Quelques remarques sur les référentiels de numérisation de la BnF *

Denis Roegel †

7 janvier 2014

Table des matières

*Dans ce document, nous avons simplifié les références vers le site de Gallica. Les contenus des encadrés marqués « **G**: » sont les dernières parties de liens commençant par « `http://gallica.bnf.fr/ark:/12148/` ».

†maître de conférences, Université de Lorraine & LORIA, `roegel@loria.fr`.

1 Introduction

Début décembre 2013, Laurent Duplouy, le responsable de la numérisation à la BnF, a rendu public les référentiels de numérisation de la BnF, afin qu'ils puissent notamment servir aux prestataires de numérisation.

Ces documents sont particulièrement intéressants et il s'agit peut-être de la première fois qu'une grande bibliothèque rend publics ses référentiels. C'est donc un point très positif, et cette publication devrait permettre une analyse critique, des échanges et des améliorations qui pourront bénéficier à la BnF, mais aussi à d'autres acteurs du domaine.

2 Les référentiels

Les référentiels de numérisation sont disponibles à l'adresse suivante :

`http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.numerisation_referentiels_bnf.html`

Il y a actuellement huit référentiels :

- Numérisation des documents opaques
- Numérisation des documents transparents
- Format de fichier image
- Enrichissement des métadonnées
- OCR
- Traitement des tables
- ePub 3
- Livraison de document numérique

Ces référentiels couvrent toute la chaîne de numérisation de la BnF, depuis la numérisation elle-même, en passant par les choix techniques (formats d'image), la transcription par OCR, l'ajout de métadonnées et de tables (tables des matières et index), la production du format ePub pour liseuses et tablettes, et les modalités techniques de livraison des numérisations.

On ne peut que féliciter la BnF pour cet effort d'ouverture et encourager d'autres bibliothèques à faire de même. Il faut cependant noter que la BnF pourrait poursuivre ses efforts et être plus ouverte sur l'autre versant de la numérisation, à savoir l'utilisation des données numériques, c'est-à-dire toutes les questions liées aux interfaces, où beaucoup reste encore à faire et à améliorer.

3 La compréhension de la numérisation et les choix

3.1 Le choix de la résolution

Outre les remarques détaillées qui figurent plus loin, j'aimerais évoquer ici ce qui reste à mes yeux l'écueil fondamental dans la chaîne précédente. Comme on peut le voir dans une partie des documents du référentiel, la BnF prend souvent le point de vue de la présentation des documents, c'est-à-dire de l'interface, lorsqu'elle devrait se concentrer sur la question de la conservation.[1] Ainsi, on lit par exemple que « Les documents de très grands formats pourront être numérisés en 150 dpi ». Pour des supports transparents, il faudrait, selon le référentiel, numériser à une résolution moindre des transparents $4'' \times 5''$ que des diapositives 24×36. Le raisonnement sous-jacent imagine que la vue sera examinée globalement et que c'est la perspective globale qui doit déterminer la résolution. Or, s'il est vrai que la perspective globale détermine la résolution de *présentation* (à taille de présentation fixe), cela n'implique absolument rien pour la résolution de numérisation. Avec ce raisonnement, on arrive très facilement à des numérisations illisibles, ou très limites.

Il est facile de comprendre que le raisonnement précédent n'est pas correct, car certains documents ne sont pas destinés à être vus à la fois globalement et dans le détail. Une grande carte Michelin, par exemple, ne peut être vue dans son ensemble qu'à 1.50 m ou 2 m de distance, mais à cette distance on ne pourra pas lire les détails. Une telle carte, et d'autres documents qui seraient encore plus grands, ne sont pas des posters, mais de grands documents destinés à être examinés de près, donc partiellement. La taille du document ne doit donc jouer aucun rôle sur la résolution de numérisation. Ce qui importe, c'est le contenu.

