Using JPEG2000 for Enhanced Preservation and Web Access of Digital Archives – A Case Study

James S. Janosky
Aware, Inc., Bedford, MA, USA

Rutherford W. Witthus
University of Connecticut

Abstract

JPEG2000: The New Standard for Digital Archiving

The JPEG2000 standard (ISO 15444-1) provides the advantages of advanced wavelet compression to digital archives while eliminating the concerns associated with proprietary compression and file formats. JPEG2000 allows archivists to preserve culturally significant digital objects using lossless compression while making the collection more accessible to a wider audience.

From a single master JPEG2000 image, one can extract a highly compressed image for transmission and display it in a web browser. The layered file format supports extracting any desired image size or quality. Tiling, Progressive Display, and Client-Side Region of Interest can be combined to provide for effective viewing of archive-quality files over a limited bandwidth. Compliance with an ISO standard and embedded support for multiple types of metadata each help ensure that the archive content outlives the systems that created it.

Using Charles Olson's Melville Project at the University of Connecticut as a case study, this paper demonstrates the capabilities of a JPEG2000 Image Server and discusses how the JP2 and JPX files can be used to support multiple types of metadata for such archives.

Introduction

JPEG2000 is a relatively new international standard for image compression developed by the ISO/IEC JTC1 SC29 Working Group 1, also known as the Joint Photographic Experts Group (JPEG). JPEG2000 was designed to take advantage of new mathematical techniques to improve still image compression by providing better image quality at high compression ratios, lossy and lossless compression with a single codec, error resilience for noisy channels, and region of interest coding.

JPEG2000 uses a wavelet transformation, which makes it fundamentally different from the previous JPEG image compression standards. Since the wavelet transform is performed over the entire image, a JPEG2000 image does not exhibit the blocky artifacts common in highly compressed traditional JPEG images. JPEG2000 will also generally yield twice as much compression for the same image quality as JPEG.

The advanced functionality of JPEG2000 derives from the layered file format and the resulting ability to extract portions of the compressed image code stream for viewing. These portions can be used to progressively display an image as each data layer arrives, effectively reducing the required transmission time. Similarly, a JPEG2000 image can be viewed without fully decoding the image.

The advantages of JPEG2000 for digital archiving include:

1. Open standards that "future proof" data and encourage collaboration
2. Rich support for metadata within the compressed image files, including XML schemas (e.g. EAD, METS, MARC, NISO, PDF, etc.)
3. Support for lossless and lossy decompression
4. Efficient remote viewing of archive-quality images through tiling and progressive decoding of resolution levels

This paper presents the Aware JPEG2000 Image Server and its functional components. It goes on to discuss selection of various JPEG2000 encoding options used to maximize the efficiency of the JPEG2000 Image Server. It then presents the Melville Project case study: an actual implementation of the JPEG2000 Image Server.

Aware JPEG2000 Image Server

The JPEG2000 standard enables random access to the compressed image code streams. The Aware Image Server uses this feature to extract and decode the minimum amount of data necessary for viewing and to provide interactive zooming on the selected image. A View Window is used to “zoom in” on a particular region of the image. A Navigation thumbnail indicates the region selected for viewing using an overlaid graphical box, and the lowest resolution layer of the entire image may be viewed in a larger, separate window. The requested resolution level of the region in the View Window is extracted, and only this

much smaller image is transmitted to the client. Native image quality is preserved during the zoom process by utilizing the multi-resolution format of JPEG2000 images. The zoom process involves server-side extraction of incrementally higher resolution data that is contained within the archived JP2 file. Because the view window is of constant size, the same amount of data is transmitted for each zoom level.

The Aware Image Server user interface is a web page that:

- Retrieves and displays the thumbnails (an extracted resolution level, not a separate image),
- Retrieves the view window image (extracted region and resolution level data),
- Retrieves and formats metadata from the JP2 or JPX image files, and
- Assembles the various components.

All data is extracted from the single master compressed image file, eliminating the need to create and maintain multiple versions of each digital object (e.g. thumbnails, archives, viewing resolutions, printing resolutions, etc.).

