https://doi.org/10.2352/issn.2168-3204.2017.1.0.6

TIFF in Archives: A survey about existing files in memory institutions

Peter Fornaro, Lukas Rosenthaler, and Erwin Zbinden; Digital Humanities Lab, University of Basel; Martin Kaiser, KOST_CECO, Bern (Switzerland)

Abstract

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 for handling images and data within a single file. 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. Therefore TIFF is a complex file format that can be problematic for the use in archives, even though it is still the most common option for most GLAM institutions.

The aim of the TI/A initiative was to find a proper subset of tags for the use of TIFF in archival environments. To select proper features in such a recommendation, it was necessary to analyses existing files first. In this paper we present the results of two surveys that have been done in this context:

A) The analysis of about 4 million TIFF files stored as digital assets in memory institutions. The files represent a large variety of TIFF formats, regarding e.g. compression schemes, quantization depth and date of creation.
B) A survey about the number, use and relevance of digital files in archives, museums and libraries. The survey was done in the context of an ongoing project of the Swiss government to find a sustainable strategy for archiving digital cultural heritage objects.

Motivation

Digital files are part of most assets of galleries, libraries, archives and museums (GLAM institutions). One of the important media types are certainly image files, either created by retro digitization or born digital. For most media types some de facto standards exists. In the case of images in GLAM institutions this is the Tagged Image File Format (TIFF) [1]. TIFF is a well-known, robust, flexible image file format for handling image- and meta-data within a single file. The flexibility of TIFF allows also to include various types of metadata that are important for archival applications such as:

- IPTC - Metadata (International Press and Telecommunications Council): Contextual metadata to cover the needs in the field of photo-journalism and press-photography [2].
- EXIF - Metadata (EXchangeable Image File format): A metadata standard for technical information about the capturing device and the capturing process
- ICC - Profiles (International Color Consortium): Embedded information for the definition of color transforms and color spaces.

TIFF also supports different compression schemes, like LZW or even JPEG. Due to this robustness, its flexibility and simplicity TIFF is for many GLAM institutions the final format for digital masters. In addition TIFF also follows the basic principles of "good file formats" for archiving:

- The format itself is well documented
- The format is widely used
- The format does in principle not contain proprietary or patented elements (algorithms etc.)
- From a technical point of view the format is simple and robust

However: There are some options within the TIFF standard that are rarely used and not supported by typical applications. Furthermore some applications remove important meta information without notification or warning, so that the removal stays unrecognized until this specific meta information should be used for any application. On one hand TIFF certainly offers adequate features for digital preservation and it supports highest quality demands, like high tonal resolution. On the other hand it allows such a rich set of features and variability of specifications, that in detail two files might not be of the exact same format specification, just because one element of the workflow has not handelt the tags of the two files in the same way. Luckily the basic technology of the format is rather simple so that the format can be checked for consistency and correctness but this process must be executed, which is n.

From the point of preservation the file format is of major importance [3][4]. Todays approach of digital data migration to overcome obsolescence works very well in the case that hardware technology that is becoming out-dated. This problem can be solved by a systematic copy process to migrate data onto a new data carrier [5]; a process that is called bit-stream preservation. In the case of the file formats, that is the logical structure that encapsulates digital data (e. g. digital images), obsolete is a much more demanding problem [6]. A file format defines the meaning of the bits within the bit-stream and is thus essential for correct interpretation and proper rendering of the coded data. A format migration is a much more complex process than creating a plain copy of a bit-stream, by copying it to a new data carrier [7]. A format migration can easily result in the loss of important metadata, due to improper transformation into new code. The success of bit-stream preservation is simple to verify, eg by the creation and comparison of a hash code of the source and the destination file. In the case of a format migration this is not possible because the two hashes will be different by definition. Therefore it is necessary to verify the stability of a file format for long-term preservation of digital data.

Within the European PERFORMA project a software has been developed to check the quality of existing TIF files [8]. The software allows full customization of the features that shall be checked. The

project is a collaboration of the Digital Humanities Lab of the University of Basel and the University of Girona in Spain.

