Paper: The Use of JPEG2000 in the Information Packages of the OAIS Reference Model

Robert Buckley, Jim Reid and William Stumbo, Xerox Research Center Webster, Webster, New York

Abstract

Two topics that have been prominent at recent IS&T Archiving Conferences have been OAIS and JPEG2000. OAIS stands for Open Archival Information System; JPEG2000 is a wavelet-based image compression standard from the JPEG committee. While use of JPEG2000 is growing in image archiving systems, its use is seldom described in the context of the OAIS model. This paper will describe the use of JPEG2000 in the context of the OAIS Information Packages and their requirements. It will report on the practical use of JPEG2000 data and files in an OAIS-based image archiving system, with a special focus on the choices made for JPEG2000 content in the Archival Information Package (AIP), and how they affect downstream performance when content is disseminated to the consumer in the OAIS model in the form of a Dissemination Information Package (DIP). The encoding parameters can affect both the ability of the AIP to meet the requirements of the archival system and the efficiency of the transformations between AIPs and DIPs, especially when the DIPs are not precomputed but derived on demand to meet consumer requests.

OAIS and Information Packages

The OAIS Standard (ISO 14721:2003) defines the Reference Model for an Open Archival Information System [1]. The Reference Model is a framework for understanding and applying the concepts needed for the long-term preservation of digital information. It covers the full range of functions associated with an archive: ingest, storage, data management, access, administration and planning. While the reference model is often pictured as a set of interconnected processes (see for example Figure 4-1 in [1]), our interest here is in the data flow configurations in an OAIS-based system. So we will focus instead on the main data objects in an OAIS-based system, shown in Figure 1. These are the three Information Packages for Submission, Archive and Dissemination (SIP, AIP and DIP), and the Descriptive Information.

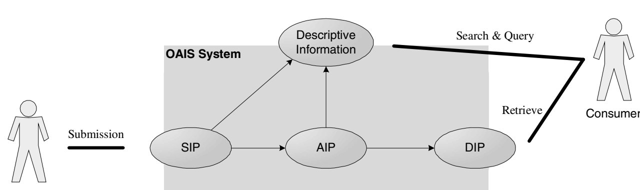

***Figure 1**. Data Objects in the OAIS Reference Model*

Figure 1 also shows the base data flow configuration in an OAIS-based system from SIP to AIP to DIP, involving the three variants of an Information Package. Input is delivered to an OAIS system in the form of a Submission Information Package (SIP), a packaging of the incoming information that is digestible by the archival system and used to construct one of more AIPs. The heart of an OAIS archive is the Archival Information Package (AIP). It is the variant of the Information Package that is stored and preserved by the OAIS. The Descriptive Information is used to discover the AIP of interest in response to a consumer request. When consumers request information from the archival system, it is provided to them in the form of a Dissemination Information Package (DIP), which is derived by transforming and converting one or more AIPs in a way identified by the Order Agreement, an OAIS construct specifying details of the delivery from the archive to the consumer. Of the four data objects in Figure 1, three are exposed outside the system: SIP, DIP and the Descriptive Information.

JPEG2000

JPEG2000 is an open wavelet-based ISO image compression standard [2]. Compared to existing methods, JPEG2000 has features and capabilities that make it attractive for image archiving and access [3]. It defines a single algorithm for both lossy and lossless compression and supports the creation of tiled and packetized codestreams, where a single packet contains the compressed data from one quality layer at a specific position of one resolution of a given color [4,5]. Therefore a single codestream can support multiple quality and resolution views of a compressed image, enabling progressive display and scalable rendering, which simplifies image management by reducing the need to maintain multiple image derivatives.

JPEG2000 supports "smart" decoding so that a user need only retrieve and decompress those parts of the JPEG2000-compressed image to obtain the quality, resolution, size, and portion of the image required for the application at hand. The smart decoding capabilities of JPEG2000 enable fast access to image subsets and the generation of image derivatives on demand and on the fly to meet customer requests for pan and zoom views.

The JPEG2000 standard also defines file formats that are capable of representing single or multiple images [2,6,7]. These formats provide generous metadata support and enable preservation by allowing metadata to be packaged with the codesteam.