Plus précisément, deux aspects jouent ici un rôle. Le premier, c'est qu'un document original comme un texte imprimé ou une carte sont destinés à être regardés dans certaines conditions, notamment à une certaine distance. Cette distance, jointe aux capacités optiques de l'œil, détermine une résolution au-delà de laquelle il n'est pas nécessaire d'aller. En pratique il est difficile de distinguer des détails beaucoup plus petits qu'un 300^{e} de pouce à une cinquantaine de centimètres et si par sécurité on double cette valeur, une résolution de 600 dpi suffira à produire un facsimilé offrant le même confort de lecture. On pourrait donc ici parler des

1. Le concept de « numérisation durable », indissociable de celui de la conservation, a été analysé de manière très détaillée sur `http://locomat.loria.fr`.

besoins du lecteur.

Le second aspect, c'est celui du contenu informationnel du document. Un document qui reproduit une scène n'est pas un original (malgré ce que laisse entendre le référentiel des documents transparents), mais est un document qui a une relation avec un autre document, un paysage, ou des personnes. Cette relation est obtenue par une chaîne qui influe sur la qualité du résultat. L'optique de l'appareil de saisie, les mouvements de l'appareil ou d'éléments de la scène (y compris des lumières) perturbent l'acquisition et dégradent celle-ci, introduisant des facteurs limitants. D'autre part, le support d'acquisition lui-même a des limites, par exemple liées à la taille des grains, ou d'autres facteurs. Ces considérations s'appliquent aussi à la numérisation elle-même. En fin de compte, dans la reproduction on a une limite dans la densité des détails. Par conséquent, dans un tirage photographique, ou une diapositive, il existe une résolution maximale au-delà de laquelle on ne pourra plus tirer de détails provenant de la scène. On pourra tirer de nouveaux détails du support, mais ceci n'est en général pas ce qui est recherché. On pourrait ici parler des *besoins du document*.

Dans tous les cas, que le support soit un imprimé ou une reproduction photographique (opaque ou transparente), ce support est ou bien visualisé dans un contexte qui permet d'en déduire une résolution, ou bien doit être vu comme une reproduction, auquel cas il est souhaitable d'extraire de ce support tout ce qui a trait à la scène originale. Dans les deux cas, ce n'est pas la taille du support qui doit intervenir, mais sa nature, éventuellement le processus par lequel le support a été obtenu (certains transparents sont des transparents contacts, mais d'autres non, et ceci a un impact sur la définition), et, souvent, le mode de visualisation.

Cette distinction entre les besoins du document et les besoins du lecteur ne semble pas comprise par le service de numérisation de la BnF. Il faut noter que je l'avais déjà évoquée dans mon étude sur la numérisation durable.

3.2 Le choix du format image

Un autre écueil, moins grave, est celui qui consiste pour certaines bibliothèques, dont la BnF, à vouloir à la fois utiliser les technologies les plus récentes, et à croire que les technologies un peu plus anciennes sont caduques. Ainsi, il y a vingt ans, la BnF a fait le choix du format TIFF, qui n'était déjà pas particulièrement justifié. Je suppose que ce format a été choisi car il permet un stockage « sans pertes ». On pouvait alors numériser et on était sûr que ce que l'on stockait était bien le résultat de la

numérisation et non une version dégradée de cette numérisation. Même si ce raisonnement est correct, il oublie tout ce qui est en amont de la chaîne, et notamment les limitations optiques ou autres qui peuvent avoir un impact plus important que les pertes dues à un changement de format.

Les référentiels de numérisation n'évoquant à aucun moment les pertes en amont de la numérisation (sans parler des phénomènes de bruit dans les capteurs), je suis enclin à penser que le choix du format JPEG2000 s'est fait uniquement sur l'examen de la dernière étape de la chaîne, c'est-à-dire sur le passage de l'image acquise par le capteur au fichier stocké. Mais il est tout-à-fait possible d'avoir de très mauvais fichiers JPEG2000 et ce format ne garantit absolument rien. Il présente certes des avantages sur le JPEG, notamment l'affichage incrémental et une compression un peu plus grande à qualité égale, mais ces avantages ne sont pas exceptionnels, et ils ne peuvent à eux seuls justifier un changement de format. Gagner 20% de place mémoire n'est aujourd'hui pas une raison suffisante pour passer de JPEG à JPEG2000. Cela dit, choisir l'un ou l'autre est quasiment inconséquent. Par contre, le TIFF non compressé est évidemment bien plus gourmand en place. J'ai moi-même utilisé du TIFF et je l'ai remplacé par du JPEG il y a plus de dix ans, sans pertes visibles, y compris au niveau des détails. Je n'ai pas l'intention de passer au JPEG2000 et je vois très mal ce qui le justifierait aujourd'hui. Et quand bien même cela deviendrait un jour préférable, je pourrais toujours convertir mes images JPEG en JPEG2000.

