https://doi.org/10.2352/issn.2168-3204.2017.1.0.157

JPEG2000 as a preservation format for digitization: lessons learned from a library

Laurent DUPLOUY; Bibliothèque nationale de France; Paris, France.

Abstract

This article attempts to present the methodology used to respond to questions and issues raised by the adoption of JPEG2000 format at the National Library of France for mass digitization. It attempts to describe particularly the methodology used to define a compression ratio for heritage digitization.

Finally, it presents lessons learned after two years of mass production.

Background

JPEG2000 (shorter: *JP2*) is considered for more than 15 years the best image format in itself, and memory institutions are constantly assessing it as a candidate for long-term preservation. [1]

The complexity of *JP2*, combined with the lack of tools to manipulate or validate it, also combined with unfortunate experiences, have prevented a lot of cultural memory institutions to adopt this format [1][2].

On the other side, some important benefits had to be considered: the efficiency of the compression algorithm, its customization, the fidelity to the original image…

The Bibliothèque nationale de France (BnF, the National Library of France) owns more than 2 Petabytes digital library in uncompressed TIFF and produces 15 to 20 million images per years: using a format that enables to reduce the volume and its growth has become a sustainability challenge. But the gain with lossless compression of JP2 is not sufficient in the case of digitization (ca. 1 for 2), therefore the BnF have to consider the lossy compression, and have reached:

1. the best shape of the algorithm for its digitization
2. the appropriate compression rate(s) to use. In addition, the fact that JP2 was already used as a dissemination format had to be taken in consideration.

Adoption limits

The BnF has been studying the *JP2* format since 2006. It took time to adopt this format for four main reasons.

The first reason was the difficulty of mastering the format. In the context of preservation, the BnF divides formats into 4 categories (from the least to the best mastered). Preservation igitization involves being in the strongest. This means to have at least one format expert who monitors the format, to have tools to validate the format regarding to the standard but also regarding to precise implementation at the BnF (see below). Given the complexity of the format, these requirements could not be reached until 2013.

The second reason is the socialization of the format. The BnF evaluates a format in terms of penetration, quantification of users and developers communities. More developers, more tools, more users, and vice versa. In this kind of green cycle, the risks are reduced and the maintenance costs of a format are limited. This format needed time to be adopted by a large user community. It is clear that if institutions of memory have adopted *JP2 format*, this is not yet the case for the general public. It considerably restricts the user community.

The third reason was the lack of validation and format characterization tools. Still for preservation reasons, the BnF enacts strict production rules. Here again, the objective is to control the contents which are produced by external service providers most of the time. This control reduces the risk of loss or degradation of future migration (transformation). With the arrival of *JPYLYZER* in 2012, this point was lifted. *JPYLYZER* makes it possible to verify the validity of the file and extract its characteristics. The latter can be confronted with a usage profile and thus makes it possible to obtain a uniform production in terms of technical choices and metadata.

Finally, the fourth reason is that the original version of *JP2* (Part 1) first version contained two elements of ambiguity. They have led to different understandings and potentially incompatible implementations in tools supporting *JP2* [2] [3]. This was a terrible threat for long term preservation.

1. The strict application of the standard prohibits in theory the inclusion in *JP2* format of ICC profiles of the "Display Device" type, such as Adobe 1998, ProPhoto RGB or RGB v2 used to define widely used color spaces. Only «Input Device» profiles were allowed.
2. Two metadata fields: "capture resolution" and "default display resolution" are defined, without semantics explanation. In addition, different digitization resolution units (pixels / inch or cm) were allowed without indication. So, there were two possible choices in existing tools at the time.

These two points were fully resolved in 2013 by the amendment 6 to JPEG2000 [4].

BnF's Recommendations

In 2014, an internal working group worked on the development of recommendations: quality levels, resolution levels, progression order, compression rate, etc. Here we present the main recommendations.

Compression rate

The objective was to get the best ratio with the minimum possible loss of information. We did not want to base our choice only on subjective criteria such as a comparison of the original in a controlled light environment with its rendition on screen.

In reference scientific literature, it is common to read that a Peak Signal to Noise Ratio (PSNR) of 25 or 30 permit to reconstructed correctly an image without a significant loss of information [5] [6]. In the context of the BnF requirements, the various experiments showed that the PSNR analysis did not allow

to find strictly and precisely the best ratio without any kind of visual loss. So, we did not use this method.

We categorized the collections to be treated in various groups according to the typology of the documents and the treatment characteristics:

- Specialized documents: prints, photographs, maps, glass plates, etc. Digitized in color (24 bits per pixel)
- "Exceptional" documents: printed matter or manuscript containing Illuminations, illustrations or individual brochures in color (24 bits per pixel).
- "Printed" documents are scanned in color (24 bits per pixel).
- Transparent documents scanned in grayscale (8 bits per pixel)
- Newspapers (large format) digitized in grayscale (8 bits per pixel)

For each of these groups, we selected a few documents representing the category. Then from the uncompress TIFF format image we applied on each the desired compression ratio: R8: for a bit rate of 8 which corresponds to a ratio of 3, R6: for a bit rate of 6 which corresponds to a ratio of 4, etc. to R0.25: for a bit rate of 0.25 which corresponds to a ratio of 96.