Compressed JPEG2000 images

The compressed JPEG2000 images (JP2 or JPX files) may be stored in either a file system or a database with a pointer for each image provided to the JPEG2000 Image Server. Batch processing scripts are provided to compress the images (TIFF to JP2 in this case study) and to insert metadata. If the metadata files are linked to the source images through a naming convention, they can be systematically included via scripting as part of the compression process. Metadata can also be inserted and edited at any time after the creation of the compressed image file.

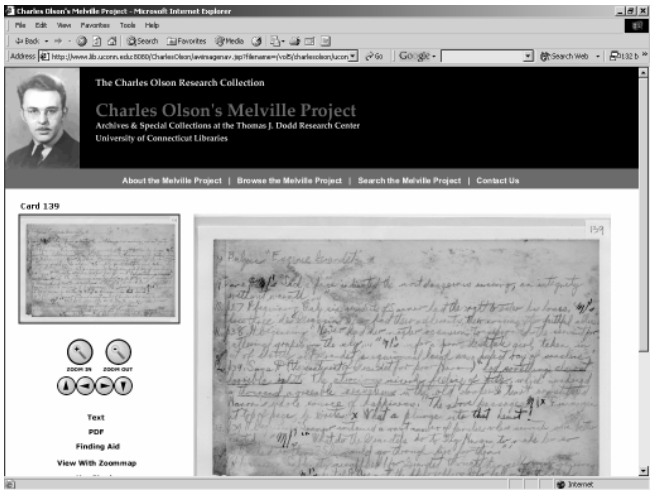

Figure 1. Screen shot of the Melville Project JPEG2000 Image Server showing the view window, thumbnail, navigation buttons, and metadata links.

Selecting JPEG2000 Encoding Options

Before creating a digital collection using JPEG2000, some basic decisions must be made to select the proper compression parameters and options. The many encoding options supported by the JPEG2000 standard provide a fine level of control over the compression process. The ideal encoding options will depend on the material in the collection and how it is likely to be used. The following sections outline factors to consider.

File Types: J2K, JP2, and JPX

JPEG2000 supports three basic compressed image file types: J2K, JP2, and JPX. A J2K file is a single compressed image code stream. The JP2 and JPX file formats are respectively designed to include basic and advanced forms of image metadata. Note that not every JPEG2000 decoder can handle JP2 files or the additional information found in JPX extensions. There are several levels of compliance defined in the standard.

JP2

JP2 files may contain one or more compressed J2K images, several types of metadata boxes, and two enumerated color spaces: sRGB and grayscale. Four types of JP2 metadata boxes are defined in the standard:

1. Intellectual Property Box: Used for carrying intellectual property rights information about the image(s) in the file.
2. XML Box: Used for vendor specific information in XML format. (E.g. NISO Z39.87, MARC, METS, etc.)
3. URL Box: Used for including an URL that can be used by an application to acquire more information about the associated image or vendor.
4. UUID Box: User defined metadata boxes used for any other information not covered by the above metadata boxes (e.g. PDF files, audio files, etc.).

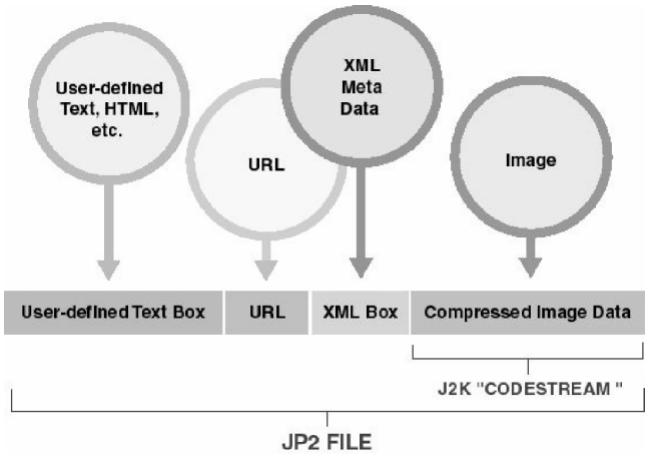

Figure 2. Diagram of typical JP2 file.

The JP2 Image Header box contains a field indicating whether or not the original color space is known. An unknown color space indication means that the color space included in the image is an approximation of the unknown original.