4 Remarques détaillées

On trouvera ci-après un certain nombre de remarques plus détaillées sur les référentiels de numérisation de la BnF. Lorsque certains documents ont donné lieu à peu de commentaires, cela veut dire soit que je n'avais pas de critiques à faire, soit que je n'avais pas une grande pratique des formats concernés.

4.1 Référentiel de livraison de document numérique [2]

Je n'ai ici pas de remarques particulières, si ce n'est que le document fait référence au format JPEG2000 (`.jp2`) dont l'emploi n'est pas justifié.

2. Version du 29 octobre 2013.

4.2 Référentiel de format de fichier image [3]

[p. 3] Le document rappelle qu'en 1993 la BnF avait choisi d'utiliser le TIFF, ce qui était déjà un choix contestable. À partir de 1996, le TIFF a été utilisé en version monopage.

On apprend que la BnF a envisagé le passage à JPEG2000 dès 2009, mais hésitait, vu la complexité du format. Je vois mal en quoi la « complexité » du format joue un rôle, puisque cette complexité est invisible. D'autre part, on sent ici une focalisation sur la difficulté à utiliser le format (ce qui est accessoire), et un mutisme total sur les motivations pour utiliser ce format. Or, il n'y a pas de bonnes raisons d'utiliser ce format, en tous cas pas de meilleures que le JPEG. Certes, ce n'est pas grave d'utiliser le JPEG2000 en format interne, il faut simplement que cela n'ait pas de conséquences dommageables à la qualité des documents visibles par les utilisateurs. Or, justement, sur certains sites, le JPEG2000 semble très mal employé (voir mon panorama de la numérisation durable sur `http://locomat.loria.fr`).

Il y a aussi une certaine confusion quant à la distinction entre le format de stockage et le format de présentation, qui n'ont pas besoin d'être les mêmes. Le fait que peu de personnes soient équipées pour le JPEG2000 ne doit pas être un frein à son emploi pour le stockage, puisqu'un autre format peut être utilisé pour la présentation, éventuellement dans une résolution moindre. Je crois que les responsables de la numérisation à la BnF devraient clarifier ces aspects.

Les avantages donnés pour employer le JPEG2000 me semblent minimes :

1. *capacité de compression sans perte visuelle importante* : à titre d'exemple, j'ai pris **G**:`btv1b52500992c/f83.highres`, un fichier JPEG couleur de 1024×1563 pixels et dont la taille est 284209 octets ; d'après la description de l'ouvrage sur Gallica, les dimensions des pages seraient de $152\,\mathrm{mm} \times 100\,\mathrm{mm}$; en convertissant ce fichier en JPEG2000 avec *ImageMagick*, sa taille croit chez moi à 300725 octets, mais c'est à qualité égale ; en passant à une qualité de 50%, j'obtiens certes un JPEG2000 de taille 114443 octets dont la dégradation n'est pas sensible ; du moins je ne vois pas de différence nette, même en zoomant ; cela dit,

3. Version du 28 octobre 2013.

en prenant le JPEG d'origine et en le dégradant à une qualité de 40%, j'obtiens un JPEG de 98725 octets, qui apparaît légèrement dégradé, mais sans que ce soit véritablement gênant ; ce qui est en fait le plus gênant dans cet exemple, c'est que le document n'a été numérisé qu'à environ 250 dpi, ce qui est très insuffisant ici ; et les dégradations observées sont quasiment invisibles dans une lecture normale de ce document, ce que j'invite chacun à vérifier ; en d'autres termes, si le problème est la taille de l'image, on peut obtenir des résultats plus que satisfaisants sans avoir recours au JPEG2000 ;