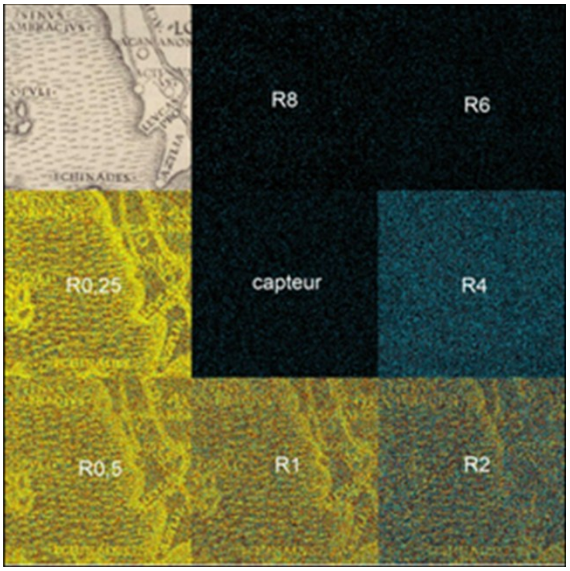

Figure 1 : part of a map (false color comparaison)

From this situation, we measured the difference between the RGB values of each pixel of each JP2 image obtained and the corresponding pixel of the original TIFF image. The result was shown in false color to highlight the differences and facilitate the comparaison (see fig. 1).

The curves obtained (see fig. 2) have two different shapes: the elephant forms (R0.25, R0.5, R1, R2 in the example) and the bell shapes (R4, R6, R8 in the example). The ratios corresponding to elephant curves have been eliminated. They generated visually perceptible defaults on the image (fig. 1), even if it was very subtle and light (R2 on fig. 1).

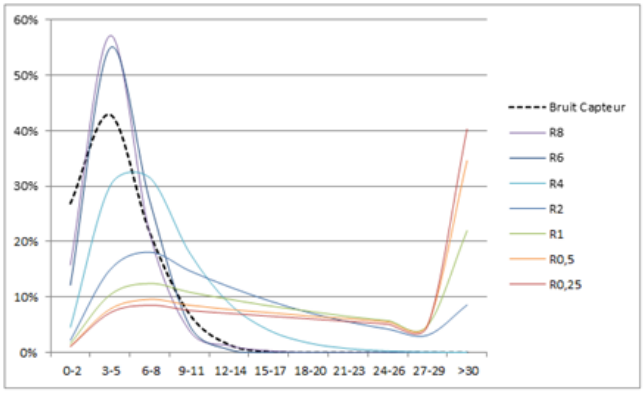

Figure 2 : errors distribution from a map

At this stage, the question was to find a discriminating criteria between the different ratios candidates. As it can be seen in the figures, the ratios of the curves in the form of bells cause diffuse errors not localized, over the whole image (see fig. 1 for the ratios R8, R6 and R4). From this point of view, the ratio R4 was a good candidate even with a large errors distribution.

In addtion, as we know, all image sensors generate noise [7] [8] [9]. To do find a reference, we scanned a gray image (RGB values: 128, 128, 128) with a SINAR 54 H (matrix of 22 mpx) and measured differences between RGB values and the known target values. The result is the zone at the center in fig. 1 and the sensor curve ('bruit capteur') in fig. 2. We refined our best ratio with this new reference. The principle is to stay in the same error zone that the noise image sensor. So, we selected the highest ratio while remaining inferior of errors level generated by the sensor itself.

In fig. 2 this is characterized by curves (R8 and R6) which have the same profile as the one of the image sensor, and in fig. 1 we can visually estimate the threshold at which the errors generated by the *JP2* compression algorithm are greater than the noise errors generated by the image sensor.

Thus, in production, it is not possible to distinguish the errors produced by the *JP2* compression algorithm from those produced by the image sensor itself. This also means that errors produced by the compression algorithm are the same types of errors than those produced by the sensor.

It is a pragmatic approach. Its limit is based on a capture material which is bound to improve.

Colorimetric profile

The color profile as described in the BnF's image standard [10] must be present in the file. As we saw earlier, the information on the ICC profile raised issues which has been solved since. Because this information is essential for color reproduction, we have decided to apply strict controls over how to record this information. It is entered in the "Color Specification" Box of the JP2 header using a "Restricted ICC" description of the "Display Device Profile" class.

Resolution levels

The resolution level determines the number of decompositions that will be applied to the image and on which a wavelet transform will be performed. This principle is at the heart of JPEG2000 compression. It should be at least 3 levels. The resolution level *n* makes it possible to obtain an image with dimensions 1 / 2^n of the original For instance with 5 levels, the dimensions of the reference image will be 1/32 of the original image. In order to have a wide flexibility in future use and to not have to reprocess the images for the diffusion, we chose for 10 levels of quality. In the metadata, it is written in a "Capture resolution box" of the *JP2* header and expressed in "pixel per meter".

