not recommended. From a technical point of view 1 or 4 bits per sample are also not recommended. However, it can be expected that such files are common in existing archives and therefore this feature must be covered to prevent an unnecessary data migration.
- Non-standardized color spaces must be forbidden. Here we have the same problem. From a technical point of view CMYK color spaces are not necessary for proper image rendering but they might be common in archives.

DOI: 10.2352/issn.2168-3204.2016.1.0.87

- Large files with a size > 1GByte can cause problems while decoding. It is not recommended to store such large files

Based on that preliminary work we will define a recommendation for the proper and safe use of TIFF in archives. It will be tried to have the recommendation document standardized by the International Standard Organization, ISO. Such a precise definition of the functionalities and their implementation in a Tagged Image file for Archives will help to increase the sustainability of the original image format drastically. In addition such a recommendation can be used in a conformance checker to have existing digital archival assets examined regarding their archival quality.

Conclusions

TIFF is a flexible file format that can be used in many different ways. The many variants of TIFF that can be generated result in a broad variety of such files that constitute the assets of archives. It is not the aim of this initiative to create a new image file standard, since it is fully based on an existing standardized file format. Thus any existing conform modern reader will be able to open TI/A files and render them correctly without problems. TI/A can be regarded as a quality label for TIFF in archival environments. However, since TIFF is very complex and offers many options, a *subset* of the different specification has to be defined for archiving. Thus the original standards are enforced with some stricter rules on how the files must be constructed for archiving. In this sense, the recommended approach is no new file format but a specific version of the original TIFF standard that is suitable for long term archiving.

It is not possible to define in this publication a final specification for TI/A but it addresses the problematic issues of TIFF and it is an invitation to become a member of the community and to discuss the different features of the format regarding its archival compatibility.

Of course the problem of partially proprietary file structures is also true for other image file formats and content like video or motion picture. One of the most important and promising image file formats is JPEG2000 [13, 14]. JPEG2000 is e.g. used in motion picture as Digital Cinema Package, DCP [15]. The above approach to limit the functionalities of TIFF can be applied to any other format, leading to archival definitions of already existing file formats like JPEG2000/A or DCP/A [16].

References

[1] HEATH T., BIZER Ch., Linked Data: Evolving the Web into a Global Data Space (2011), Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool

[2] KUNY, T. 1998. A digital dark ages? Challenges in the preservation of electronic information. *Int. Preserv. News*, 8–13.

[3] ROTHENBERG, J. 1995. Ensuring the longevity of digital documents. *Sci. Amer. 272*, 1, 42–47.

[4] GUBLER, D., ROSENTHALER, L., AND FORNARO, P. 2006. The obsolescence of migration: Long-Term storage of digital code on stable optical media. In *Proceedings of IS&T's Archiving Conference*. IS&T, 135–139.

[5] MÜLLER Florian, FORNARO Peter, ROSENTHALER Lukas, GSCHWIND Rudolf (2010), PEVIAR: Digital Originals ACM Journal on Computing and Cultural Heritage, Volume 3, Issue 1. ACM 2010

[6] GOETHALS, A, General Considerations for Choosing File Formats, Harvard University Library Last modified: 07/31/09

[7] LOEFFLER, H., (2007), Baranger, Walt, ed., *Photo Metadata White Paper 2007*, IPTC |first2= missing |last2= in Editors list (help). The white paper discusses upcoming changes to the IPTC Photo Metadata Standards

[8] PDF/A 101: An Introduction – presentation from the First International PDF/A Conference in Amsterdam

[9] JHOVE Validator; http://jhove.openpreservation.org/getting-started/

[10] KOST TIFF-Validator, http://kost-ceco.ch/cms/index.php?tiff_val_de

[11] Performa DPF Manager, http://www.preforma-project.eu/dpf-manager.html

[12] Tagged Image for Archives Standard Initiative, http://ti-a.org, September 2015

[13]WITTUS, R. W., JANOSKY, J. S. Using JPEG2000 for Enhanced Preservation and Web Access of Digital Archives – A Case Study

[14] BUCKLEY Robert, Using Lossy JPEG2000 Compression For Archival Master Files, Prepared for the Library of Congress Office of Strategic Initiatives, Version 1.1 March 12, 2013

[15] The Digital Cinema Initiatives, LLC, http://www.dcimovies.com/

[16] FORNARO, P., GUBLER D., DCP/A: Discussion of an Archival Digital Cinema Package for AV-Media, IS&T Archiving Conference Proceedings, Berlin, 2014

Biography

Peter Fornaro is deputy head of the Digital Humanities Lab of the University of Basel. He is doing research in the field of digital archiving, imaging, cultural heritage preservation and computational photography. Fornaro is teaching at the University of Basel. Besides research and lecturing he is giving consulting to companies, archives and museums. Fornaro is member of the Swiss Commission for Cultural Heritage Preservation (EKKGS).

Lukas Rosenthaler is the head of the Digital Humanities Lab of the University of Basel as well as of the Swiss National Data Curation Center for the Humanities. He is an expert for data base systems, virtual research environments, image processing and digital archiving and he is supporting open access initiatives.