Advanced Digital Image Preservation Data Management Architecture

Wo Chang; National Institute of Standards and Technology; Gaithersburg, Maryland, USA

Abstract

Digital data preservation is a difficult problem due to the high cost associated with the required technological development, rapid change and continuous anticipation of societal technology adaptation, and finally, the lack of interoperable standards for access and integration of diverse media content, software applications, and operating hardware systems over a decade-long timeline. While there is no single strategy to encompass all problems, the goal is to provide open preservation strategies between technologies. This paper presents a newly developed international standard technology called Multimedia Application Formats (MAF) from the ISO/IEC SC 29 WG 11 (MPEG) as the recommended digital image preservation data management architecture to migrate and preserve digital image content. This paper presents the implementation of such preservation strategy.

Disclaimers

NIST does not endorse or recommend any of the mentioned products, companies, or sites in this paper and such mentions do not imply that the cited products, companies, or sites are better or worse than similar products, companies, or sites.

1. Introduction

The ability to effectively preserve and retrieve born-digital data is increasingly crucial as digital technology continues to produce vast amounts of valuable and irreplaceable knowledge and information. In 2003 alone, items produced in magnetic media [1] exceeded five exabytes (10^{18}) of data, each exabyte equivalent to 50,000 times the entire U.S. Library of Congress' printed collection. Data ranges from personal collections of photos and videos from digital cameras and camcorders to scientific experimental data, national defense satellite images, and healthcare records. The big questions are: Will datasets generated in the 1980s be recognizable by software today? Will digital content created today be accessible and renderable 50 years from now? Digital data preservation and access to long-term retention archival knowledge and information impact all sectors of the industry, such as biomedical, healthcare-IT, manufacturing, libraries, publishers, scholarly societies, and the government, just to name a few.

This paper utilizes the newly developed ISO/IEC 23000 [2] (also known as "MPEG-A") framework as the data preservation strategy. MPEG-A is a recent addition to a sequence of standards that have been developed by the ISO/IEC SC 29 WG 11 Moving Picture Experts Group [3]. This new standard was developed to package existing technologies (audio-visual data content, file container and metadata) from all published MPEG standards and then combine them into the so-called "Multimedia Application Formats" or MAFs.

The outline of this paper is as follows: section 2 describes the problems of legacy image content and how to overcome them (convert them into a richer data format) while section 3 presents the ISO/IEC MAF framework and its technology components. Section 4 discusses the ADvance Multimedia Information REtrieval (ADMIRE) implementation approach and demonstrates the automatic conversion of legacy graphic file formats into the newer format so that searching, retrieving, and displaying converted images can be performed. Section 5 summarizes of the future work that can utilize the same preservation framework strategy to handle audio and video file formats.

2. Problems with Legacy Image Content

Legacy image content faces at least three major challenges: (a) lack of image content playback software, (b) increase in diversity of metadata description for each file format, and (c) lack of best practices for preservation of original quality of image content.

For digital images alone, over a hundred or even a thousand different types of graphic file formats have been created since the inception of the computer age. Each type stores graphics data in a different way. Bitmap, vector, and metafile formats are by far the most commonly used formats. Bitmap files, sometimes referred to as raster files (ex. Microsoft BMP[4], PCX[5], TIFF[6], and TGA[7]), essentially contain an exact pixel-by-pixel map of an image. Vector format files (ex. AutoCAD DXF[8] and Microsoft SYLK [9]) are useful for storing line-based elements, such as lines and polygons, or other simple geometric objects such as text, or some mathematical descriptions of image elements, rather than pixel values. Metafiles (ex. Macintosh PICT[10] and CGM[11]) can contain both bitmap and vector data in a single file. They provide a language or grammar that may be used to define vector data elements or store a bitmap representation of an image. Metafiles are frequently used to transport bitmap or vector data between hardware platforms and software applications. However, these irreplaceable valuable contents are facing obsolescence due to a lack of playback devices or players as technologies move forward.

There are also metadata problems associated with each graphic file format: (a) missing metadata description during time of image acquisition; (b) diversified metadata structures among different file formats, including home grown metadata and industry defined metadata; and (c) difficulty searching and retrieving image content due to a difference in semantic definitions between metadata schemas.

