Choisir un format d'images numériques dans le cadre de la numérisation patrimoniale[1]

Théophile Naito

Introduction

En raison des opportunités rendues possibles par les nouvelles technologies et sous la pression initiale du projet de numérisation entrepris par Google,[2] les bibliothèques et les services d'archives ont entrepris et préparent régulièrement des projets de numérisation des documents sous leur responsabilité.

Dans ce contexte, les archivistes et bibliothécaires responsables se trouvent devant des choix techniques qui ne faisaient pas partie de leurs tâches il y a quelques années encore, ce qui ne va pas sans difficulté.

Tout comme il est indispensable de disposer de compétences en paléographie lorsque l'on gère des documents manuscrits anciens, il est nécessaire de détenir un savoir dans le domaine des formats d'images numériques lorsque l'on se lance dans la production et la conservation d'images numériques.

Ainsi, la première partie de cet article aborde quelques-uns des principaux éléments d'un format d'image. En particulier, les algorithmes de compression et la notion de profil ICC (utile pour gérer les couleurs des images numériques) sont abordés.

Ensuite, l'article présente brièvement les formats d'images les plus courants dans les institutions patrimoniales.

Finalement, un processus de décision permettant d'aboutir au choix d'un format d'images en fonction du contexte est présenté, après une discussion de méthodes déjà existantes.

Il est important de noter qu'il s'agit ici de choisir un ou plusieurs formats d'images pour un projet de numérisation à venir, et qu'il ne s'agit pas de déterminer si des images déjà produites doivent être converties en un format mieux adapté.

1 Cet article est une adaptation et une mise à jour du travail de master MAS ALIS, intitulé « Images numériques matricielles à la Bibliothèque de Genève : TIFF ou JPEG 2000 ? », et écrit sous la direction d'Alexis Rivier (Bibliothèque de Genève).

2 Jacquesson, Alain : Google Livres et le futur des bibliothèques numériques. Paris 2010.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Les images numériques matricielles

Le concept d'image numérique matricielle

Une image numérique matricielle est une image, codée numériquement, obtenue par la description d'un ensemble de petits carrés de couleur unie placés l'un à côté de l'autre dans un tableau rectangulaire. L'idée est d'utiliser des carrés suffisamment petits pour que l'œil ne se rende pas compte que l'image est en réalité un assemblage de ces carrés.

On voit cet assemblage de carrés dans l'image de la figure 2 qui est un détail de l'image de la figure 1. Chacun de ces carrés est appelé « pixel », par contraction de « picture element ».

Toute image numérique n'est pas nécessairement matricielle. En effet, il est également possible de décrire une image sous une forme vectorielle. Une image vectorielle est une image composée d'éléments géométriques de base, tels des segments de droite ou des arcs de cercles. L'avantage d'une description vectorielle, pour les images s'y prêtant bien, est double. D'une part, cela permet d'obtenir des fichiers de taille modeste puisqu'il est bien plus court de spécifier quelques caractéristiques (taille, couleur, position) d'un nombre relativement faible d'objets géométriques, que de spécifier la couleur d'une très grande quantité de pixels. D'autre part, il est possible de changer l'échelle d'une image vectorielle facilement et sans perte de qualité (dans le cas d'un zoom par exemple), alors que cela n'est pas possible sans perte de qualité pour une image matricielle. Cela se voit dans l'image de la figure 2, qui montre qu'un zoom important fait apparaître les pixels.

Toutefois, les descriptions vectorielles sont bien adaptées pour les images de synthèse. Pour les images naturelles obtenues à l'aide d'appareils photographiques ou de scanners de documents, une description matricielle est la règle.

Il est à noter que les images matricielles sont souvent appelées images bitmap ou images raster.

Formats d'images numériques matricielles

La description des pixels d'une image matricielle est souvent accompagnée par d'autres informations concernant l'image. Citons la date de création de l'image, la dimension de l'image, des informations concernant les couleurs utilisées dans l'image, l'auteur de l'image. Il est évident que toutes ces informations doivent être organisées. De même, les données décrivant l'image doivent être structurées et codées d'une manière bien définie. Dans le cas contraire, aucun logiciel ne pourrait lire l'image concernée. Cette organisation et ce codage sont déterminés par un format de fichier. Pour les images matricielles, il existe une très grande quantité de formats.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Figure 1 : Fleurs

Figure 2 : Fleurs (détail) : image formée par des petits carrés de couleur homogène

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

A titre d'exemple,[3] une organisation simple en deux blocs, que l'on retrouve dans certains formats d'images matricielles est celle représentée dans la figure 3.

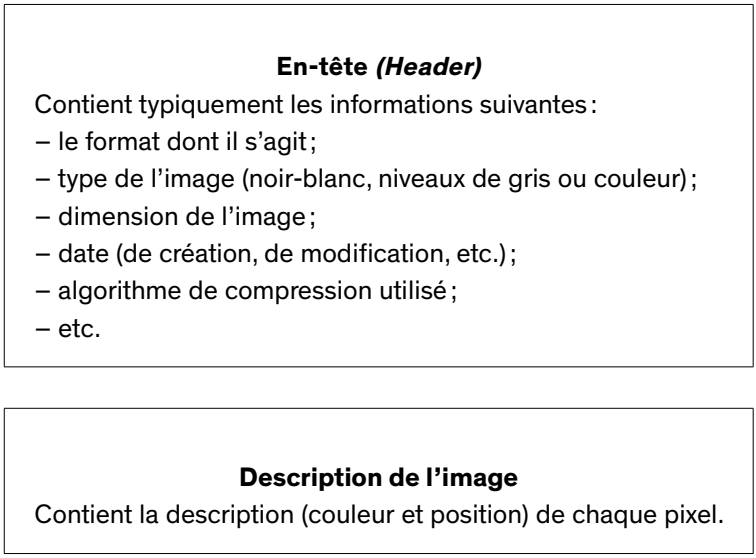

Figure 3 : Organisation en deux blocs d'un format d'image matricielle

D'autres formats adoptent une structure plus complexe, mais qui a l'avantage de s'adapter à de nombreux besoins. C'est le cas du format TIFF, dont on donne brièvement une idée de la structure. Un fichier respectant ce format débute toujours par un en-tête (header), qui contient quelques informations de base et qui pointe vers un répertoire (image file directory). Celui-ci contient l'essentiel des métadonnées liées à l'image, et il indique où se trouvent les données décrivant cette image et où se trouve le prochain répertoire s'il y en a un autre. On peut relever que les données relatives à une même image peuvent être structurées de différentes façons, puisque le répertoire peut être situé avant ou après les données décrivant l'image.

3 Murray, James D. ; Van Ryper, William : Encyclopedia of graphics file formats. Sebastopol Calif. 1996. Disponible en ligne www.fileformat.info/mirror/egff/index.htm (Partie 1 et 3), www.fileformat.info/format/all.htm (liste de formats permettant l'accès aux articles de la partie 2), (consultées le 22 juillet 2013). Ce livre présente de nombreux exemples et des détails supplémentaires.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Algorithmes de compression[4]

Décrire une image matricielle nécessite une grande quantité de données. En effet, puisqu'une image est composée d'un très grand nombre de pixels qui ont tous une couleur parmi un nombre de couleurs qui peut être gigantesque, le poids d'une image peut être très important. Par exemple, à une résolution de 300 ppp,[5] une image de 10 cm x 15 cm est composée de plus de 2 millions de pixels. Si chaque pixel peut avoir une couleur dans un ensemble de plus de 16 millions de couleurs (un cas tout-à-fait usuel pour les images naturelles), il est alors nécessaire de disposer de 24 bits par pixel.[6] Un calcul montre que l'on obtient un fichier d'un poids de plus de 6 Mo.

C'est considérable, et cela pose la question du stockage des images lorsqu'elles sont en grande quantité, et aussi celle de leur transmission à travers un réseau. Pour y répondre, une intense activité de recherche est menée dans le domaine de la compression des données numériques. En effet, on peut (et il faut !) se demander s'il est possible de décrire une image de manière plus économique. Les succès dans ce domaine de recherche sont grands, et il est aujourd'hui courant de compresser efficacement les images matricielles avec toute sorte d'algorithmes de compression.

Essentiellement, il existe deux types d'algorithmes de compression : les algorithmes « sans perte », et les algorithmes « avec perte ». Les algorithmes sans perte permettent de conserver la totalité de l'information originale, alors que les algorithmes avec perte ne permettent pas de retrouver l'image originale. L'utilisation de ce deuxième type d'algorithme permet des taux de compression spectaculaires.

Mais la compression est un sujet délicat dans le monde des archives et des bibliothèques puisqu'elle est souvent considérée comme un élément à éviter dans le cadre de l'archivage à long terme.[7]

De sorte à permettre une meilleure appréhension du sujet, quelques algorithmes simples et standards sont brièvement présentés dans la suite de cet article.