Discussions of JPEG2000 almost always describe JPEG2000-compressed image data in a JPEG2000 file format, such as JP2, defined in Part 1 of the standard [2]. When discussing JPEG2000, it is useful to distinguish between the compression algorithm and the file format, both of which were standardized by the JPEG committee as part of the standard. Just as other file formats can accommodate JPEG2000-compressed data, JPEG2000 file formats can accommodate non-JPEG2000 and even uncompressed image data. The focus here however will be on JPEG2000-defined formats containing JPEG2000-compressed image data.

JPEG2000-based Data Flow Configurations

The capabilities of JPEG2000 position it as a reasonable choice for a universal archival image format. Because JPEG2000 is especially advantaged for storage and distribution, this paper will focus on it use for archiving and dissemination. It will review the various JPEG2000 use models currently employed, casting them in terms of data flows using AIPs and DIPs.

Archival Information Package

The AIP consists of the information that is the focus of preservation, accompanied by a set of metadata sufficient to support the archive's preservation and access services. The archived information and its associated metadata represent a single logical package within the archival system. This single logical package may map onto a single physical file, but is not required to do so.

An AIP could also be an Archival Information Collection, whose content information is the aggregate of multiple AIPs. However, for the purposes of this discussion, we will restrict ourselves to AIPs that are Archival Information Units, whose content information represents a single image.

Using JPEG2000 as the format for AIPs provides several advantages. First, using JPEG2000 offers the system the choice of lossless or lossy compression within the same compression framework. And using the JPEG2000 file format offers an alternative to TIFF for uncompressed data. Lossless compression is often the reflexive choice for an AIP, but an examination of requirements may show that loss is tolerable, especially when the larger sizes of lossless files can become onerous. Another advantage is that multiple derivatives can easily be obtained from a JPEG2000-compressed AIP file without always requiring that the entire image be accessed and decompressed first.

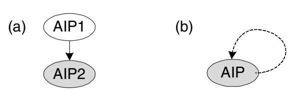

***Figure 2**. Conversion Data Flows*

Figure 2(a) illustrates a data flow where the contents of an archive are migrated from one format to another. In our data flows, JPEG2000-encoded packages are shown shaded. An example of this scenario would be an organization with a large archive of images in lossless TIFF format. The organization would like to migrate this archive to JPEG2000 without losing any image quality. This could easily be done by a simple application which takes each Archival Information Package, extracts the lossless TIFF file and converts it to JPEG2000. The application could easily be extended to perform a validation, ensuring the two images are identical on a pixel by pixel basis. Abrams has reported on a large scale automated migration of GIF, JPEG and TIFF files to JPEG2000 [8]. While he did not specifically mention the OAIS model in his paper, some of the issues he reported would apply here just as well.

Figure 2(a) can also describe the situation where there is an archival master, and derived from it by a conversion process, a production master, with both packages retained. For example, in Phase 1 of the National Digital Newspaper Program, each newspaper page is delivered as a lossless TIFF archival master and a JP2 production master containing lossy JPEG2000-encoded image data. The production masters are designed to serve as the surrogates of the TIFF master files for day-to-day use and client access [9].

One aspect of performing this type of conversion is the need prior to conversion to gain an understanding of how consumers would like to access the archival information. This will influence the encoding of the JPEG2000 files. The encoding used is important when the AIP is transformed to a DIP in response to a consumer request. For performance reasons, it is desirable to make one of the easily derived subsets of the JPEG2000 codestream correspond to the anticipated resolution and quality of the DIPs that will be needed. For example, if being able to quickly display a screen-sized image is important, then the number of resolution levels would be selected so that the lowest resolution level subband would be XGA size or smaller, with the compressed data organized in resolution-major progressive order.

Of course, user expectations and requirements can change over time. Figure 2(b) represents a data flow that accommodates these changes. Once it becomes clear that the format of the JPEG2000 files needs to be revised to best meet new usage patterns, the JPEG2000 files can be converted or re-encoded using parameters that create files better tuned for the current environment.

Dissemination Information Package

The DIP packages the information supplied to an end user as the result of an Access request. The DIP can correspond to one or more AIPs, or to part of an AIP. The Access component of the archival system converts information from the internal archive format to a DIP using a currently-supported format for delivering the content to the consumer.