3. ISO/IEC Multimedia Application Formats

To overcome the problems of the diverse legacy file formats and their multiple metadata structures, one solution is to utilize the MPEG-A framework. This framework is designed to provide extensible metadata in the form of 'xml boxes' along with timed

media information for presentation in a flexible, extensible format that will be able to facilitate interchange, management, editing, and presentation of the media. The presentation content (ex. images) may be 'local' to the system, or may be remote from over a network or other streaming delivery mechanism. Currently, there are a number of MAFs developed within this new standard, ranging from Music Slideshow MAF to Broadcast Streaming MAF. The relevant MAF for digital images is the ISO/IEC 23000-3 Photo Player MAF (PPMAF) [12] that can be adopted for digital image preservation. The following subsections describe the technology components used within PPMAF.

ISO/IEC MPEG-7 Metadata

The MPEG-7 [13] standard, also known as "Multimedia Content Description Interface," aims at providing standardized core technologies allowing for description of audiovisual data content in multimedia environments. This technology is being designed by a range of experts including broadcasters, manufacturers, content creators, publishers, intellectual property rights managers, telecommunication service providers, academia, government, etc., to:

- Define a rich set of standardized tools to describe audiovisual content,
- Create good storage solutions, high-performance content identification, fast, accurate, personalized filtering, searching, and retrieval data structure/formats,
- Enable both human users and automatic systems to process the encoded audiovisual content descriptions.

Basically, MPEG-7 provides standardized metadata structure and its attributes/values definition to describe the audiovisual content, as shown in Figure 1.

Figure 1: *MPEG-7 Standard Content Description*

ISO/IEC MPEG-4 File Format

MPEG-4 File Format (MP4-FF) [14] is based on the ISO Base File Format [15]. It is an object-oriented file container that can be formed as a series of objects, called "boxes" as shown in Figure 2.

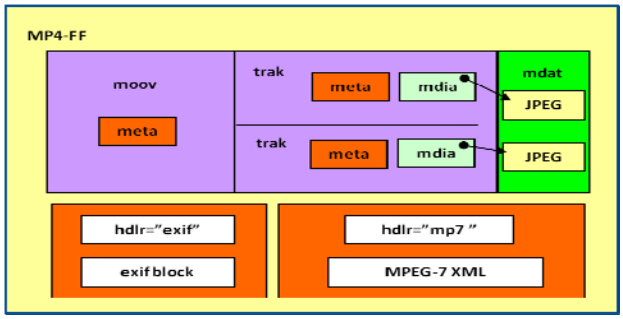

Figure 2*: MPEG-4 File Format (MP4-FF)*

All data are contained in boxes; there are no other data within the file. This includes any initial signature required by the specific file format. All object-structured files shall contain a file type 'ftyp' Box where:

"ftyp" box – a file level file type identifier that represents the version of an application.

"moov" box – a presentation box that stores metadata and related data tracks.

"meta" box – a box that contains descriptive or annotative metadata.

"trak" box – a container box for a single track of a presentation.

"mdat" box – a box that carries the actual image data.

MP4-FF directly supports MPEG-7 metadata that is stored under the "meta" box. When it is used, the handler-type can be either 'mp7t' for textual metadata in Unicode format or 'mp7b' for binary metadata compressed in the binary format. In this case, the binary XML box contains the configuration information immediately followed by the binarized XML. When the format is textual, there is either another box in the metadata container 'meta', called 'xml', which contains the textual MPEG-7 document, or there is a primary item box identifying the item containing the MPEG-7 XML. On the other hand, when the format is binary, there is either another box in the metadata container 'meta', called 'bxml', which contains the binary MPEG-7 document, or a primary item box identifying the item containing the MPEG-7 binarized XML. MP4-FF provides other powerful features such as audiovisual synchronization and content protection with access control, however, they will not be covered in this paper.

Image File Formats

Currently, PPMAF only supports the ISO/IEC 10918 JPEG file format because all photos taken from new digital cameras would provide a nice set of data acquisition metadata (see *Preservation of Metadata* below). Since the concept of packaging image content (regardless of what file format is used) and metadata into a rich file container is an essential strategy for digital data preservation management, the MAF framework would therefore be worth exploring.