Pour ce faire, rappelons que toute information numérique se présente sous la forme d'une suite de 0 et de 1. Une telle suite est appelée un mot dans l'alphabet $\{0, 1\}$. Le nombre de 0 et de 1 qui forment un mot est appelé la longueur de ce mot. Un algorithme de compression a comme objectif de remplacer un mot contenant une

4 Pu, Ida Mengyi : Fundamental data compression. Amsterdam 2006. Salomon, David ; Motta, Giovanni : Handbook of data compression. Londres 2010. Le travail d'I. M. Pu est une excellente introduction aux algorithmes de compression, alors que celui de D. Salomon et G. Motta tend vers une description de tous les principaux algorithmes existants. Toutes les informations de cette partie de l'article, et bien plus encore, peuvent être trouvées au moins dans l'un de ces livres.

5 Pixels par pouce ; un pouce valant 2.54 cm.

6 En effet, $2^{24} = 16\,777\,216$.

7 Büchler, Georg ; Kaiser, Martin (CECO) : Kolloquium Datenkomprimierung bei Bild, Audio, Video. Berne 2009.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

information d'intérêt par un mot d'une longueur plus faible. De plus, ce remplacement doit se faire sans perte d'information, ou avec une perte acceptable. Pour obtenir cet effet de compression, l'idée est de supprimer toutes les formes de redondance qui apparaissent.

Il est à noter qu'il est impossible de définir un algorithme de compression capable de remplacer n'importe quel mot par un mot de longueur plus faible. Plus précisément, cela est impossible sans perte d'information. En d'autres termes, pour tout algorithme de compression sans perte d'information, il existe au moins un mot que l'algorithme n'arrive pas à remplacer par un mot de longueur plus petite (en fait, il est possible de prouver que pour tout algorithme sans perte, il existe un nombre infini de mots qui ne peuvent pas être compressés en des mots de longueur plus faible).

Cela veut dire qu'avant de définir un algorithme de compression, il faut déterminer quel type d'information il s'agit de compresser. Certains algorithmes sont efficaces pour compresser du texte, alors que d'autres sont particulièrement efficaces pour compresser des images. Et parmi les algorithmes efficaces pour compresser les images, certains sont spécifiquement construits pour compresser des images qui représentent du texte en noir/blanc, alors que d'autres sont efficaces pour des images non-textuelles. En effet, tout type de données a un genre de redondance particulier dont il s'agit de tirer profit au mieux. Comme premier exemple, voyons l'algorithme de compression suivant.

Codage par plages – Run-length encoding (RLE)

Cet algorithme de compression vise typiquement les images en noir/blanc, ou tout autre type d'information dont la représentation numérique comporte une majorité de 0 (et de 1) qui se suivent. Par exemple, le mot

00000000000000000000000000000001111100000000111100000000000000000000000

pourrait être une partie du codage d'une ligne de pixels dans une image qui représente du texte, en noir, sur un fond blanc. Les 1 représentent les pixels en noir, moins fréquents que les pixels en blanc. L'idée de l'algorithme RLE est de tirer profit des grandes plages de 0 (et de 1), en écrivant

32 * 05 * 09 * 04 * 24.

Ce mot doit être compris comme signifiant « 32 zéros, puis 5 uns, puis 9 zéros, puis 4 uns, puis 24 zéros ». En écriture binaire, il devient :

00011111100001000000 1000001100010111 ;

sachant que l'on a simplement écrit les nombres 31, 4, 8, 3, 23 en binaire sur les 7 derniers bits de chaque groupe de 8 bits (les espaces sont là pour faciliter la lecture de ces groupes). Le premier bit de chaque groupe de 8 bits indique s'il s'agit d'une

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

plage de 0 ou de 1. Il est évident que l'algorithme a effectivement permis d'obtenir un effet de compression : au lieu d'être écrite sur 74 bits, la même information a pu être compressée sur 40 bits.

La faiblesse évidente de cet algorithme réside dans le fait qu'il est incapable de compresser efficacement des mots qui contiennent peu de grandes plages de 0 et de 1. Dans ces cas, l'algorithme augmente la longueur des mots ! Un autre point important est que la ligne originale de 0 et de 1 peut aisément être récupérée à partir du mot obtenu par cette méthode de compression. Il n'y a aucune perte d'information dans ce processus.

Codage de Huffman

Très souvent, chaque caractère d'un alphabet est codé avec le même nombre de bits. C'est le cas de la norme de codage ASCII, qui attribue 8 bits à chaque caractère textuel. C'est également le cas lorsque l'on code un ensemble de couleurs sans compression. On emploie fréquemment 24 bits pour coder une couleur. Toutefois, autant dans le cas d'un texte que d'une image, certains caractères apparaissent plus souvent que d'autres. L'idée de David Albert Huffman, en 1952, a été de coder les caractères apparaissant fréquemment avec un nombre plus faible de bits en suivant un processus rigoureux. Ainsi, le caractère apparaissant le plus souvent dans la langue française, « e », sera codé sur un seul bit lorsqu'il s'agit de coder des textes écrits dans cette langue.

Pour un texte particulier, il est possible que ce ne soit pas le caractère « e » qui apparaisse le plus souvent. Pour coder ce texte, il faudrait d'abord calculer les fréquences de chaque caractère pour en déduire le codage approprié. Deux difficultés se présentent dans cette situation. D'une part, pour que le décodeur puisse lire le texte, il faut que le codeur lui transmette la table de codage.[8] Cela augmente le poids du fichier et réduit l'efficacité de la compression. D'autre part, le calcul préliminaire des fréquences augmente le temps de codage, ce qui peut être malvenu. Pour éviter ces deux difficultés, c'est une version adaptative du codage de Huffman qui est utilisée. L'idée est de commencer le codage sans utiliser de compression, et ensuite d'adapter la table de codage à chaque caractère codé. Cela se fait de façon à ce que le décodeur puisse reconstruire la table de codage au fur et à mesure du décodage, sans que le codeur n'ait besoin de lui fournir une information.

De la même manière, comme la fréquence d'apparition des couleurs dans une image dépend de celle-ci, c'est un codage adaptatif qui est souvent utilisé pour coder les images.

8 Une table de codage est la correspondance entre chaque caractère et le code qui le représente. Par exemple, e = 1, s = 01,…, est une table de codage.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Compressions « Groupe 3 » et « Groupe 4 »

Ces deux algorithmes ont été pensés pour la transmission de documents en noir (l'information) et blanc (le papier) par fax. Ce sont des standards publiés par l'Union Internationale des Télécommunications (UIT), et leur dénomination officielle est UIT-T T.4 (compression « Groupe 3 ») et UIT-T T.6 (compression « Groupe 4 »).[9]

La norme « Groupe 3 » consiste en fait en deux algorithmes de compression différents : une compression unidimensionnelle et une compression bidimensionnelle. Le premier de ces algorithmes considère chaque ligne de pixels de manière indépendante des autres lignes, ce qui fait le caractère unidimensionnel de cette méthode. La compression se fait en deux temps. D'abord, il s'agit de procéder à un codage par plages, et ensuite c'est un codage de Huffman qui est appliqué en considérant chaque plage comme un caractère d'un alphabet. Il ne s'agit pas d'un codage adaptatif, mais d'un codage basé sur un set de documents représentatif défini par l'UIT. A titre d'exemple, le tableau de la figure 4 est une partie de la table de codage définie par le standard UIT-T T.4.

Longueur de la plage blanche	Code	Longueur de la plage noire	Code
0	00110101	0	0000110111
1	000111	1	010
2	0111	2	11
3	1000	3	10
4	1011	4	011

Figure 4 : Extrait du Tableau 2/T.4 – Codes de terminaison [10]

La compression bidimensionnelle distingue différentes situations. Dans le meilleur des cas, il y a une forte redondance verticale en raison de la nature des documents visés (les documents transmis par fax) et l'idée est de coder une ligne de pixels par

9 Pour compliquer les choses, ces algorithmes sont aussi connus sous le nom de CCITT T.4 et CCITT T.6. En effet, c'est l'organe de l'UIT connu sous le nom de Comité consultatif international téléphonique et télégraphique (CCITT) qui est à l'origine de ces standards. Aujourd'hui, le CCITT n'existe plus, remplacé par le Secteur de la normalisation des télécommunications de l'UIT (UIT-T). De plus, on peut aussi écrire « ITU » (anglais) au lieu d'écrire « UIT » (français). Ces deux standards sont disponibles gratuitement sur le site web de cette organisation.

10 Secteur de la normalisation des télécommunications de l'UIT : Recommandation UIT-T T.4. Genève 2004, 6.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

rapport à la ligne de pixels qui se trouve immédiatement au-dessus. Lorsqu'une telle redondance n'existe pas sur toute ou une partie d'une ligne, alors c'est le codage unidimensionnel décrit ci-dessus qui est employé. De plus, la compression « Groupe 3 » est définie de sorte à pouvoir supporter des erreurs de transmission. Une des mesures prises dans ce cadre est le codage d'une ligne sur deux (ou sur quatre) selon la méthode unidimensionnelle, ce qui permet d'éviter qu'une seule erreur se propage dans toute la suite du document et le rende incompréhensible.