Problem

In this project one of the aims was to specify a subset of features (allowed and recommended tags) for TIFF for archiving. We therefore proposed a subset of the functionality of TIFF that is fully compatible with the de-facto TIFF standard itself but marks some tags as **required**, some as **optional** and some as **problematic** 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 and its "relative" the PDF/A [9] format for archival purposes. With such a recommendation it can be ensured, that different institutions follow the same guideline while checking the files, e.g. with the software, developed in the project, "DPF Manager". Community building is also an important aspect in the context of digital preservation. Therefore the process of the definition of such a recommendation has been started with an initiative, called TI/A (Tagged Image for Archives). TI/A offers a web platform for discussions of experts and it initiated a standardization process for the use of TIFF in archives that shall be accredited by the International Standard Organization (ISO). The initiative addresses two different cases:

1) Existing files that need to be monitored to find obsolete "tags" within the file, so that a migration can be started early enough.
2) Support in the selection of proper features if new files are written. This later case is also important for the definition of a sustainable media standard.

As a starting point for such a standardization process, it is necessary to get a proper image of the real situation out there. If we speak for example about functionalities of an image file format, e.g. the right choice of image resolution or other quality aspects, there is nothing like one or an absolute truth. It is in all cases a question of the application and the future use of the digital image files. Sometimes it is even a question of in-house policies or best-practices adopted from other institutions. There are certainly technical aspects that are necessary from today's point of view but it can be expected that they changed already and will change with time. Certain features of a file format might have been a requirement a couple of years ago but changed with the advance of technology, so that they became obsolete today. In the TI/A initiative it was clear that taking into account existing files and their technical specification is very important, because the initiative addresses mainly such files that are already existing in archives and need to be checked for problematic technical aspects. Problematic means, that files do make use of tags that render them useless in future. Technical specifications that seem to be absolutely necessary today, maybe were not of big importance some years ago One reason because storage space was much more expensive. Such an alert then starts further processes, like a migration from the "bad" TIFFs to a "correct" ones. To find the definition of correctness we focussed on two sources of knowledge:

1) The average of the opinion of the experts in this field, gained from the discussions on the web-platform.
2) The result of an analysis of existing TIF-files in GLAM institutions.

The analysis of the existing files is then merged with the input of the experts in the field of digital preservation to find the best possible recommendation. Unfortunately it is not easy to get access to existing archival assets nor is there a straightforward way to analyze the tags used in TIF files. The solution was to test some TIFF tag extraction programs, including a solution that was developed in-house that analyses all possible tags of existing TIFFs and stores the results as a list of tags that can be processed to get a so called "feature histogram".

A second aspect addresses the purpose of digital files in archives. It makes a big difference if files are used as digital masters that replace the originals or if they are just dissemination copies. To find answers to those questions, a survey was needed that represents a wide variety of memory institutions and the application of their digital assets.

Approach

To be able to get access to real image data we worked in close collaboration with KOST-CECO. This group of experts is operating a platform to bring together knowledge from archives and experts in archiving to return it as best-practices to archives and any other institution that stores digital assets that needs to be preserved. KOST-CECO is for example offering a catalogue of file formats and the evaluation of them for archival purposes. By the help of and in cooperation with KOST-CECO we were able to get access to assets of large memory institutions in Switzerland:

- The Swiss National Archives
- Staat-Archive Basel-Stadt
- Staat-Archive St. Gallen

Due to the fact, that we wanted to analyze "hot data", we had to copy the assets to an independent infrastructure within the archives. On this infrastructure, simply consisting of a NAS storage and a linux workstation, we started a number of already existing tools to get as much information from the image data files as possible. In such a way we gathered the information about the image files plus the information about the differences in file analysis of the various tools in the same process. We used the following tools for the file analysis:

- **Jhove [10]**
  Jhove is a common software for the analysis of multiple types of files. If is a format-specific digital object validation API written in Java.

- **checkit_tiff**
  The tool evaluates TIFFs based on "rules" defined in a configuration profile. The configuration profile is human readable.

- **exiftool**
  ExifTool is a platform-independent Perl library plus a command-

line application for reading, writing and editing meta information in a wide variety of files.

- **exiv2**
  Exiv2 is a C++ library and a command line utility to manage image metadata. It provides fast and easy read and write access to the Exif, IPTC and XMP metadata of digital images in various formats.

- **DPF Manager**
  DPF Manager is a advanced TIFF conformance checker for digital preservation. It is the software that is developed in the PERFORMA project.

- **Tiffhist**
  Tiffhist is a software developed by the Digital Humanities Lab of the University of Basel. It simply reads all tags in TIFFs and writes all the information in a log file.