4. ADMIRE Implementation Approach

The ADMIRE system utilizes the PPMAF concept by packaging converted legacy image content into the international standard of image file format along with their associated metadata into the MP4-FF as shown in Figure 3. One nice feature of PPMAF is that it provides embedded thumbnails plus the flexibility of storing the actual image content locally or externally via a URL. In this way, small portable devices can easily view image content without suffering much from large volume of high-resolution image content. The ultimate goal of ADMIRE is to place the packaged image content into a set of distributed archival repositories so that platform independent devices (PDAs, cell phones, desktops, etc.) can access them via network communication protocols and services. The following subsections will describe the detailed implementation approach.

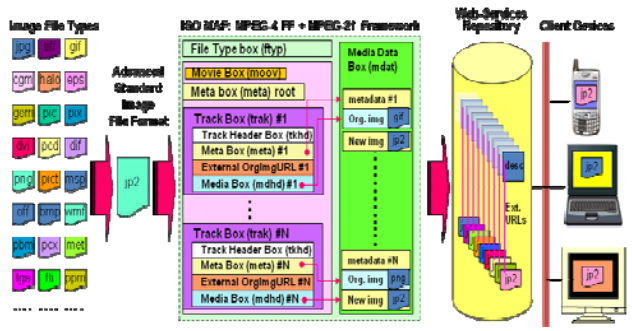

***Figure 3**: ADMIRE Digital Image Preservation Data Management Architecture*

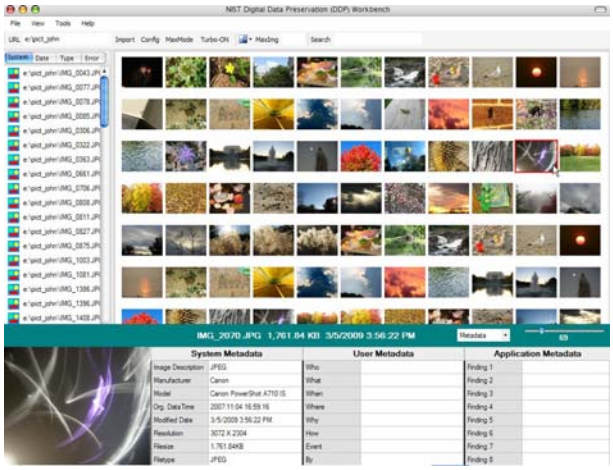

***Figure 4**: ADMIRE System*

Conversion of Legacy Graphic File Formats

Preserving the original image content is by far the most important task, and picking the right graphic file format for the image content to be converted into, is a critical matter. There are a number of popular graphic file formats that have emerged from lossy and lossless compressions as technologies advanced. The choice for preserving image content would be using lossless compression file formats so there will be no throwing away bits during the conversion process. Despite the fact that the JPEG file format provides high quality compression, is widely adopted from digital camera manufacturers, and greatly supported from many browsers and applications, due to its nature of lossy compression (which is different from the JPEG-LS (lossless) [16]), the JPEG image file format will not be suitable for preservation applications.

However, JPEG2000 [17], the successor of JPEG file format, which was developed in 2000, incorporates all operation modes of JPEG plus efficient lossless compression to form a unified compression architecture. JPEG2000 offers other nice features such as: (a) continuous-tone and bi-level compression of image components (e.g. R, G, or B) each from 1 to 16 bits deep; (b) progressive transmission by pixel quality (more data received, better quality of image) and resolution (more data received, image size increased); and (c) random code-stream access and processing which enable decompression of region of interest.

At the writing of this paper, while we are in the process of analyzing the quality of a variety of different graphic file formats (GIF, TIFF, JPEG, JPEG2000, etc.) for the ADMIRE implementation (see Figure 4). We just used the JPEG file format as a starting point so the rest of the ADMIRE system can be put together. ADMIRE is a Windows-based system using Visual Studio 2008 C++ platform. One of the major decision factors of using C++ (vs. Java and other languages) for the implementation of ADMIRE is because C++ libraries are widely available for file conversion, image processing tools, and codec compression tools compared to other programming languages. Furthermore, C++ is a platform-independent language and does not require any specific environment setup like Java Virtual Machine (JVM) in Java. Currently, ADMIRE utilizes the ImageMagick[18] C++ library as the primary conversion tool to convert legacy file formats into JPEG format. In theory, ImageMagick can convert close to a hundred different graphic file formats.