2. *capacité d'extraire un niveau de résolution sans générer des images intermédiaires* : est-ce si important ? dans la mesure où la BnF ne voudra sans doute pas toujours donner accès aux images originales, ne vaudra-t-il pas mieux séparer le format de stockage et le format de présentation ? d'autre part, une partie de ce discours suppose implicitement une interface d'affichage avec des éléments partiels zoomés ou des affichages mosaïques, alors que ces affichages font justement partie des tares de certaines interfaces actuelles ; plutôt que de choisir un format qui ne se justifie pas tant que cela, les bibliothèques comme la BnF devraient plutôt améliorer leurs interfaces qui sont encore ultra-primitives, non conviviales, peu efficaces, etc. ;
3. *forte expressivité des métadonnées* : rien n'oblige à mettre les métadonnées dans le fichier image, on peut les stocker dans un autre fichier ; cela dit, la spécification XMP part 3 indique de manière détaillée comment le XMP peut être stocké dans les JPEG, JPEG2000 et de nombreux autres formats, si bien que l'argument d'expressivité invoqué ici me semble faible.

Par contre, l'affichage incrémental des images n'est pas mis en avant. Même si le JPEG2000 a certainement des avantages sur le JPEG ou d'autres formats, le fait qu'il soit possible de convertir les anciens formats en JPEG2000 ne semble pas rendre ce nouveau format indispensable. On a l'impression que la BnF souhaite être à la pointe de la technologie, mais en réalité elle ne porte l'effort que sur un aspect de la chaîne, en négligeant d'autres points essentiels. C'est un peu comme si dans la chaîne du froid on ne portait ses efforts que sur les véhicules d'acheminement, et pas sur les autres étapes de la chaîne.

[p. 4] Je trouve dommage que la presse soit numérisée en qualité inférieure à celle des livres, on stocke trois fois moins d'information par pixel pour la presse que pour les livres, alors qu'à résolution égale,

vu la taille plus petite des polices de la presse, il faudrait au contraire inverser ce rapport, ce n'est pas très logique ; cela l'est aussi peu du fait de la présence de gravures et autres illustrations dans la presse, mais le référentiel de numérisation des documents opaques indique p. 9 que celles-ci seront numérisées en couleur, si ces illustrations sont en couleur (i.e., autres qu'en noir et blanc) ; je pense que cela ne suffit pas.

4.3 Référentiel de numérisation des documents opaques [4]

[p. 4] La numérisation est présentée comme technique de conservation et de diffusion, donc implicitement permet de remplacer l'original d'un document ; pour que ceci soit effectif, il faut que ceux qui numérisent comprennent la notion de « quantité d'informations » dans un document, car cela détermine comment on doit le numériser ; or, cette notion n'est aujourd'hui pas encore assimilée par les bibliothèques qui raisonnent trop en termes d'interfaces.

[p. 5] « Les règles … n'évolueront qu'à la marge en fonction des évolutions techniques du domaine ». Ceci montre que la BnF n'a pas encore compris que les règles ne dépendent pas que de l'évolution technique du domaine, mais aussi de la compréhension que la BnF aura des besoins des documents et des besoins des utilisateurs, points qui sont indépendants de l'« évolution technique du domaine ».

[p. 7] Le document insiste sur le fait que la numérisation de « conservation » reproduit au plus près le document original. Ceci ne pourra se faire que lorsque la BnF aura réfléchi aux besoins des documents en liaison avec leur contenu informationnel.

La numérisation « de diffusion » est présentée avec un objectif sensé être flatteur, ce qui est une interprétation pour le moins curieuse. S'il doit y avoir une différence entre la conservation et la diffusion, c'est dans les questions de rapidité d'accès et de présentation, aspects qui rendront de moins en moins nécessaires une différence entre deux formats, du fait de l'accroissement des débits. Le point essentiel de la diffusion ne réside pas dans les formats d'image, par exemple, mais dans l'ergonomie des interfaces.