Cette façon de faire limite l'efficacité de la compression puisque l'algorithme renonce volontairement à exploiter la redondance verticale sur tout le document. Pour remédier à ce fait lorsqu'une résistance aux erreurs n'est pas nécessaire, la compression « Groupe 4 » reprend le même processus que la compression bidimensionnelle « Groupe 3 », en supprimant certains des mécanismes utiles dans des environnements propices aux erreurs. En particulier, l'entier du document est codé suivant le codage bidimensionnel, ce qui permet un taux de compression environ deux fois meilleur.

Algorithmes de compression basés sur un dictionnaire (algorithmes de la famille LZW)

Une famille d'algorithmes d'un type différent existe. Ce sont les algorithmes basés sur l'utilisation d'un dictionnaire. La stratégie consiste à parcourir le mot à compresser à la recherche de chaînes de caractères qui apparaissent dans un dictionnaire. Lorsqu'une telle chaîne de caractères est trouvée, elle est remplacée par le numéro d'index[11] de la chaîne. Par exemple, admettons qu'il existe un dictionnaire pour le français contenant 99 999 mots au plus. Et supposons que le mot :

— « algorithme » arrive en 2415ème position dans ce dictionnaire ;
— « compression » arrive en 10 324ème position ;
— « de » arrive en 11 112ème position ;
— « dictionnaire » arrive en 12 956ème position ;
— « un » arrive en 74 454ème position.

Dans ce cas, la phrase

« L'algorithme de compression LZW utilise un dictionnaire. »

peut être codée de la manière suivante :

« L'*2'415*11 112*10 324*LZW*utilise*74 454*12 956*. »,

car « L' », « LZW », « utilise » et « . » ne se trouvent pas dans le dictionnaire considéré.

Examinons l'efficacité de cet algorithme. La phrase qu'il faut coder est constituée de 56 caractères. En utilisant le code ASCII, on constate que 448 bits sont requis pour coder ce texte. Calculons maintenant le nombre de bits nécessaires

11 Il s'agit du nombre permettant de situer la chaîne dans le dictionnaire.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

lorsque l'on profite du dictionnaire. Comme il faut 17 bits[12] pour coder les nombres jusqu'à 100 000, on doit utiliser 18 bits pour coder les mots qui se trouvent dans le dictionnaire. En effet, il faut 1 bit supplémentaire pour indiquer si ce qui suit correspond à un index ou à un mot codé en ASCII. Les mots présents dans le dictionnaire étant au nombre de 5, cela nous donne 90 bits. De plus, les quatre chaînes de caractères codées en ASCII nécessitent respectivement 22, 30, 62 et 14 bits, puisqu'il faut 1 bit pour indiquer que ce qui suit est codé en ASCII, 5 bits pour indiquer la longueur du code ASCII, puis le code ASCII. On arrive à un total de 218 bits. On constate qu'il a effectivement été possible de compresser efficacement la phrase proposée.

Ci-dessus, c'est un dictionnaire traditionnel qui a été utilisé. Pour augmenter l'efficacité de l'algorithme, les dictionnaires utilisés contiennent en réalité des chaînes de caractères[13] qui n'ont pas nécessairement de signification. De plus, la conception à l'avance d'un dictionnaire peut être irréalisable lorsque les données ne sont pas des données textuelles. C'est typiquement le cas des images. Il n'est pas évident qu'il soit possible de trouver des arrangements de couleurs qui reviennent régulièrement dans toutes les images. Pour cette raison, il est nécessaire de définir des algorithmes capables de construire un dictionnaire différent pour chaque image. Pour éviter la transmission du dictionnaire entre le codeur et le décodeur de l'image, il s'agit de le construire selon une procédure permettant au décodeur de reconstruire le dictionnaire au fil du décodage.

Par exemple, le codeur peut démarrer avec un dictionnaire vide ou par défaut (constitué des lettres de l'alphabet par exemple). Ensuite, en cours de codage, le codeur cherche dans le dictionnaire le mot le plus long qui correspond à la chaîne de caractères qu'il doit coder. Supposons que la prochaine chaîne de caractères à coder soit « couleurs » mais que seul le mot « coule » soit déjà présent dans le dictionnaire. Dans ce cas, le codeur code uniquement le début, à savoir « coule », du mot « couleur », et le codeur précise que la lettre suivante est le u. Ensuite il ajoute le mot « couleu » au dictionnaire, et il s'attaque à la prochaine chaîne de caractères à coder. Dans notre cas, il s'agit de « rs » et de ce qui suit. Ainsi, le dictionnaire est en constante évolution, en fonction des redondances qui apparaissent.

Parmi les algorithmes faisant appel à une procédure de ce genre, on trouve les algorithmes précurseurs LZ77 et LZ78, publiés par A. Lempel et J. Ziv en 1977 et 1978 respectivement. Par la suite, une variante de ces algorithmes a été publiée en 1984 par T. Welch. Cet algorithme est connu sous le nom de LZW et est probablement le plus fameux des algorithmes de la famille LZ, qui regroupe les variantes de LZ77 et de LZ78.

12 En effet, $2^{17} = 131\,072$.

13 Ou des chaînes de 0 et de 1.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Plutôt qu'une efficacité ciblée sur un type de donnée très particulier, les algorithmes de la famille LZ ont tendance à être efficaces sur des données de tout type. Cette caractéristique en fait un algorithme « tout-terrain », très utile lorsqu'il s'agit de compresser différents types de données ensemble. Toutefois, lorsqu'il s'agit de compresser un type de données bien précis, il existe souvent un algorithme plus efficace.

Algorithmes de compression JPEG, JPEG 2000 et JBIG2

Les méthodes de compression JPEG (pour les images en couleur ; compression avec perte), JPEG 2000 (pour les images en couleur, compression sans ou avec perte) et JBIG2 (pour les images noir/blanc, compression sans ou avec perte) font appel à des mathématiques avancées, et ne peuvent donc être résumes dans le cadre du présent article. Ces algorithmes bénéficient de techniques récentes et sont particulièrement efficaces.

Gestion des couleurs

Pour certaines utilisations, les couleurs des images ne sont pas très importantes, voire superflues. Ainsi, la fidélité exacte des couleurs est sans valeur pour le lecteur intéressé uniquement par le texte d'un document. Pour ce genre d'usages, il peut même être avantageux de ne pas prendre en compte la couleur et de se contenter d'une numérisation en noir/blanc. Le poids des images numérisées en noir/blanc, nettement inférieur au poids des mêmes images numérisées en couleur, est un atout important. Dans le même genre de cas, l'impression est également une fonction pour laquelle une numérisation sans couleur peut être profitable. Par exemple, cela permet de réduire l'utilisation du toner des imprimantes.

Mais les bibliothèques et les archives peuvent tout à fait être confrontées à des situations dans lesquelles le respect des couleurs originales est important. Cela peut être le cas lorsqu'un utilisateur publie des images de documents. Par exemple, des images peuvent être publiées dans un catalogue d'exposition. Il peut aussi arriver que quelqu'un ait comme projet la création d'un fac-similé qui soit le plus fidèle possible au document original. Et si la numérisation a comme but de créer une copie pouvant remplacer le document original, alors la fidélité de la reproduction est primordiale. En effet, une institution patrimoniale peut estimer que certains originaux sont dans un tel état que la création d'une copie est nécessaire pour assurer la pérennité de l'information. Pour ces situations, il est nécessaire de savoir comment les formats d'images gèrent les couleurs.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Espace de couleurs et profil ICC

Aujourd'hui, de nombreux appareils gèrent des couleurs. Les imprimantes, les écrans, les appareils photographiques et les scanners sont dans ce cas. Dans le but de transmettre fidèlement les couleurs entre ces appareils, différents standards ont été émis par les organismes compétents. En l'occurrence, il s'agit de la Commission Internationale de l'Eclairage (CIE) et de l'International Color Consortium (ICC).

Le concept d'espace de couleurs est important. Un espace de couleurs est une correspondance définie entre un ensemble de couleurs et un ensemble de nombres. Cette correspondance permet de coder les couleurs avec des nombres. Plus précisément, il s'agit de faire correspondre un triple[14] de nombres à chaque couleur. Par exemple, l'espace RVB (RGB en anglais) décompose une couleur C en un rouge R, un vert V et un bleu B. A la couleur C est associé le triple (r,v,b) où r est le nombre associé à la couleur R, v est le nombre correspondant à la couleur V et b est le nombre décrivant la couleur B. Concrètement, (255,0,0) correspond au rouge primaire alors que (110,11,20) décrit la couleur grenat.[15] Il est toutefois nécessaire d'être très prudent : il existe un grand nombre d'espaces RVB. Parmi les espaces de ce type qui sont souvent utilisés se trouvent sRGB et Adobe RGB (1998).

Cela veut dire qu'un même triple peut très bien correspondre à des couleurs différentes en fonction de l'espace de couleurs considéré. De plus, deux espaces différents ne couvrent pas nécessairement le même gamut.[16] Par exemple, l'espace de couleurs sRGB a un gamut plus petit que l'espace Adobe RGB (1998), ce qui veut dire qu'il existe des couleurs que l'on peut décrire dans l'espace Adobe RGB (1998) alors qu'elles ne peuvent pas être décrites par l'espace sRGB.

