Automated Preservation: The Case of Digital Raw Photographs

Stephan Bauer and Christoph Becker

Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at/~becker

Abstract. In digital preservation, a common approach for preservation actions is the migration to standardized formats. Full validation of the results of such conversion processes is required to ensure authenticity and trust. This process of *quality assurance* is a key obstacle to achieving scalability for large volumes of content. In this article, we address the quality assurance process for the preservation of born-digital photographs and validate conversions of raw image formats into standard formats such as Adobe Digital Negative. To achieve this, we rely on a systematic planning framework. We classify requirements that have to be evaluated according to their measurement needs. We extend an existing measurement framework using a combination of tools, image similarity algorithms, and purpose-built plugins. By combining metadata extraction, image rendering and comparison, and perceptual-level quality assurance, we evaluate the feasibility of automating the core part of quality assurance that is often the most costly part of preservation processes.

1 Introduction

In our society so strongly relying on digital data and communication for all purposes, digital preservation has become an important challenge: Only a functioning and appropriate technical environment can create a meaningful interpretation of digital objects. In the case of static content, where data structures specify the full semantics of the intellectual meaning of data streams, a common solution is to convert data into different representations to enable rendering in the environment of a certain user group. For preserving images, this corresponds to conversion to a standardised image format with long expected life time and widespread, solid tool support. These *migration* processes need to be fully validated to ensure that the intellectual meaning of the content is fully and authentically preserved. The migration itself can be relatively easily automated and parallelised to achieve scalability on large volumes of content. The key bottleneck lies in verifying the migration results.

Today, professional photographers take literally tens of thousands of pictures per month[1]. They normally rely on a so-called *raw* format that stores the raw

[1] http://www.nationalgeographic.com/educator-resources/presidents-photographer/window-into-white-house/

C. Xing, F. Crestani, and A. Rauber (Eds.): ICADL 2011, LNCS 7008, pp. 39–49, 2011.

data recorded by the camera's imaging sensor. The fundamental difference between the raw data and the interpreted data in a JPEG file is that the raw data itself needs interpretation in order to create a meaningful and presentable image, much like a negative film requires developing and printing to produce a photograph. Born-digital photographs captured in raw formats thus present a particular challenge, since the semantics of data captured by the sensors in digital cameras require complex interpretation processes to deliver a meaningful and authentic rendering. Without a trustworthy evaluation of different possible migration paths, no decision and conversion process can be trusted to deliver authentic and usable photographs. There is a vast amount of objects involved in large collections of raw camera data stored in proprietary formats containing specific camera profiles and the inherent uncertainties of migration. Hence, it is clear that such a conversion into standardised digital still image formats can only be successful with reliable and meaningful methods for automated Quality Assurance (QA). QA in this context refers to the process of measuring the significant properties of content and delivering measures of similarity according to specified criteria of interest in a structured form.

This article leverages a preservation planning framework to address this key issue in providing automation and scalability for the preservation of born-digital photographs. We elicit requirements that have to be evaluated and classify them according to their measurement needs. By extending an existing measurement framework using a combination of tools and purpose-built plugins for metadata extraction, image rendering and comparison, and perceptual-level quality assurance, we demonstrate automated QA for preserving digital photographs. We discuss experiment results and outline challenges lying ahead.

The remainder of this paper is organised as follows: Section 2 provides an overview of related work in image preservation and comparison. Section 3 discusses challenges and requirements posed by born-digital photographs. Section 4 demonstrates how to leverage an existing evaluation framework and extend it to achieve automated QA using existing software and new tools. The results of a case study are presented in Section 5. Finally, Section 6 draws conclusions and gives a short outlook on future work.

2 Related Work

In the digital libraries field, the discourse about preservation of images has largely focused on migration strategies converting scanned image content to standardised still formats such as TIFF-6 or JPEG 2000 [8]. A recent contribution discusses issues with JPEG 2000 as a preservation format [10]. Considering born-digital photographs, it must be noted that the term *raw format* does not describe one specific format. Each major vendor has multiple different types, most of which are partially proprietary and not standardized. To increase standardization, Adobe created an open raw format called Digital Negative (DNG).