JPX

Baseline JPX files may contain everything in a JP2 file as well as a limited sub set of the extensions found in Part 2 of the standard. Baseline JPX supports 8 of the 17 restricted color spaces, full ICC color profiles, and additional types of metadata boxes. Baseline JPX files may contain more than one color space, each with its own approximation and precedence. The approximation field is used to indicate how well a color specification approximates the actual color space of the image, ranging from exact to poor. If more than one color space is present, the precedence field is used to suggest a priority depending on the capabilities a particular decoder. Baseline JPX also adds the ability to include an Output ICC profile for commercial printing and proofing systems. Full JPX may include other extensions such as image integrity verification, image history, geo-referencing metadata, additional restricted ICC profiles, vendor defined color profiles, multiple composite layers, etc.

Tiles and Resolution Levels

JPEG2000 images should be compressed in tiles and multiple resolution levels for the most efficient use by an image server. Resolution levels are power-of-two reductions of the original image. If the image is tiled, the resolution levels will consist of power-of-two reductions of every tile. The tile size specified during compression will determine the number of available resolution layers. A tile size of 1024 x 1024 pixels yields 6 resolution layers by default with the Aware JPEG2000 Image Server used in this case study.

Table 1. Resolution level size per tile

Resolution Level	Size in pixels
1	1024 x 1024 (full image tile)
2	512 x 512
3	256 x 256
4	128 x 128
5	64 x 64
6	32 x 32

An image may have additional resolution levels down to a 1 x 1 pixel layer. Users may want to create additional resolution layers during encoding of very large images. Large images with many tiles will benefit from additional resolution levels since — at a minimum — the smallest resolution level from every tile must be decoded. Users may also set a specific target size, target compression ratio (target bit rate), and target quality for each layer. These features can be used to control the image quality available in each layer, which is particularly useful if access to the digital collection is to be restricted.

The layered file format of JPEG2000 also simplifies repository management, since multiple versions of each digital object (thumbnail, web version, print master, etc.) do not need to be maintained.

Lossy or Lossless Compression

The JPEG2000 standard supports both lossy and lossless image compression. The "parsable" bit stream and file format allows any region, resolution level, quality layer, color channel, or combination of these parameters to be extracted from a single master image. Images can be encoded losslessly and then decoded either losslessly or lossily by extracting the appropriate number of layers needed for a particular use. JPEG2000 allows highly compressed derivative images to be quickly extracted without decoding the entire file. For example, a losslessly compressed master image can be stored for preservation and reference. From this master file, a medium-quality image can be extracted at a 30:1 compression ratio and transmitted for browsing, and a high-quality image can be extracted at a 10:1 compression ratio to be viewed for most research. The full lossless image is also available. It is in this way that the quality scalability of JPEG2000 elegantly supports remote viewing and access of large, losslessly compressed image file. Starting with lossy compressed images will reduce the storage requirements but will limit the maximum image quality of the archived file.

Compression Ratio

Lossless compression typically yields compression ratios between 2:1 and 3:1. The higher compression ratios available with lossy compression can be used to further reduce storage costs and improve the performance of the JPEG2000 Image Server, since lossy files are smaller and require less data to be transmitted. As with any lossy compression algorithm, higher compression ratios will trade reduced file size for image quality. Generally, JPEG2000 can be used to compress images twice as much as traditional JPEG for the same image quality. Lossy compression ratios should be selected based on the type of material in the collection, the condition of the material, and the needs of the users.

Case Study

Over the past two years, Archives & Special Collections at the Thomas J. Dodd Research Center at the University of Connecticut in Storrs has worked on a project funded by the Gladys Krieble Delmas Foundation to clean and make accessible a series of hand-written cards produced by the poet Charles Olson during his effort to transcribe the marginalia in hundreds of books owned by Herman Melville. Due to extensive water damage to Olson's note cards, this important and valuable collection has been unavailable to researchers until now. Terms of the grant stipulated that the collection be publicly displayed. The project aspired to provide an online display of the collection to make it available to the widest possible audience.

Prior to beginning the digital project, the individual cards were separated, dry surface cleaned, humidified, and placed in clear polyester (Mylar-3) 3-sided pocket enclosures. This process was thoroughly documented. The cards were scanned as 600 dpi color images and stored as TIFF files. The TIFF digital images were then compressed to JP2 files in a batch process. A 10:1 compression ratio was used, providing excellent image quality while significantly reducing the storage requirements. The original archival TIFF images may also be compressed using lossless JPEG2000 at a later date for long-term storage, thereby eliminating the need to store the large TIFF files.