Examinons un cas concret qui rappelle que chaque appareil gère son propre espace de couleurs. Prenons le cas d'un utilisateur qui visionne sur un écran une image numérisée à l'aide d'un scanner. Supposons que certains pixels de l'image aient été codés comme du rouge primaire par le scanner : (255,0,0) dans l'espace de couleurs RVB du scanner. Si cette image est fournie sans aucune précaution, l'écran affichera le rouge primaire (255,0,0) correspondant à son propre espace RVB. Et cet espace n'a aucune raison d'être le même que celui du scanner. Par conséquent, l'image que l'utilisateur visionne sur l'écran ne correspond pas, en termes de couleurs, au document qui a été numérisé.

De sorte à pouvoir maintenir la fidélité des couleurs tout au long d'une chaîne allant de la production des images à leur visualisation, on utilise un espace de couleurs intermédiaire et standardisé. Différents espaces de couleurs standardisés par

14 C'est le cas le plus fréquent, mais d'autres modèles sont possibles. Le modèle CMYK, utilisé par les imprimantes, fait usage de quadruples.

15 Selon fr.wikipedia.org/wiki/Liste de couleurs (consultée le 22 juillet 2013).

16 Sous-ensemble de l'ensemble des couleurs.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

la CIE existent. Par exemple, CIE L*a*b*, datant de 1976, est un tel espace. Ce standard décrit l'ensemble des couleurs qu'un humain perçoit. De plus, cet espace est conçu pour que la distance mathématique entre deux triples associés à deux couleurs corresponde à la différence de perception visuelle entre ces deux couleurs. A titre d'exemple, il y a la même différence de perception entre les couleurs représentées par (50, 100, 150) et (50,100, 200) et entre les couleurs représentées par (50,100, 200) et (50,150, 200).

L'existence d'un tel espace standard permet de respecter les couleurs tout au long d'une chaîne. Mais pour cela il faut introduire le concept de profil ICC, introduit par l'ICC et reconnu conjointement par l'Organisation Internationale de normalisation (ISO) et par l'ICC (ISO 15076–1). Un profil ICC établit la correspondance entre un espace de couleur d'intérêt (par exemple celui d'un scanner ou d'un écran) et un espace de couleur standardisé (CIE L*a*b* par exemple).

Retournons à l'exemple de l'utilisateur qui visionne sur un écran une image numérisée par un scanner. Cette fois, supposons que le scanner est muni d'un profil ICC. De même, supposons que l'écran dispose de son propre profil ICC. A nouveau, regardons le cas du pixel dont la couleur est codée (255,0, 0) par le scanner. Lorsque l'écran affiche ce pixel, il va suivre le processus suivant : il cherche d'abord à quelle couleur dans l'espace CIE L*a*b* correspond (255,0,0), en utilisant le profil ICC du scanner. Appelons « rouge » cette couleur. Puis il s'agit de voir à quoi correspond ce « rouge » dans l'espace de couleurs de l'écran, en utilisant le profil ICC de ce dernier. Ce pourrait être (252,10, 15) par exemple. Maintenant, le respect des couleurs est assuré tout au long de la chaîne !

Comme le rouge, codé (255,0,0), du scanner pouvait être représenté par l'écran, il a été possible de conserver cette couleur sur le chemin entre le scanner et l'écran. Toutefois, il peut arriver que le gamut de l'écran ne contienne pas le rouge indiqué par le scanner. Dans un tel cas, l'écran affiche le rouge le plus « proche » possible.

Pour que ce qui précède puisse fonctionner, il est nécessaire que les logiciels permettant l'affichage d'une image aient accès au profil ICC du scanner. Pour cela, il y a deux possibilités. L'une est d'intégrer le profil dans le fichier de l'image. Pour ce faire, il faut que l'image matricielle soit codée dans un format qui permet d'intégrer le profil ICC du scanner. La solution alternative est d'indiquer l'endroit où les logiciels peuvent trouver le profil ICC, sauvegardé de manière indépendante.

Il semble que la solution qui s'est imposée soit celle de l'intégration du profil ICC dans le fichier de l'image. Cette solution a l'avantage de lier le profil avec l'image. Cela permet de faciliter la gestion de la correspondance entre les images et le profil ICC du scanner lors de l'échange d'images, mais aussi dans le cours de l'archivage. A l'inverse, il y a aussi un désavantage puisque l'on constate que le

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

même profil est susceptible d'être conservé en de multiples copies. Ce qui augmente légèrement l'espace de stockage nécessaire. A titre d'illustration, remarquons que la Bibliothèque de Genève (BGE) conserve actuellement plus de 800 000 images dans le cadre du projet e-rara. Chacune de ces images contient un profil ICC, alors que seuls 5 profils différents ont été utilisés. Sur la base d'un poids moyen de 265 Ko par profil ICC, il en résulte un poids supplémentaire d'environ 200 Go. Ce poids correspond approximativement à 1,5 % du poids total des 800 000 images, ce qui donne la mesure de l'économie qui pourrait être réalisée en n'intégrant pas les profils dans les images.

Les formats courants d'images

Il existe un très grand nombre de formats d'images, comme cela est illustré par le travail de J. Murray et W. van Ryper.[17] Dans le présent article, on se contente d'examiner l'utilisation des formats largement acceptés dans le monde des archives, des bibliothèques et au-delà. En effet, porter son choix sur un format rarement utilisé comporte des risques importants pour des raisons évidentes.

Le format TIFF (Tagged Image File Format)

Le format TIFF, propriété d'Adobe et publié pour la première fois en 1986 par Aldus, est un format dont la principale caractéristique est une grande souplesse. Par exemple, il peut être utilisé sans compression, mais il peut aussi être utilisé avec une compression sans perte (algorithme RLE, de Huffman, LZW, Groupe 3 et Groupe 4) ou avec perte (JPEG). Il peut également être utilisé en mode « multi-page », qui permet de conserver et transmettre plusieurs images en un seul fichier. En ce qui concerne les couleurs, il peut aller jusqu'à 48 bits par pixel, voire plus.

Toutefois, cette souplesse peut aussi être un obstacle : en effet, un logiciel de visualisation donné ne décode pas nécessairement toutes les versions possibles du format TIFF. Par exemple, les logiciels de visualisation peuvent se contenter d'afficher la première image d'un fichier multipage, sans capacité de lire les images suivantes.

Le format TIFF est répandu. En particulier, le secteur des archives et des bibliothèques a fait du format TIFF (TIFF 6.0 datée de 1992) sans compression le standard de fait dans le domaine de l'archivage à long terme.[18] De même, Google fait usage de ce format pour sa bibliothèque numérique « Google Livres » selon A. Jacquesson.[19]

17 Murray/Van Ryper, Encyclopedia of graphics file formats.

18 Büchler/Kaiser, Kolloquium Datenkomprimierung bei Bild, Audio, Video.

19 Jacquesson, Google Livres et le futur des bibliothèques numériques. La remarque concernant le format TIFF se trouve à la page 58.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

La spécification du format TIFF est divisée en deux parties. La première partie définit le format TIFF Baseline, et les Archives et bibliothèques exigent souvent que les images respectent ces exigences plus restrictives.

La spécification du format TIFF, gratuitement disponible sur le site web d'Adobe, ne prévoit pas l'intégration d'un profil ICC. Mais la flexibilité du format TIFF permet tout de même l'intégration d'un profil ICC dans un fichier respectant ce format. Ceci est expliqué dans l'annexe de la spécification du format ICC.[20] Toutefois, étant donné l'absence d'indication dans la spécification du format TIFF, les logiciels capables de décoder un fichier TIFF n'en tiennent pas nécessairement compte.

Le format JPEG (JPEG File Interchange Format)

En fait, JPEG n'est pas un format mais un algorithme de compression efficace, mis au point par le comité JPEG et normalisé par l'ISO et l'UIT en 1992 (ISO/IEC 10918–1 ou UIT-T T.81). L'algorithme JPEG peut être utilisé dans divers formats, tel TIFF ou PDF. Toutefois, il existe un format spécifiquement conçu pour JPEG : c'est le format JFIF (ISO/IEC FDIS 10918–5), et c'est à lui que l'on pense lorsqu'il est question de JPEG comme format d'image. Ce format permet de gérer des images en couleurs allant jusqu'à 24 bits par pixel.

JPEG est un algorithme de compression avec perte.[21] Autrement dit, un cycle de compression avec JPEG entraîne une modification de l'image. Mais l'efficacité de l'algorithme permet de garder ces modifications dans le domaine du raisonnable : il est possible d'obtenir un taux de compression d'environ 20 : 1[22] sans que l'œil humain ne s'en rende compte ! Lorsque le taux de compression est très élevé, une image compressée avec l'algorithme JPEG peut souffrir de l'apparition de petits carrés clairement visibles. Cela provient du fait que la compression se fait sur des blocs de 8 × 8 pixels.

En théorie, le taux de compression est au choix de l'utilisateur, mais un taux par défaut est souvent proposé par les logiciels.