Preservation of Metadata

Some graphic file formats have the built-in metadata structure within their own file formats. For most of the new digital cameras, the EXIF [19] (EXchangeable Image File format, created by Japanese Electronic Industry Development Association, based on TIFF 6.0) metadata tags are embedded within the JPEG file format. The EXIF tags focus on camera device information such as focal length of lens, shutter speed, aperture value, compression bits/pixel, etc. (see Figure-5a) along with the make and model, photo creation information like date/time, and photographer's annotation comments. Another example is the DICOM [20] format, widely used in the medical industry, which is also based on the TIFF format and carries data acquisition creation information as well as patient information (see Figure-5b).

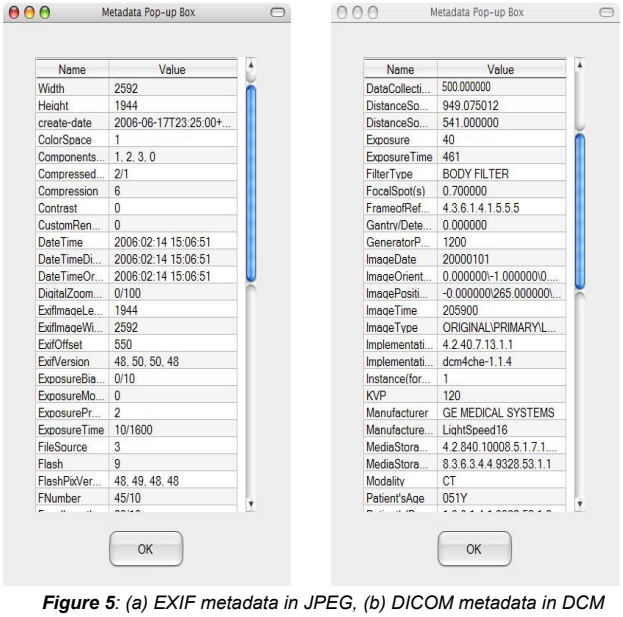

***Figure 5**: (a) EXIF metadata in JPEG, (b) DICOM metadata in DCM*

ADMIRE provides three sets of metadata as shown in Figure 6: (a) device metadata – creation information from a given device, (b) user metadata – annotation made by users, and (c) application metadata – feature extraction information made by software applications. For device metadata, ADMIRE will gather the basic creation information from each respective embedded metadata, including image description (ex. Event description), make and model of the device, original date/time, pixel resolution, and file-size of a given image. All three sets of metadata tags will be captured using MPEG-7 to form the keyword for searching for the image content.

System Metadata		User Metadata		Application Metadata	
Image Description	JPEG	Who		Finding 1	
Manufacturer	Canon	What		Finding 2	
Model	Canon PowerShot A710 IS	When		Finding 3	
Org. DateTime	2007:09:25 09:04:33	Where		Finding 4	
Modified Date	3/5/2009 3:56:29 PM	Why		Finding 5	
Resolution	3072 X 2304	How		Finding 6	
Filesize	1,519.95KB	Event		Finding 7	
Filetype	JPEG	By		Finding 8	

***Figure 6**: Times ADMIRE metadata panel*

Packaging Content using MAF Strategy

To preserve the image content, ADMIRE expands on the PPMAF concept to package the associated metadata, a thumbnail, the original image, the converted image, and the conversion process results into the MP4-FF as shown in Figure 7. For file-based storage application, all image contents will be stored in the media data box and their associated metadata will be stored in the movie box. The link between metadata and their corresponding image contents are specified by a media box. Linkage information to external references can be specified in the metadata contained in the meta box of each track box.