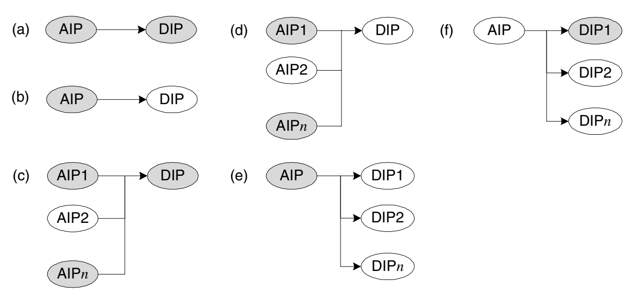

***Figure 3**. Dissemination Data Flows*

The DIP may be delivered as a JPEG2000 file as shown in Figure 3(a). It may also be delivered in a non-JPEG2000 format, such as an image format native to a web browser (Figure 3(b)). AIP-to-DIP transformations of this type would require a format conversion, with transcoding and possibly image processing to condition the image.

The data flow in Figure 3(b) is fairly common in archival systems that use JPEG2000 for an AIP. In these situations, the archival system manager obtains the benefits of JPEG2000, while consumers can continue using their browsers without downloading or installing plugins. Motivated users may want to download a JPEG2000 plugin to receive the benefits of JPEG2000 on the receiving end. In this case, illustrated by the data flow in Figure 3(a), the DIP is a subset of the JPEG2000-encoded AIP, which is delivered directly to the consumer for decompression and viewing. This data flow can be mediated by JPIP, the JPEG2000 Interactivity Protocol [10], a client-server protocol specifically designed for accessing JPEG2000 codestreams and file formats.

One system that offers users the choice of JPEG or JPEG2000 viewing is the Digital Archive of the National Archives of Japan at http://jpimg.digital.archives.go.jp/kouseisai/index_e.html.

Figure 3(c) – Figure 3(f) illustrate other possible AIP-to-DIP conversions. We note that it is possible to combine the contents of multiple AIPs into one DIP in response to a user request. Likewise, a user request may generate multiple DIPs from one AIP. Lastly, while our focus is on archives advantaging JPEG2000, we note that AIPs may contain other formats.

Submission Information Package

The Submission Information Package (SIP) is the variant of the information package that is transferred from the Producer to the OAIS archive when information is ingested into the archive. A submission can consist of a single SIP or multiple SIPs. Likewise the ingested SIP may result in the creation of a single AIP or multiple AIPs. Figure 4 illustrates some of the potential combinations. The specific combination will be driven by the items being stored into the archive and their intended usage.

An example of the submission data flow shown in Figure 4(b) is the Internet Archive. The AIP is a ZIP file containing an individual JPEG2000 file stored using one JP2 file for each book page. This approach was taken since JP2 is a single-image file format.

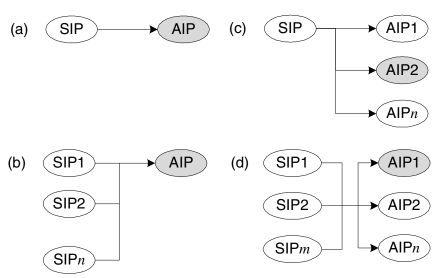

***Figure 4**. Submission Data Flows*

The SIP will typically include documents (in paper and/or electronic form) that represent the object to be archived, as well as metadata that can be used to describe and annotate the object. If the documents are transferred in paper form, they would be scanned and saved as a lossless or lossy JPEG2000 file, as appropriate, using JPEG2000 encoding parameters that are suitable for future retrieval of the asset.

If the documents are transferred in electronic form, they would be converted and saved as lossless or lossy JPEG2000 files, again as appropriate, using the same parameters mentioned above. The native color space of the submission will determine whether the color support of JP2 is sufficient or whether a color space from JPX is required [6]. It may be that many archival applications need a JPEG2000 file format profile that is intermediate between JP2, which is defined in Part 1 of the standard, and JPX, which is defined in Part 2. This profile would combine the relatively simplicity of Part 1 codestreams with the extended color support of Part 2 file formats.

JPEG2000 Encoding: Design Study

Table 1 summarizes the JPEG2000 encoding for the production masters created in Phase 1 of the National Digital Newspaper Program (NDNP). NDNP is a joint collaborative program of the National Endowment of the Humanities and the Library of Congress, intended to provide access to page images of historical American newspapers. Web access will be provided through the use of JPEG2000 production masters. The recommended format for these masters is a visually lossless, tiled JPEG2000-compressed grayscale image, with multiple resolution levels and multiple quality layers, encapsulated in a JP2 file with Dublin-Core compliant metadata [7]. Figure 5 shows a sample NDNP image.