L'efficacité de JPEG en fait un algorithme de compression et un format très répandu. Malgré la perte d'information que JPEG provoque, il est parfois utilisé pour conserver des images lorsque des questions de coûts et d'espace de stockage jouent un rôle important. Ainsi, comme exemple parmi d'autres, les Archives d'Etat de Genève (AEG) utilisent le format TIFF uniquement pour les cadastres et les plans,

20 Disponible en ligne : www.color.org/specification/ICC1v43_2010–12.pdf (Consultér le 22 juillet 2013).

21 En réalité, JPEG peut également compresser sans perte, mais l'efficacité est moindre et l'algorithme est totalement différent de celui qui compresse avec perte. De plus, peu de logiciels offrent la possibilité d'utiliser l'algorithme de compression JPEG sans perte.

22 Cela signifie que le poids du fichier compressé vaut 5 % du poids du fichier original non compressé.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

alors que les registres et les documents textuels sont numérisés et conservés selon le format JPEG.[23] En effet, les images sont considérées comme un moyen de diffusion, alors que les documents originaux sont seuls considérés comme documents à archiver pour le long terme. Ainsi, les AEG ont la possibilité de numériser une deuxième fois un document pour lequel il y aurait une demande qui ne pourrait pas être satisfaite par l'image respectant le format JPEG.

De plus, la compression JPEG est acceptée par le format PDF/A. Ce format, dont il est question plus loin dans cet article, est très largement accepté pour l'archivage à long terme.

En ce qui concerne la gestion des couleurs, comme pour le format TIFF, il est nécessaire de consulter l'annexe de la spécification du format ICC pour obtenir la manière de procéder pour inclure un profil ICC dans un fichier JPEG.

Le format JPEG 2000

A l'image de JPEG, JPEG 2000 est un algorithme de compression avant d'être un format. Toutefois, la spécification de JPEG 2000, rédigée par le comité JPEG, publiée et normalisée par l'ISO et l'UIT en 2004 (ISO/IEC 15444–1 et UIT-T T.800) contient la définition d'un format d'image en annexe. C'est de ce format, nommé JP2, que l'on parle lorsque JPEG 2000 est considéré comme un format.

Parmi les éléments différenciant JPEG 2000 de JPEG se trouve l'amélioration de la qualité visuelle des images. En effet, on obtient une qualité légèrement supérieure en utilisant JPEG 2000 pour un taux de compression équivalent. D'autre part, JPEG 2000 permet tant la compression sans perte qu'avec perte. La différence réside uniquement dans le degré de compression souhaité.

De plus, JPEG 2000 a été pensé pour faciliter toute une série de fonctionnalités. Par exemple, la transmission et la visualisation d'images à travers un réseau sont en principe facilitées. Toutefois, ce format ne s'est pas encore imposé aussi largement que JPEG, qui lui est encore très souvent préféré.

Dans le domaine de l'archivage à long terme, JPEG 2000 connaît un succès croissant,[24] et il est accepté par les formats PDF/A-2 et PDF/A-3.

L'inclusion d'un profil ICC dans un fichier JPEG 2000 est prévue par la spécification de JPEG 2000, avec certaines restrictions.[25]

23 Dunant-Gonzenbach, Anouk (AEG) : Politique et bonnes pratiques de la numérisation aux AEG. Genève 2010.

24 Van der Knijff, Johan : JPEG 2000 for Long-term Preservation : JP2 as a Preservation Format. In : D-Lib Magazine 2011.

25 Voir le blog de J.Van der Knijff dans le cadre de l'Open Planets Foundation. L'article discutant la question des profils ICC se trouve à la page www.openplanetsfoundation.org/blogs/2013–07–01-icc-profiles-and-resolution-jp2-update-2011-d-lib-paper (consultée le 22 juillet 2013).

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Le format PDF (Portable Document Format)

Le format PDF est un format créé par Adobe. Il a été normalisé par l'ISO en 2008 (ISO 32000–1). Le format PDF n'est pas un format d'image. C'est un format de type plus général dont le but est la description de documents, de sorte à pouvoir représenter un document donné exactement comme l'auteur l'a conçu. C'est un format très couramment utilisé, en particulier pour échanger des documents. Ainsi, de nombreuses bibliothèques numériques utilisent ce format.

Il existe une version du format, appelée PDF/A, qui a été standardisée par l'ISO en 2005, 2011 et 2012 (ISO 19005-1 pour PDF/A-1, ISO 19005-2 pour PDF/A-2 et ISO 19005-3 pour PDF/A-3), et dont le but est de satisfaire le mieux possible aux exigences de l'archivage à long terme. Ces trois standards sont valides en parallèle, l'un ou l'autre peut être utilisé selon les besoins.

Il est courant de convertir des documents numériques au format PDF/A lorsque l'on veut archiver ces documents pour le long terme, à l'image de la pratique des Archives fédérales suisses qui exigent le format PDF/A pour les documents bureautiques.[26]

Le format PNG

Le statut légal de l'algorithme de compression LZW, qui était protégé par des brevets, a incité un groupe ad hoc à créer un format pouvant remplacer le format GIF. C'est ainsi que la spécification de PNG a été publiée en 1996. Ce format est un standard depuis 2004 (ISO/IEC 15948). Quelques-unes des caractéristiques de ce format sont:

— un algorithme de compression sans perte efficace, surtout pour les images de synthèse ;
— des couleurs pouvant aller jusqu'à 48 bits par pixel ;
— de ne pas être restreint par des questions légales.

Le format PNG est souvent reconnu comme un format apte à l'archivage à long terme. Mais en raison de la concurrence des formats déjà vus ci-dessus, il est rarement utilisé dans ce cadre.

Le choix d'un format d'images pour un projet de numérisation : méthodes existantes

Différents auteurs se sont intéressés à la manière d'évaluer et de choisir un format d'images parmi d'autres formats. Par exemple, la Bibliothèque nationale des Pays-Bas [27] et le Centre de coordination pour l'archivage à long terme de documents élec-

26 Archives fédérales suisses : Formats de fichier adaptés à l'archivage. Normes et standards pour l'archivage de documents numériques. Berne 2007.

27 Rog/van Wijk, Evaluating file formats for long-term preservation. La Haye sans date.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

troniques (CECO)[28] ont étudié et établi de telles méthodes d'évaluation, orientées vers la conservation des images à long terme.

Une tendance se dégage de ces études : l'idée générale est de déterminer des critères importants pour la conservation à long terme, et ensuite de noter les différents formats envisagés vis-à-vis de ces critères. Cette façon de faire permet d'établir aisément un classement des formats en fonction des notes attribuées. De plus, le procédé pour arriver à ce résultat semble relativement facile à appliquer, puisqu'il suffit de suivre une « recette de cuisine ». Celle-ci consiste à confronter chaque format aux différents critères pour en tirer une note.

Par exemple, les critères définis par le CECO sont les suivants : « Ouverture du format », « Licence libre », « Diffusion », « Fonctionnalités », « Implémentation », « Densité de mémorisation », « Vérifiabilité », « Bonnes pratiques » et « Perspectives ». La signification plus précise de chacun de ces termes est expliquée dans le travail du CECO.

Ensuite, ce dernier attribue une note à chaque format et ce pour chaque critère. Finalement, un calcul de moyenne est fait en tenant compte du facteur de pondération attribué à chaque critère. Cette moyenne permet d'établir le classement suivant pour les formats d'images matricielles, étant précisé que les formats qui ne sont pas indiqués n'ont pas été évalués.

1.	TIFF 6.0 sans compression et PDF/A-2	Note : 1,51.
3.	JPEG, JPEG 2000 et DNG	Note : 0,89.
6.	PNG	Note : 0,73.

L'étude de la Bibliothèque nationale des Pays-Bas établit une méthode similaire, dotée de quelques raffinements. Ainsi, les sept critères *(Openness, Adoption, Complexity, Technical Protection Mecanism (DRM), Self-documentation, Robustness, Dependencies)* sont chacun partagés en différentes caractéristiques. Par exemple, le critère *Openness* admet les trois caractéristiques suivantes : « Standardisation, Restrictions on the interpretation of the file format » et « Reader with freely available source ». Finalement, un type de moyenne pondérée est établi à partir des notes qui correspondent aux diverses caractéristiques. Cette moyenne permet d'aboutir au classement suivant, seuls les formats indiqués ayant été évalués.

1.	TIFF (Baseline 6.0 sans compression)	Note : 84,8.
2.	PNG 1.2	Note : 78.

28 KOST : Catalogue des formats de données d'archivage (Cfa, version 3). Berne 2013. Ce catalogue est disponible en ligne : www.kost-ceco.ch/wiki/whelp/Cfa (consultée le 22 juillet 2013).

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

3.	JP2 (JPEG 2000 Part 1) lossless	Note : 74,7.
4.	JP2 (JPEG 2000 Part 1) lossy	Note : 66,1.
5.	Basic JFIF (JPEG) 1.02	Note : 65,4.
6.	TIFF 6.0 LZW	Note : 65,3.

On constate que ces deux classements sont différents. Si l'établissement d'une méthode universelle, valable dans toute situation, était le but de ces deux études, alors cette constatation serait une contradiction. Mais les études de la Bibliothèque nationale des Pays-Bas [29] indiquent que les facteurs de pondération doivent être établis en fonction des situations particulières. Dans le même esprit, les recommandations du CECO précisent bien qu'il n'existe pas un unique format valable dans toute situation, et qu'il est nécessaire de tenir compte de l'application prévue pour faire un choix.