The output of the programs has been stored in log files for later analysis (see below). Due to the size of the assets processed, the whole analysis needed a couple of weeks to be finished. The most detailed information about the files is delivered by the Tiffhist program. This is a small footprint command line program that scans sequentially all possible tags in TIF files. If a tag number is found, the tag number and its values are stored in a text file. Like this we get full information about the content and the technical specification of the files.

Besides the technical analysis of the tags we did a survey about the purpose and the size of the digital assets in about 110 memory institutions in Switzerland (244 were asked). The survey has been done web-based where 12 simple question have been asked, addressing the following aspects:

A. Size of the asset
B. File formats
C. Existence and use of a media standard (File Format specification
D. Location of the assets
E. Preservation strategy
F. Relevance of the assets (e. g. replacing the original)
G. Use of standards (national and international)

The digital domain has become increasingly important in recent years, so that many processes without digital data would be unthinkable in today's society. This development affects not only largely all business processes, archives and museums, but also the whole area of the protection of the cultural heritage and the preservation of monuments. The long-term and sustained preservation of digital objects requires new strategies and methods, which are very different in comparison to the preservation of physical objects. In the context of the revision of the *inventory for particularly protected cultural goods*, the question arises as such digital inventories should be included.
In order to be able to derive correct and effective measures, a detailed knowledge of the initial situation is important. For this purpose, the *Federal Commission for the Protection of Cultural Heritage* (EKKGS) compiled a catalog of questions and did a survey with Swiss institutions. This survey serves as the basis and analysis of the current situation within the framework of the Swiss Cultural Heritage. With the help of the survey an overview of the data landscape in Swiss archives, museums and libraries is to be created. The following questions are fundamental: What is the significance of digital data within collections? Is the data storage dynamic and how is the digital objects worked? What knowledge can be learned about the infrastructure used to store / use digital collections?
Depending on the results of the survey, an extension of the inventory is planned for digital objects under special protection. Currently, the inventory includes physical objects such as buildings, archaeological objects, collections, or parts thereof. In a subsequent step, practical protection measures for "digital cultural goods" are to be derived. The survey also provides an important overview over the type images data files and the formats used. Therefore we can use the results of the survey for the work with TI/A and the guidelines for the use of this file format in archival environments.

Results

TAG Analysis

The Tiffhist program produced an output of 566 log-files each containing roughly 7'000 to 10'000 TIFF-tag datasets. The tiffhist datasets (4'065'951 in total) were extracted from the log-files and after some data validation written to a database and could be analyzed. A wide variety of tags where evaluated, not only the TIFF revision 6.0 baseline and extension tags, but also the tags defined by the ISO-Standard TIFF/EP and EXIF- and private tags. This allowed an analysis of real world usage of TIFF tagged images in existing archives.

Is is no surprise that most of the files represented either black and white or grayscale (3'343'776) images. Files representing RGB, Palette or YCC colors were less often found (RGB images: 664928, Palette images: 55916, YCC images: 1202, CMYK images: 103. There were no images with a transparency mask nor with CIELab, ICCLab or ITULAB color).

The results showed that most of the images were 1 bit (black and white) images (2'082'010), many images were 8 bit (1'871'169), and there are also some 16 bit images (81'231).

It also turned out that a biggest part of the images where uncompressed (2'065'958). Surprisingly almost the other half of images were Group 4 Fax compressed (1'866'963). LZW compression unexpectedly had only 83'312 shows. Group 3 Fax compression was found 20'041 times. CCITT 1D compression was found rarely (564). New style JPEG compression was found 27'520 times, whereas old style JPEG compression as defined in TIFF rev. 6.0, was not found at all. The same applies for PackBit compression which was not found.

Quite interestingly the tag TIFF/EPStandardID was not found at all. An ICC profile was found in 158’125 images.

At the beginning digitization was performed with scanners, later on digital cameras where used. With this development the use of EXIF tags became widespread.

This is only a short overview of the most important questions, which mostly resulted from an analysis of the TIFF baseline required tags 258 (BitsPerSample), 259 (Compression) and 262 (PhotometricInterpretation).

The results of the TIFF analysis were interesting. The best practices and the technical possibilities have fundamentally changed

over the last 25 years. While today 16bit RGB images are regarded as standard, the early files often have been stored bi-tonal. However compression schemes like LZW or JPEG were rather rarely used.

Survey

The survey about the purpose of digital files in GLAM institutions was also interesting. The following results could be gained from the survey:

Question 1: Are digital data (primary sources, digital reproductions, photographs ...) a significant part of the collection of your institution?