The Aware JPEG2000 Image Server dynamically generates thumbnails, low-resolution images, and high-resolution images from the master JPEG2000 encoded image. The images were compressed using 1024 x 1024 tiles, six resolution levels, and a "progressive by resolution" (RLCP) progression order. The JPEG2000 compressed image code streams were first ordered by resolution (R), then quality (L), color channel (C), and finally by position (P). Technical metadata from the scanning process were systematically included in the JP2 files during compression. Quantitative metadata for both individual items and the collection as a whole were added later using the metadata editing functions.

Four metadata boxes were included with each JP2 image: technical metadata, a text transcription of each card, a PDF file containing a text transcription, and the short Encoded Archival Description (EAD) finding aid. The scanner setting for each digital image was inserted into an XML metadata box as text in each JP2 file. A second XML metadata box was used to contain the textual transcription of each hand written card. A user-defined metadata box (UUID) was created to store PDF files as metadata. This provides users with a transcription as close to the original card as possible, including position and emphasis of words and sentences. Finally, the shortened EAD finding aid was inserted into a third XML metadata box. Even though the EAD describes the entire collection, a modified EAD was inserted into each JP2 compressed image file to provide context for the individual digital objects whose provenance would otherwise disappear and to allay concerns that an image may become disassociated from its corresponding metadata. While this did increase the size of the resulting files, it addressed the disassociation problem and simplified the operation of the image server. The collection is now smaller and simpler than it was since it is not necessary to store, maintain, and track multiple versions of each image.

The Aware JPEG2000 Image Server web page was customized to maintain the look and feel of the library's web site. Headers, branding, and background information were added to further integrate the JPEG2000 Image Server. An index page and additional web pages describing the project were created and a search function was integrated. XSL Stylesheets were created to format the metadata for display by the Aware JPEG2000 Image Server.

This case study illustrates that an Aware JPEG2000 Image Server can be used to effectively provide broad web access to a large, fragile collection. The scalability and interactive zoom features of the Aware JPEG2000 Image Server make it possible to present higher quality images on the web than would otherwise be possible, supporting detailed study without further endangering this fragile collection. The extensive built-in support for storing metadata within the same file as the image also greatly simplifies management of the collection. The preservation goals of the project are met by using a standards-based image format and metadata schema. The standards-based approach helps ensure the longevity of the collection and largely eliminates the need for future data migration.

While the University of Connecticut is still adding material to the online collection, the first images can be viewed at the following web site:

http://charlesolson.uconn.edu/Works_in_the_Collection/Melville_Project/browse.cfm

The University of Connecticut plans to add additional collections in the future as well as host other digital preservation projects. Because the standards-based approach of the JPEG2000 Image Server works so well in collaborative efforts, *Connecticut History Online*, the premier electronic image provider of historical images of Connecticut, will be processing its large-format materials using the Aware product.

Conclusion

JPEG2000 offers significant advantages for digital archives. As an open international standard with a lossless compression option, JPEG2000 is a superior format for the preservation of digital objects. The highly flexible format allows archivists to simplify repository management by reducing the number of versions of each digital object that must be maintained. Various types of metadata can be inserted directly into the JP2 or JPX image files, ensuring that the image is never separated from its associated metadata. By taking advantage of some of the advanced features of JPEG2000, the JPEG2000 Image Server enables efficient remote viewing of archival quality digital images. The interactive zooming features provide a rich way to view culturally significant material previously inaccessible to researchers and the public.

Biographies

James Janosky has 15 years technical business development and sales experience. Since joining Aware, he has helped develop the market for JPEG2000, focusing on digital archives, medical imaging, geo-spatial imaging, and embedded digital image processing. Mr. Janosky has worked closely with several major universities and library service vendors to develop digital archiving strategies using JPEG2000. Mr. Janosky has given presentations on JPEG2000 at the 2003 ALA Midwinter Technical Showcase and the 2002 CIL Conference.

Rutherford Witthus is the Curator of Literary and Natural History Collections at the Thomas J. Dodd Research Center at the University of Connecticut in Storrs. He is also in charge of the automation efforts in archives and special collections. Mr. Witthus has been involved in EAD projects, JPEG2000 development and implementation, and works with the technical Committee of *Connecticut History Online.*