En effet, les deux méthodes décrites ci-dessus sont imparfaites. Ces méthodes visent à simplifier le plus possible l'évaluation et le choix d'un format en ramenant tous les critères sur une seule dimension. Ainsi, en additionnant des nombres correspondant à divers critères, on établit un moyen permettant de comparer des critères qui peuvent ne rien avoir en commun. Une telle simplification est discutable.

Une autre observation que l'on peut faire est que les processus définis par les méthodes brièvement présentées ci-dessus impliquent de noter les formats selon divers critères. Mais la façon dont il s'agit d'attribuer les notes est subjective. Elle dépend des personnes chargées de cette tâche. Or, les compétences et l'expérience de ces gens influent certainement sur les notes distribuées.

En résumé, les méthodes d'évaluation de formats vues plus haut donnent un faux sentiment de rigueur. Elles permettent d'obtenir des chiffres et d'en déduire un ou plusieurs formats plus adaptés que les autres, alors qu'il n'est pas possible de définir ces chiffres de manière unique. Il est d'ailleurs utopique de vouloir créer une méthode absolument rigoureuse, puisque le choix d'un format dans le domaine de la conservation à long terme n'est pas seulement une question de compétence et de réflexion, mais aussi un pari sur l'avenir.

Le choix d'un format d'images pour un projet de numérisation : Méthode alternative (arbre de décision)[30]

Les limites des méthodes discutées plus haut incitent à réfléchir à une méthode mieux adaptée. Le point essentiel est de pouvoir distinguer entre différentes situations, et il est par conséquent nécessaire de renoncer à un format unique qui serait valable dans tous les cas.

29 Gillesse, Robert ; Rog, Judith ; Verheusen, Astrid : Alternative File Formats for Storing Master Imagesof Digitisation Projects. La Haye 2008.

30 L'arbre de décision est visuellement représenté à la fin de cet article. Il s'agit des figures 5 et 6.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

D'ailleurs, certaines institutions renoncent au format TIFF [31] pourtant généralement indiqué comme le format à privilégier, comme cela est illustré par les études citées ci-dessus. Bien qu'il soit possible d'analyser ces choix divergents comme des erreurs, il semble beaucoup plus raisonnable de voir dans ces exemples une preuve de la nécessité de reconnaître que selon les cas, il s'agit de choisir des formats différents.

Dans la suite de cet article, on propose une méthode d'évaluation en arbre, qui permet de distinguer différentes situations en plusieurs étapes. Pour faciliter le choix, le processus décrit ci-dessous limite les possibilités à quelques formats dont les qualités sont telles que les risques sont aussi faibles que possible.

L'utilisation et les limites de la méthode en arbre

— Cet article n'a pas pour objectif d'étudier tous les formats d'images existants, ce qui serait impossible, mais il propose un processus permettant d'effectuer un choix raisonnable en fonction du contexte. Les formats et les algorithmes de compression proposés ici peuvent être considérés comme aptes à l'archivage à long terme, comme cela est brièvement expliqué plus haut.

Toutefois, d'autres formats et d'autres algorithmes de compression peuvent également être envisagés en cas de nécessité. Ainsi, cet article prend le point de vue d'une institution engagée dans des travaux de numérisation de documents originaux, ce qui lui laisse un contrôle entier sur le choix du format. Mais on peut imaginer une situation différente : une institution patrimoniale peut tout-à-fait se voir proposer des images numériques produites hors de son contrôle. Dans une telle situation, il vaut mieux commencer par étudier l'aptitude à l'archivage du format proposé plutôt que de convertir les images dans le format précédemment choisi par l'institution pour ses propres travaux.

— Le processus décrit ci-dessous, et permettant d'aboutir à un choix de format d'images n'est pas décrit en détail et laisse une liberté considérable aux personnes engages dans le choix d'un format d'images. En effet, il est impossible de prévoir toutes les situations et ce ne sont donc que les grandes lignes qui sont décrites ci-dessous. Certaines étapes nécessitent un travail d'analyse et de réflexion important, qu'il n'est pas possible d'éviter.

— Tous les formats dont il est question dans cet article sont des formats laissant une marge de manœuvre plus ou moins grande. Parmi les éléments le plus souvent discutés se trouvent l'algorithme de compression, qui peut soit être

31 C'est le cas des AEG, comme nous l'avons vu précédemment. Mais c'est aussi le cas du service Bibliothèque et Archives de l'UIT, qui conserve des fichiers PDF et non des fichiers TIFF (Service Bibliothèque et Archives de l'UIT. About the Digitization Programme & History of ITU Portal. Genève 2012). De plus, diverses institutions patrimoniales utilisent le format JPEG 2000, comme cela a été mentionné précédemment.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

choisi (cas du format TIFF et du format PDF), soit être réglé à un niveau qui est à la convenance du producteur (cas de JPEG et de JPEG 2000), et la résolution de l'image, qui ne dépend pas du format. Mais il y a en réalité une grande quantité d'autres éléments qui doivent être déterminés.

Par exemple, les métadonnées, de tout type, sont des informations qu'il est possible d'intégrer directement dans le fichier d'une image et qu'il est nécessaire de sélectionner préalablement. Par la suite, il faut trouver le moyen de les inscrire dans le fichier au moment de la production des images (dans le cas contraire, seules les métadonnées automatiquement inscrites sont présentes dans les fichiers).

D'autre part, les données qui décrivent une image peuvent être organisées de différentes manières, selon les possibilités offertes par les formats et selon le choix du producteur. Ainsi, et à titre d'illustration, TIFF donne la possibilité d'organiser les images en tuiles (il s'agit de diviser une image en plusieurs petites images rectangulaires, de sorte à faciliter l'accès à une région de l'image) au lieu de l'organisation en lignes qui est en principe l'organisation par défaut. Pour sa part, JPEG 2000 permet de choisir entre différentes organisations qui facilitent l'une ou l'autre utilisation. Par exemple, il est possible d'organiser les données d'une image de sorte à faciliter la transmission de cette image à une résolution plus faible que celle de l'image originale. L'idée est de placer d'abord les données permettant de construire l'image à une faible résolution, puis celles permettant de construire l'image à une résolution moyenne, et finalement les données permettant d'obtenir l'image originale de haute qualité. De cette façon, un utilisateur du web pourrait visualiser une copie de faible résolution sans avoir à attendre le chargement de toutes les données de l'image originale.

Il est donc très important de ne pas se contenter de choisir un format et un éventuel algorithme de compression, mais de déterminer aussi les caractéristiques, avec tous les détails nécessaires, que l'on souhaite pour les images numériques qu'une institution veut produire. Ces choix doivent faire l'objet d'une spécification, telle celle de la Bibliothèque nationale des Pays-Bas pour les caractéristiques techniques,[32] et tel le travail effectué par les AEG et les Archives de la Ville de Genève pour les métadonnées.[33]

Première étape

La première étape, qui peut aussi être comprise comme une sorte d'étape préliminaire, consiste à regrouper toutes les informations nécessaires. Il s'agit de détermi-

32 Van der Knijff, Johan (Bibliothèque nationale des Pays-Bas) : JPEG 2000 compression specifications for KB digitization projects. La Haye 2011.

33 Voir le blog d'Anouk Dunant Gonzenbach : present-hieretdemain.tumblr.com (articles du 13 et du 28 mai 2013, consulté le 22 juillet 2013).

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

ner les éléments permettant de caractériser les documents originaux, les objectifs de la numérisation, les ressources disponibles, et éventuellement certaines conditions spéciales.

— Déterminer les caractéristiques des documents originaux est évidemment nécessaire pour définir l'aspect technique de la numérisation, et en particulier le format des images numérisées. Parmi les éléments importants se trouvent la couleur (noir-blanc, niveaux de gris, couleur) et le contraste des documents originaux, le type de document (textuel, iconographique, etc.), et la finesse des détails.
— La numérisation peut répondre à deux objectifs principaux. Il peut s'agir de diffuser un ou plusieurs documents de manière facilitée, ou il peut s'agir de prendre une mesure de conservation pour préserver au mieux un document nécessitant un traitement particulier. Et en mettant les choses au pire, un document numérique peut remplacer un document original dont l'existence est menacée à court terme. Dans ce dernier cas, il est important de réaliser que le remplacement se fait au prix d'une perte importante puisqu'une image numérique ne peut être qu'une approximation du document original.
— Les ressources à disposition jouent évidemment un rôle important aussi. Il est nécessaire d'en dresser un inventaire, quel que soit le type (ressources en finances, personnel, temps, infrastructure, organisation, etc.).
— Enfin, il peut arriver qu'il soit nécessaire de tenir compte de conditions spéciales, qui sont aussi susceptibles d'apparaître dans la suite du processus de choix.

Deuxième étape

Durant la deuxième étape, il s'agit de déterminer si les images numériques doivent être conservées pour le long terme ou non, en s'appuyant sur les informations mises en évidence lors de la première étape.