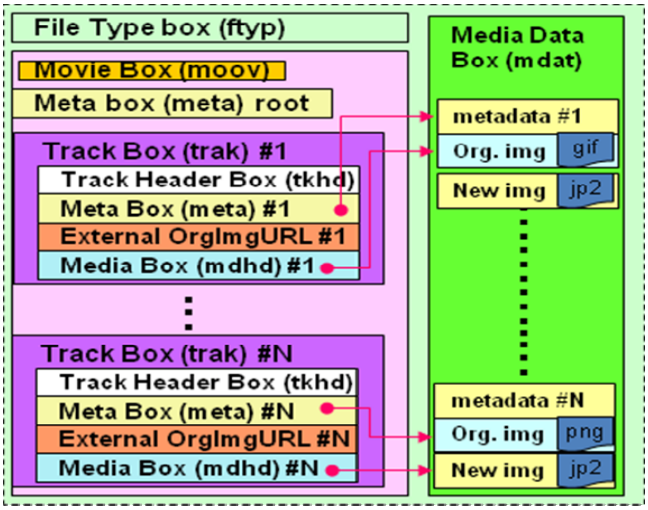

***Figure 7:** ADMIRE metadata panel*

In the mp4 file, there are two kinds of descriptive metadata files. The first one is the *collection-level* metadata (Figure 8a) at the root level, which provides event description about a set of given images with pointers to the individual images. The second one is the *item-level* metadata (Figure 8b) at the track level, which provides image description (ex. image resolution, file-size, image location either local or external via URL, etc.) for a given image. Conversion process results will also be stored at the item-level metadata.

On-going Work

Currently, we are in the process of integrating the Kakadu [21] C++ library for JPEG2000 conversion and mapping the MPEG-21 framework as the implementation approach for the OAIS (Open Architecture Information System) [22] reference model. As stated earlier, JPEG2000 offers a rich set of features especially the progressive transmission, which will be suitable for the distributed archival systems. It allows portable device clients to efficiently access image content repositories. For OAIS, it is important to adopt this reference model since it provides a clear separation between *Producer* and *Consumer* via the well-defined functionalities for the Submission Information Package (SIP) between *Producer* and *Ingest*, Archival Information Package (AIP) between *Ingest* and *Archival Storage*, and Dissemination Information Package (DIP) between *Access* and *Consumer*. There are also package description database updates between *Data Management* and *Access* for query request and query result sets.

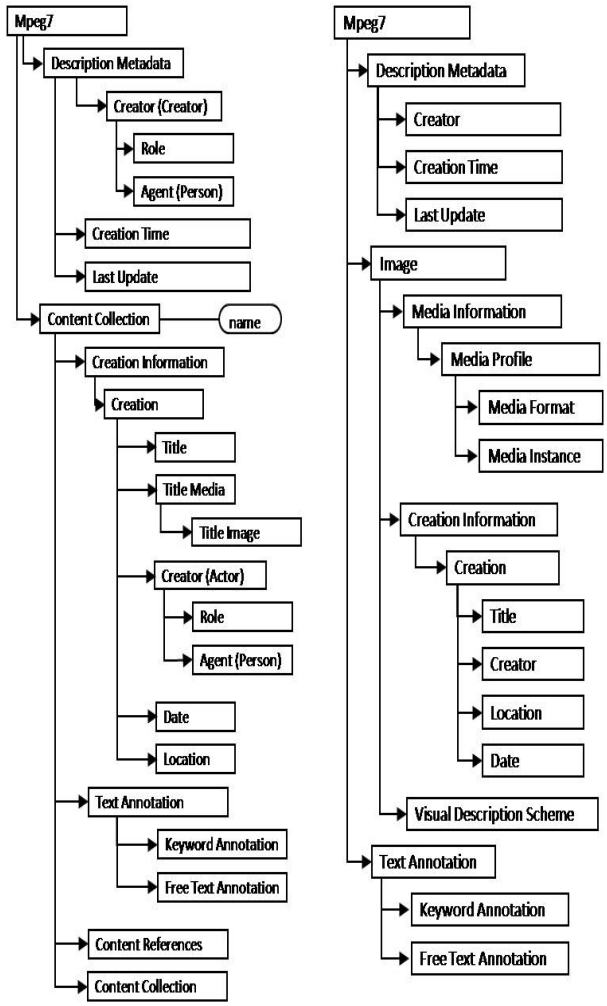

Figure 8: (a) Collection-Level Metadata, (b) Item-Level Metadata

5. Conclusion

This paper presents the advanced digital image preservation data management architecture, namely by adopting the ISO/IEC MAF as the preservation strategy. The implementation approach includes utilizing existing file format conversion tools such as ImageMagick and JPEG2000 Kakadu libraries, where JPEG2000 is used as the converted standard file format and the MAF as the standard package file container for storing metadata and image content.