Table 1. JPEG2000 Encoding

Parameter	Value
Overall	Visually lossless (8:1)
Metadata	RDF/Dublin Core
Number of components	1 (sGray, ICC)
Component Transform	None
Tile size	1024 x 1024
Wavelet Filter	9-7
Number of levels	6
Number of layers	25
Progression Order	RLCP
Resolution	300 or 400 dpi
Start of Packet markers	No
Precincts	No
Codeblock size	64 x 64
Coder speed-up	None

The key design choices in this case study are the overall decision to choose a lossy coding, and the selection of the number of resolution levels and quality layers. The number of components was set by the grayscale nature of the application, and the JP2 file format required the use of either a calibrated gray scale space, based on the sRGB color space, or a restricted ICC profile.

The encoding is visually lossless, which means it is lossy, of course, and it is not possible to exactly reconstruct the original image from the production master. However, the differences are either not noticeable or do not adversely affect the intended use of the production master. After some visual testing, it was found that 8:1 compression was a good compromise between file size and visual screen presentation quality for full size images.

The number of resolution levels was selected so the lowest resolution level image for the largest expected page image size was approximately QVGA size, which is 320 by 240. The dimensions

SENATOR PLATT REVIEWS LIFE

PEARY'S VESSEL WELL UNDER WAY

HORSE RUNS AWAY; FOUR ARE HURT

Thermometer Climbs; Poor Mortals Suffer

Goldenberg's

Seventh and K Streets

Sacrificing Shirt Waist Suits

$2.98

$3.98

Summer Suits to Order at Half Price

MERTZ and MERTZ CO., 906 F St.

Figure 5. *Sample NDNP Image*

of the original image shown in Figure 5 are 6306 by 8997, so 6 resolutions levels would suffice in this particular case.

$$\lceil \text{Log2}\ (8997/240) \rceil = \lceil 5.23 \rceil = 6 \qquad (1)$$

Using layers makes it possible to obtain reduced-quality versions of the production master. At maximum resolution and for full quality, all layers would be decompressed. However, at lower resolutions, higher compression ratios are possible without objectionable visual artifacts. Multiple layers optionally allow higher compression ratios at these lower resolutions. An advantage is that the fewer quality layers decompressed, the faster decompression is completed, which can lead to savings in decode times at lower resolutions where the reduced quality is not noticeable. The design in Table 1 uses 25 quality layers, selected so that the logarithms of the layer bit rates are close to being uniformly distributed between maximum and minimum values, corresponding to bits rates of 1.0 and 0.015625 bits per pixel.

Figure 6 shows the effect of reduced layer decompression on total decompression time for the encoding described in Table 1. Total decompression time is the time to access and decode the decompressed file. (Figure 7 shows just the block decode time.) The x-axis is the number of levels that the decompressed image is reduced below full resolution, from 0, which is full resolution, to 4, which produces an image one-sixteenth the size of the original. The y-axis is the log base 2 of the total decompression time. Each series on the plot shows how decode time varies with the number of layers that are decoded. What this plot shows is that for some reduced resolution level images, such as Reduce=2, decompressing 15 layers instead of the full 25, can more than halve the total decompression, and almost halve the decode time. The resulting decompressed images show little difference in quality. As more resolution is requested, then it is accompanied by more quality layers in the decompression stage.

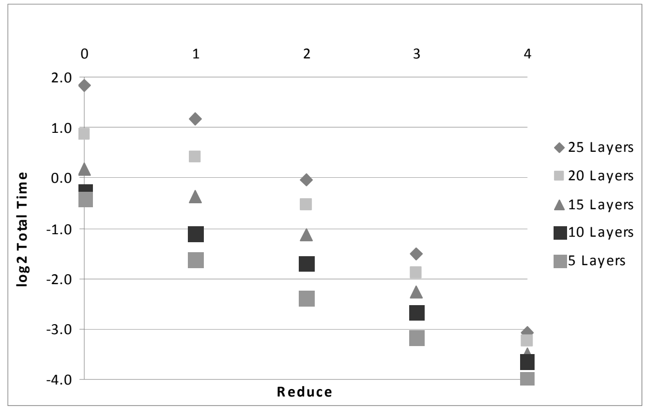

Figure 6. *Log Total Decompression Time vs. Resolution for Different Quality Layers.*