En raison du coût de la numérisation et en raison de l'intérêt des documents numérisés, les images sont souvent produites pour être conservées pour une durée sans limite dans le temps. Il est à noter qu'il y a alors deux cas différents. D'une part, on peut se trouver dans le cadre de l'archivage à long terme, qui nécessite un système d'archivage de haute qualité. Il va de soi que les documents numériques dont le rôle est de remplacer des documents originaux en dégradation rapide doivent bénéficier d'un tel système. D'autre part, il peut s'agir de documents qui ne sont pas destinées à l'archivage mais qu'il est tout de même nécessaire de conserver pour le long terme. C'est le cas des documents numériques destinés uniquement à la diffusion, parce que seuls les documents originaux sont archivés, ou parce que l'institution concernée archive parallèlement des copies numériques de haute qualité.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Mais il peut aussi arriver que les documents numérisés ne soient pas conservés au-delà d'une date bien définie et proche dans le temps. Par exemple, lorsqu'une institution offre un service de numérisation au public, il peut être jugé que les images créées dans ce cadre ne présentent pas un intérêt suffisant pour nécessiter une conservation à long terme. Alternativement, le coût de la conservation du document original et du document numérique peut être considéré comme trop important pour permettre la conservation du document numérique, sachant que le document original peut être numérisé une deuxième fois si nécessaire.

Troisième étape

Il est nécessaire de déterminer si les images seront en noir/blanc ou non. En effet, certains formats et certains algorithmes de compression sont spécifiquement conçus pour des images en noir/blanc, respectivement pour des images en couleur. Ainsi, la méthode de compression Groupe 4 est spécifiquement conçue pour les images en noir/blanc, alors que les algorithmes de compression JPEG et JPEG 2000 peuvent être utilisés uniquement pour les images en couleur.

Quatrième étape : les images ne sont pas conservées pour le long terme

S'il n'y a pas d'exigence particulière, un bon choix est de produire des images TIFF avec compression Groupe 4 (images noir-blanc à caractère textuel), des images TIFF avec compression LZW (images noir-blanc sans caractère textuel) ou des images JPEG avec niveau de compression d'environ 8/12 sur l'échelle de Photoshop[34] (images en couleur).

Au contraire, s'il y a des besoins spéciaux, alors le format le mieux adapté doit être choisi, quel qu'il soit. En effet, puisque les images ne doivent être conservées que pour une courte période, il n'y a pas besoin de prendre en compte la question de la durabilité de ces images. Il suffit de produire des images répondant parfaitement au besoin du moment.

Quatrième étape : les images, uniquement destinées à la diffusion, sont conservées pour le long terme

Lorsque les images ont la diffusion pour seuls buts, alors les formats et les méthodes de compression conseillés sont les mêmes qu'au point ci-dessus, à savoir des images TIFF avec compression Groupe 4 (images noir/blanc à caractère textuel), des images

34 Claerr, Thierry ; Westeel, Isabelle : Manuel de la numérisation. Paris 2011.
La compression JPEG ne définissant pas de niveaux de qualité officiels, le présent article prend les niveaux de qualité de Photoshop comme référence, comme cela est fait dans ce Manuel de la numérisation. Mais il va de soi que chacun peut définir le niveau désiré selon une autre échelle.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

TIFF avec compression LZW (images noir/blanc sans caractère textuel) ou des images JPEG avec niveau de compression d'environ 8/12 sur l'échelle de Photoshop (images en couleur). Mais comme il y a un objectif de conservation à long terme, il est déconseillé d'utiliser un autre format ou une méthode de compression différente sans une évaluation approfondie.
Essentiellement, il y a deux cas possibles. Il peut s'agir d'images dérivées à partir d'images archivées de haute qualité. Mais il peut aussi s'agir d'images pour lesquelles il n'existe pas de copies d'archivage. En effet, une institution peut décider de créer des images exclusivement destinées à la diffusion, en s'appuyant sur le fait que les documents à archiver sont les documents «papier» originaux. Comme exemple d'un tel cas, nous avons déjà vu l'exemple des AEG.

Quatrième étape : les images sont destinées à l'archivage
Lorsque l'on choisit de créer des images avec un objectif d'archivage à long terme, il faut déterminer si l'institution concernée veut, ou non, utiliser un algorithme de compression. De manière générale, la compression est vue comme un élément à éviter autant que possible dans le domaine de l'archivage à long terme. Pour cette raison, le standard de fait est le format TIFF sans compression. Toutefois, devant la masse des images produites actuellement et dans le futur, de nombreuses institutions se tournent vers l'utilisation de la compression.

Si une institution renonce à utiliser le format TIFF sans compression, alors il est recommandé d'utiliser TIFF avec compression Groupe 4 (images noir/blanc à caractère textuel), TIFF avec compression LZW (images noir/blanc sans caractère textuel) et JPEG 2000 sans perte de qualité (images en couleur).

Le choix d'un format d'images pour un projet de numérisation : commentaires concernant l'arbre de décision

Conditions spéciales
Il peut arriver qu'une institution soit soumise à des conditions spéciales. Ainsi, les Archives nationales d'Australie n'acceptent pas le format TIFF comme format d'archivage, en raison de son développement par une entreprise unique qui en a gardé la propriété (Aldus, aujourd'hui reprise par Adobe). Les Archives nationales d'Australie ont par conséquent choisi le format PNG comme format d'archivage.[35]

35 Cunliffe, Allan (Archives nationales d'Australie) : Dissecting the Digital Preservation Software Platform, 2011.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Concernant les formats

Comme le montre l'exemple des Archives nationales d'Australie, il est possible de jeter son dévolu sur un format différent de ceux exposés dans l'arbre de décision discuté plus haut.

En plus du format PNG, rappelons l'existence de JBIG2 (définit par le comité JPEG), un algorithme de compression plus efficace que l'algorithme de compression Groupe 4 pour les images noir/blanc, mais dont l'utilisation semble rare dans le domaine patrimonial. Dans le cas où une institution aurait la responsabilité d'une quantité d'images noir/blanc telle que cela poserait des problèmes, alors il est recommandé d'étudier l'utilisation de JBIG2.

En ce qui concerne les images couleur, notons la possibilité d'utiliser le format JPEG 2000, compression avec perte, en remplacement du format JPEG. Toutefois, en raison de l'acceptation moindre du format JPEG 2000, il vaut mieux renoncer à l'utiliser pour la diffusion des images actuellement. A moins, évidemment, de disposer ou de construire un environnement spécifiquement construit à son intention, à l'image de ce qui a été fait par les Archives nationales du Japon.[36]

Et il est possible d'utiliser le format TIFF avec la compression LZW au lieu d'utiliser le format JPEG 2000 pour l'archivage à long terme des images couleur. La compression LZW est un algorithme moins complexe (mais aussi moins efficace) que l'algorithme JPEG 2000, ce qui peut être considéré comme un avantage dans une perspective à long terme.

Concernant le format PDF

Le format PDF dispose d'atouts qui peuvent le rendre incontournable. Il permet de regrouper d'une manière agréable un ensemble d'images dans le même fichier. Typiquement dans le cas des livres, cette qualité est appréciable : un lecteur préfère certainement recevoir un unique fichier plutôt que de devoir gérer un fichier par page.

Une autre qualité du format PDF est qu'il permet de « cacher du texte derrière une image ». Lorsqu'un document textuel est numérisé, il est souvent très utile de pouvoir disposer du texte. Pour cela, il faut préalablement saisir le texte manuellement ou utiliser un logiciel d'OCR (Optical Character Recognition). Ensuite, il est possible de sauvegarder séparément le texte et l'image, mais il est aussi possible de sauvegarder le texte « derrière l'image ». En procédant de cette deuxième manière, on garde le lien entre chaque caractère textuel et le lieu où il se trouve dans l'image, ce qui est assurément précieux dans certaines situations. Or, le format PDF permet cette opération.

36 www.digital.archives.go.jp/index_e.html (consultée le 22 juillet 2013).

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

Pour ces qualités, et en raison de l'existence de versions du standard PDF qui sont précisément définies pour l'archivage à long terme, ce format peut également être utilisé. En particulier, cela peut être nécessaire pour la diffusion des images. En revanche, dans le cadre de l'archivage à long terme, les qualités présentées ci-dessus semblent moins décisives. Dans ce dernier cas, il est nécessaire de procéder à un examen détaillé de la situation pour déterminer si le format PDF/A est une bonne solution.

Contraintes pratiques

Il est crucial de ne pas perdre de vue les contraintes pratiques tout au long des étapes décrites ci-dessus. On peut illustrer cette nécessité avec le cas du format JPEG 2000. Les scanners de livres et les appareils photos ne peuvent pas toujours créer des images qui respectent ce dernier format.[37] Il faut alors commencer par créer les images dans un autre format (sans perte), puis convertir les images en JPEG 2000. Or, une conversion est un traitement délicat qui ne peut pas être fait sans précaution. En effet, certains logiciels suppriment des métadonnées dans le cours de l'opération, ce qu'il faut éviter au moins lorsque l'on se trouve dans le cadre de l'archivage à long terme. De plus, des accidents de toute sorte sont possibles, et il est donc raisonnable de mettre en place un contrôle des images obtenues par conversion avant de supprimer les fichiers originaux.[38]

Contradictions possibles

Il est possible d'aboutir à une contradiction dans le cours du choix d'un format selon la méthode en arbre. Ainsi, on peut imaginer qu'une institution, qui souhaite archiver des images qui sont codées sans compression, n'ait pas les ressources suffisantes à la création et à la gestion d'un système d'archivage permettant d'intégrer toutes les données liées à ces images.