Quality Level

The number of quality levels makes it possible to present an degraded image without waiting to decompress the entire image. This possibility may be useful for adjusting the data flow to be transmitted with regards to a possible bit rate. For instance, this can be the case for quickly delivering an image to a mobile device (tablet, smartphone). Moreover, up to a certain level of resolution, it is not necessary to obtain the details which are not essential for viewing the image.

The quality levels are not useful in the context of the BnF's digital library (Gallica). But we found that the quality level does not affect performances while generating and treating images. Also, we chose for 10 quality levels in order to leave an open choice to future applications.

Progression order

This parameter determines the order in which packets are stored according to four criteria: quality level (L = layer), resolution level (R = resolution), color (C = component) and position (P = position or Precincts). Different progression orders are possible. All operations are in the in the color space (Y, Cb, Cr) luminance / chrominance. The progression order is important for the exploitation of the images. Indeed, the choice of the order is driven by ones' display needs. Therefore it is a question of finding the best compromises. A bad choice of the progression order does not prevent exploitation of the image nor limits the possible functionality with the JPEG2000 file but forces the decompression algorithm to read more packets in order to obtain the desired effect and to increase processing times.

For the BnF to favor the improvement of the color components is not of great interest. Similarly, the progressive improvement of quality is not a goal. Indeed, we seek to present the best quality possible. The resolution level for quick access to a given resolution level and the position for extracting image portions correspond to features required for viewing in Gallica. In this perspective, the RPCL order is to be preferred.

Lessons learned from mass production

With mass digitization, it is necessary to set up automated quality control process. We quickly see what means, process and quality control recommendations are needed to control mass production (several million pages per year). The recommendations issued by the working group have become contractual requirements of our public procurement contracts. Its choices have been summarized in an online reference document [10]. The flip-flop, that is to say the transition from a production of images to the uncompressed TIFF file format to a production in the JPEG2000 file format was carried out over the 2014-2015 period; At the option of renewing the operations of digitization of collections: printed, press, specialized (manuscripts, prints, etc.), microfilms, ...

As part of the implementation of our specifications for the production of JPEG2000 files, we found that the providers (for the most part) were not prepared for this type of processing, ie they had neither the technical knowledge nor the tools to produce JPEG2000 files that meet our requirements. This observation led us to strongly involve ourselves in the implementation.

Strict control of production is a major challenge for the BnF. Thus, all images are automatically checked before being ingested in the preservation system (SPAR) and the Gallica digital library. In this case, the efficiency of the JPEG2000 algorithm was problematic. Indeed, it is not possible to predict the size of the file precisely. Images with low information content are very well compressed by the JPEG2000 algorithm sometimes more than the given to achieve and without more loss; ie. a blank will be heavily compressed. This ability, which is very useful in general, makes quality control very complex. So, How to ensure that the required parameters have been reached? To solve this problem, first we set margins of tolerance (5%). Second, images whose compression level is out of tolerance range, we require lossless compression (wavelet transformation 5-3 integer, Reversible Component Transform). This process allows us to ensure that we control our production, which is a major challenge for long-term preservation. Obviously, the price paid is to have images in lossless compression which is not optimal. Except that the images concerned are few in number (less than 5%) and that they have low information content (usually images with large uniform parts) and therefore very compressed as we have seen.

Conclusion

After more than two years of experience, we can conclude that it is possible to have a mass production of JP2 files resulting from digitization in a completely controlled way.

References

[1] Johan van der Knijff, JPEG 2000 for Long-term Preservation: JP2 as a Preservation Format, doi:10.1045/may2011-vanderknijff (2011)

[2] Johan van der Knijff, JP2 OPF File Format Risk Registry http://wiki.opf-labs.org/display/TR/JP2#JP2-Formatissues (visited march 20, 2017)

[3] Robert Buckley, JPEG 2000 as a Preservation and Access Format for the Wellcome Trust Digital Library (2009)

[4] ITU-T, Information technology – JPEG 2000 image coding system: Core coding system : Amendment 6: Updated ICC profile support, bit depth and resolution clarifications (03/2013)

[5] David Salomon, Data Compression: The Complete Reference (03/2013)

[6] David Taubman and Michael Marcellin, JPEG2000 Image Compression Fundamentals, Standards and Practice, ISBN-13: 978-0792375197 (2002)

[7] R.I. Hornsey, Noise in Image Sensors, https://ece.uwaterloo.ca/~ece434/Winter2008/Noise.pdf, (2008)

[8] Boyd Fowler, Dan McGrath, and Peter Bartkovjak, Read Noise Distribution Modeling for CMOS Image Sensors (2013)

[9] Boyd Fowler and Xinqiao (Chiao) Liu, Charge Transfer Noise in Image Sensor (2007)

[10] http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.numerisation_referentiels_bnf.html#SHDC__Attribute_BlocArticle4BnF

Author Biography

Laurent DUPLOUY is head of the digitization service of the National Library of France