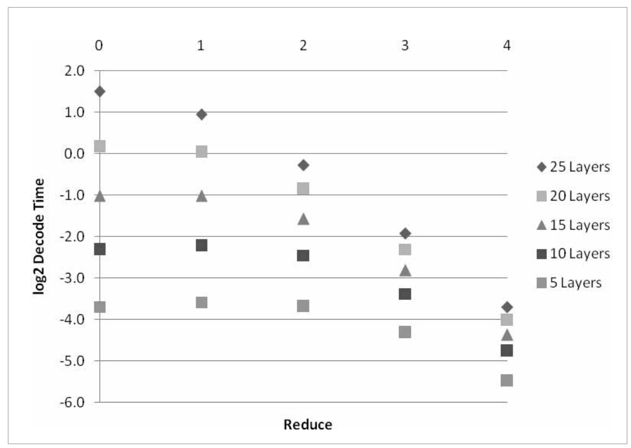

Figure 7. *Log Block Decode Time vs. Resolution for Different Quality Layers.*

The values plotted in Figures 6 and 7 were obtained using Kakadu expand to measure end-to-end and block decoder times on a Dell Optiplex GX620 with a Pentium 4 3.0 GHz microprocessor, 1 Gigabyte of RAM and a 3 Gigabyte cache. This function measures elapsed time, so the times measured demonstrate the characteristics and idiosyncrasies of the Kakadu operation, which can depend somewhat on what other applications are running at the same time. However, it was found that running two dozen applications only increased the times measured by a few percent, less than 10, compared to only one application running.

The compressed file used in the tests reported in Figures 6 and 7 did not use any coder speed-up functions. In a subsequent test, the compressed file was created using the CoderBypass option, for a small additional savings in decode time.

Conclusions

This paper has illustrated the use of JPEG2000-based data flows in an OAIS-based image archiving system. It has focused in particular on the choices made for JPEG2000 content in the Archival Information Package, and how they can positively affect downstream performance when content is disseminated to the consumer in the OAIS model.

References

[1] ISO 14721:2003, Space data and information transfer systems -- Open archival information system – Reference model (2003); also available as CCSDS 650.0-B-1, CCSDS RECOMMENDATION FOR AN OAIS REFERENCE MODEL, January 2002

[2] ISO/IEC 15444-1:2004, Information technology – JPEG 2000 image coding system: Core coding system (2004)

[3] R. Buckley & F. Frey, Preserving Images, oe Magazine, 4.1, 21 (2004)

[4] D. Taubman and M. Marcellin , JPEG2000: Image Compression Fundamentals, Standards and Practice. (Kluwer Academic Publishers, Dordrecht, 2002)

[5] M.W. Marcellin, M.J. Gormish, A. Bilgin and M.P. Boliek, An Overview of JPEG-2000, Proc. IEEE Data Compression Conference, Snowbird, UT, pp. 523-541 (2000)

[6] ISO/IEC 15444-2:2004, Information technology – JPEG 2000 image coding system: Extensions (2004)

[7] ISO/IEC 15444-6:2003, Information technology – JPEG 2000 image coding system -- Part 6: Compound image file format (2003)

[8] Stephen Abrams, Automated Migration for Image Preservation, Proc. IS&T Archiving Conference, pg. 113 (2006)

[9] Robert Buckley and Roger Sam, JPEG 2000 Profile for the National Digital Newspaper Program (2006), available at http://www.loc.gov/ndnp/pdf/NDNP_JP2HistNewsProfile.pdf

[10] ISO/IEC 15444-9:2005, Information technology – JPEG 2000 image coding system: Interactivity tools, APIs and protocols (2005)

Author Biography

Robert Buckley is a Research Fellow in the Xerox Research Center Webster in Webster, New York. He is the Xerox representative on the US JPEG 2000 committee and was the Project Editor for Part 6 of the JPEG 2000 standard. He currently chairs the CIE Technical Committee on Archival Color Imaging. He has given several invited talks on JPEG 2000 to the cultural heritage community and has consulted on its use in archiving applications.

Jim Reid received a Bachelor's degree in Math from the University of Rochester in 1983, where he also attended graduate study in Electrical Engineering. Jim has spent the past 23+ years as a computer programmer, primarily in the research labs of both Kodak and Xerox, where he is currently employed. Jim's expertise is in the areas of object-oriented programming and image processing.

William Stumbo is a member of the research staff in Xerox Research Center Webster in Webster, New York. He has held a variety of technical positions in product development and research with Xerox . He completed his Bachelor's degree in Computer Science at Purdue University in 1983 and received a Masters degree in Computer Science from Rochester Institute of Technology in 1993.