Dans une telle situation, la contradiction doit être levée en modifiant le contexte de départ (les objectifs, les ressources, les conditions spéciales), ou en évaluant une deuxième fois les conclusions auxquelles les réflexions ont abouties.

Cela signifie qu'à chaque étape il est possible de revenir à la première étape pour débuter une nouvelle fois le processus de choix avec un contexte modifié.

37 A titre d'illustration, la BGE dispose actuellement de trois scanners de livres de dernière génération, et seul le format TIFF (sans compression) est pris en charge par les trois appareils. Ce qui montre aussi que les difficultés dépassent le seul cas du format JPEG 2000.

38 Van der Knijff, Johan et al.: Improved validation and feature extraction for JPEG 2000 Part 1: the jpylyzer tool. Copenhague 2012.

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

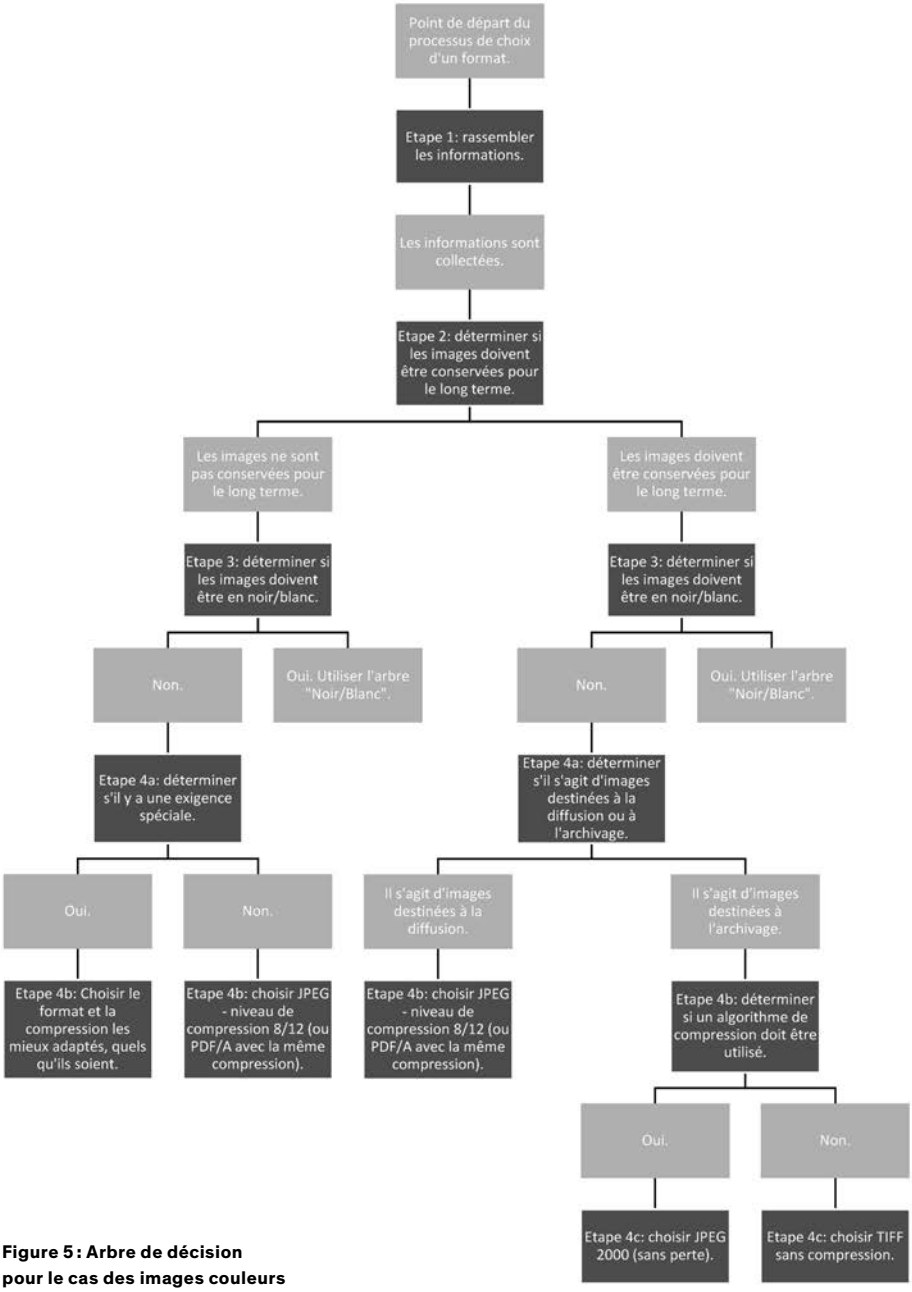

Figure 5 : Arbre de décision pour le cas des images couleurs

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7

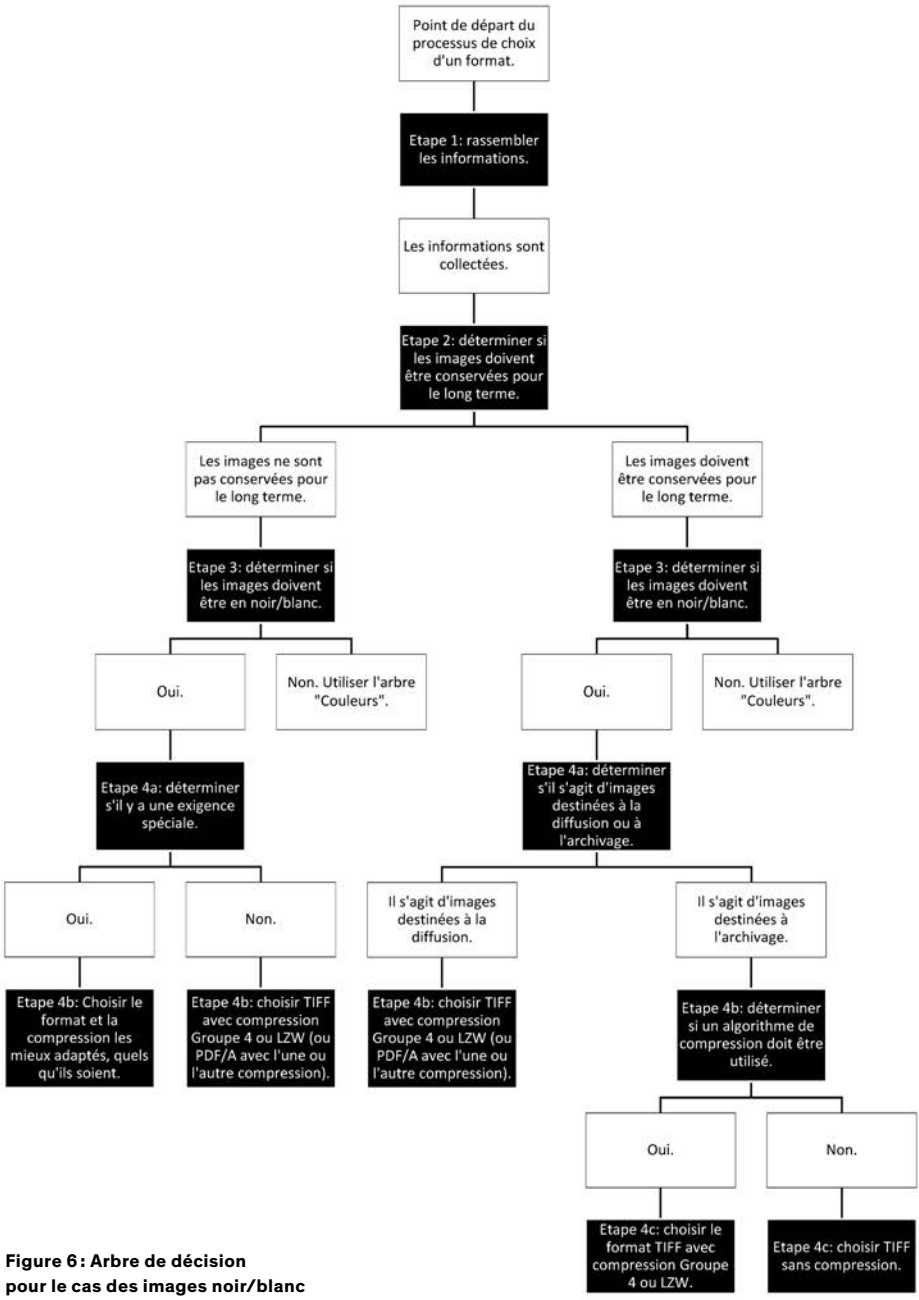

Figure 6 : Arbre de décision pour le cas des images noir/blanc

Informationswissenschaft: Theorie, Methode und Praxis, 2014 - http://dx.doi.org/10.18755/iw.2014.7