Approximately 65%, and thus the majority of respondents, indicate that digital data is a significant part of the collection. For 30.36%, digital data hardly plays a role, 4.46% have neither affirmed nor denied the question.

Question 2: How big is the digital part of your collection?

A majority of 57.27% hold a digital stock, which comprises more than 10,000 properties. Approximately 18% of the institutions own more than 1,000 properties, some 11% own more than 100 properties. The size of the population is not known for 13.64% of respondents.

Question 3: For what reasons does your institution create, use, or store digital data?

In this question, it should be borne in mind that the respondents had the opportunity to select several answers. The reasons for working with digital data are very diverse. For most institutions, digital data are an important aspect in the context of mediation (89.91%) as well as further processing in the context of scientific research and projects (79.82%). However, none of the possible answer options could be emphasized by a clear majority. Similarly, the replacement of the original (65.14%), the support of the inventory (66.97%) and the documentation of restoration work (48.62%) represent important processes which would be difficult to implement without the use of digital data.

Question 4: Where are the digital data stored?

In this question, it should be borne in mind that the respondents had the opportunity to select several answers. With a clear majority, the digital objects are stored internally in the institution in about 90% of respondents. At 20.72% of the institutions, the data are stored on an external, national server provider. In this case, it must be borne in mind that the digital objects are stored both internally and externally at some facilities (this applies to 23 institutions). Only 7.21% use a cloud service to store the data. In 2.70% of the institutions, the location is unknown.

Question 5: Who is responsible for the creation of the digital images?

In this question, it was also possible to specify several answers.

Most institutions, around 91%, produce the digital objects themselves. Approximately half of the respondents (51.35%) refer to external service providers, although here too it is important to note that the institutions themselves and external service providers are involved in the production of the digital data (this applies to 48 institutions). 1.80% of the interviewees are not aware of the responsibility for the creation.

Question 6: Is there a media standard in which the individual file formats and their properties are specified in detail?

More than half of respondents (55.45%) have a media standard with regard to digital archiving, in which the file formats and their properties are described in detail. Approximately 33% have no standard and around 12% of respondents are unclear.

Question 7: Which file formats are used to store the digital image data?

In this question, it should be borne in mind that the respondents had the opportunity to select several answers. Around 86%, and thus the majority of the institutions, use the file format TIFF for storing digital objects. 78.38% use the picture format JPG. It is striking that most institutions use both TIFF and JPG for storage (this applies to 74 institutions). Nearly 30% of the interviewees use other storage formats, which were not given as an answer to this question. It is to be assumed that in the collections of some institutions in addition to pictures may be synonymous videos or sound recordings in digital form. 17.12% use JPG2000 format, RAW with 15.32% and PNG with about 9%. For 2.70% of the interviewees the used storage formats are unknown.

Question 8: Is the process of digital archiving based on national or international standards and recommendations?

Approximately 51% of the institutions are geared towards digital archiving according to their own guidelines. 31.82% are based on international standards and recommendations, 17.27% of respondents follow recommended procedures at national level.

Conclusions

As the questionnaire clearly shows, digital images play an important role in today's GLAM institutions. The survey shows also, that most of the institutions use TIFF, many of them as replacement for the original. The tag analysis shows the variability of the TIFF tag usage. Given the importance of the TIFF format in today's memory institutions it is of utmost importance that readability in the far future will be guaranteed.

References

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

[2] LOEFFLER, H., (2007), Baranger, Walt, ed., *Photo Metadata White Paper 2007*, IPTC: The white paper discusses upcoming changes to the IPTC Photo Metadata Standards

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

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

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

[6] FORNARO, P., ROSENTHALER, L. 2016. Long-term Preservation and Archival File Formats: Concepts and Solutions, In *Proceedings of IS&T's Archiving Conference*. IS&T.

[7] 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.
[8] Performa DPF Manager, http://www.preforma-project.eu/dpf-manager.html
[9] PDF/A 101: An Introduction – presentation from the First International PDF/A Conference in Amsterdam
[10] JHOVE Validator; http://jhove.openpreservation.org/getting-started/

Biography

Peter Fornaro is member of the management team 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 is member of the management team 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.

Erwin Zbinden is senior scientist at the Digital Humanities Lab of the University of Basel. He is expert for digital preservation of photographic material in the analog and digital domain.

Martin Kaiser is member of the KOST-CECO team. After studying linguistics and a postgraduate degree in computer science, he worked as a project manager in several projects in the field of information retrieval and document management.