For future work, within the image file format space, there is the soon to be standardized ISO/IEC FDIS 29199 JPEG XR (Extended Range) image coding system [23] which offers similar capabilities as the JPEG2000 but is more suitable for high dynamic range applications with a smaller memory footprint. As part of the preservation strategy, it is important to identify and explore whether if other emerging standard technologies can be applied to better preserve those irreplaceable image contents. Outside the image space, it is equally important to explore the same preservation strategy and approach in handling audio and video content. This will reduce the level of energy and effort to investigate and deploy the already identified technologies and yet provide standardized and interoperable infrastructure between image, audio and video content. Since the MAF, MPEG-7 metadata, and MPEG-4 FF were designed with multimedia content, there may be a good fit to use these technologies as a vehicle to capture preservation metadata, content, and conversion process results in a unified way.

6. References

[1] How much information: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/magnetic.htm

[2] Multimedia Application Formats (MAF): http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=42010&ICS1=35&ICS2=040

[3] Moving Picture Experts Group: http://www.chiariglione.org/mpeg/

[b] http://www.autodesk.com/techpubs/autocad/dxf/dxf2002.pdf

[4] http://en.wikipedia.org/wiki/BMP_file_format

[5] http://en.wikipedia.org/wiki/PCX

[6] http://en.wikipedia.org/wiki/Tagged_Image_File_Format

[7] http://en.wikipedia.org/wiki/Truevision_TGA

[8] DXF: http://en.wikipedia.org/wiki/AutoCAD_DXF

[9] http://en.wikipedia.org/wiki/SYmbolic_LinK_(SYLK)

[10] http://en.wikipedia.org/wiki/PICT

[11] http://en.wikipedia.org/wiki/CGM

[12] Photo Player MAF (PPMAF): http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ic.htm?csnumber=43606&ICS1=35&ICS2=040

[13] ISO/IEC JTC1/SC29/WG11N6828 – MPEG-7 Overview:http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[14] MPEG-4 File Format: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38538

[15] ISO/IEC Base File Format: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51533

[16] JPEG-LS: http://www.jpeg.org/jpeg/jpegls.html

[17] JPEG2000: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=33877

[18] ImageMagick: http://www.imagemagick.org/script/index.php

[19] EXchangeable Image file Format for digital still cameras: EXIF Version2.2, JEITA CP-3451, Standard of Japan Electronics and Information Technology Industries Association ISO/IEC

[20] DICOM: http://medical.nema.org/

[21] Kakadu C++ Library: http://www.kakadusoftware.com/

[22] OAIS: http://public.ccsds.org/publications/archive/650x0b1.pdf

[23] JPEG XR: ISO/IEC JTC 1/SC 29/WG1 N 4918

Author Biography

Wo Chang is currently serving as manager of the Digital Media Group at National Institute of Standards and Technology (NIST). In these duties Mr. Chang oversees several key projects including digital data archival and preservation, motion image quality, and multimedia standards. In the past, Mr. Chang was the Deputy Chair for the US National Body for MPEG (INCITS L3.1) and chaired several key projects for ISO/IEC SC 29 WG11 (MPEG), including Content-based Search Framework, Multimedia Application Formats, MPEG-7 Profiles and Levels, and co-chaired the ISO/IEC SC 29 WG1 (JPEG) JPEG Search project. Mr. Chang was one of the original members of W3C's SMIL (Synchronization Multimedia Integration Language) Working Group and developed one of the SMIL reference software. Furthermore, Mr. Chang also participated in the Internet Engineering Task Force for the protocols development of Session Initiation Protocol, Real-time Transport Protocol, Real-Time Streaming Protocol, and Resource-Reservation Protocol. Mr. Chang's research interests include digital data preservation, content metadata description, digital file formats, multimedia synchronization, and Internet protocols.