Nous apprenons ici qu'actuellement la BnF met en consultation la version de conservation. Elle envisage éventuellement une distinction dans le futur.

4. Version du 8 novembre 2013.

[p. 10] Concernant les résolutions, la BnF semble considérer que les petits originaux nécessitent une résolution plus grande que les grands. Elle écrit ainsi « Les documents de très grands formats pourront être numérisés en 150 dpi », ou encore « Tout document dont le format est inférieur à A6 est numérisé à 600 dpi », sous entendant que ce qui est plus grand sera en général numérisé à 400 dpi. On comprend assez facilement les motivations de la BnF, à savoir que les petits documents peuvent faire l'objet d'agrandissements plus grands que les grands documents. Mais la BnF passe une fois de plus totalement à côté de la question des besoins du document (qui est centrale à la conservation et à la reproduction aussi proche que possible de l'original) pour penser en terme de présentation. Cette façon de voir les choses, si elle se maintient, conduit inexorablement à la perpétuation de la mauvaise qualité numérique. Un plan A0 n'a aucune raison d'être numérisé en une résolution moindre qu'un fragment de papier A7, par exemple. Il est vrai que si on présente ce document globalement à l'écran, il ne sera pas nécessaire de numériser le plan en A0. Mais alors, on peut se demander si ce plan avait besoin d'être en A0 à l'origine. Ce que la BnF devrait plutôt distinguer, ce sont les grands originaux destinés à être vus globalement ou de loin (comme les affiches publicitaires), et les grands originaux destinés à être vus de près. Les tailles sont les mêmes, mais les besoins ne le sont pas. Lorsque la BnF comprendra cela, la numérisation s'en trouvera améliorée. (En passant, il y a des endroits dans le référentiel où on a des amorces de compréhension, par exemple p. 6 du référentiel de numérisation des documents opaques, où une distinction est faite entre deux types de supports transparents.)

La BnF propose de numériser les monnaies à 1200 dpi, je pense que la plupart du temps ce n'est pas nécessaire. À mon avis, 600 dpi suffisent, en tous cas c'est mon expérience avec les monnaies. Cela dit, mieux vaut rester à 1200 dpi que de descendre à 300 dpi.

La numérisation des estampes à 800 dpi devrait aussi être analysée. Ce qui importe une fois de plus, c'est la « quantité d'information » de l'original, et notamment ce que l'on veut numériser. Veut-on numériser seulement l'intention de l'auteur, ou les irrégularités dues au support, voire le support lui-même, par exemple ses fibres ? Le niveau de numérisation dépend aussi du type de dessin, s'il contient des traits les besoins ne sont pas les mêmes que s'il contient des taches.

[p. 11] Je m'étonne que la BnF accepte qu'une seule charte colorimé-

trique soit utilisée lorsque la numérisation d'un document s'étend sur plusieurs jours. Il faudrait de manière évidente utiliser au minimum autant de chartes que de jours.

[p. 15] La numérisation des monnaies, si elle se fait par rotation, entraîne une perte d'information (disposition d'une face par rapport à l'autre), à laquelle il faudrait remédier d'une manière ou d'une autre.

Le problème de l'alignement de page est un problème difficile. L'alignement sur le bord inférieur n'est pas forcément adéquat. Il se peut que le bord inférieur ne soit pas bien aligné avec l'illustration du contenu. Je pense que dans ce cas, il faudrait produire deux numérisations, une alignée sur la page, une autre sur le contenu, voire éventuellement plus. Il faut en tous cas éviter d'avoir à tourner les illustrations par la suite, comme je l'ai montré dans un article de `http://locomat.loria.fr`. (voir aussi **G**:`bpt6k6353531h` qui a ou bien été numérisé sans aligner, ou bien subi une rotation)

[p. 17] Les documents de grand format sont le plus souvent des illustrations ou des figures techniques, et devraient être systématiquement numérisées en 600 dpi. La BnF souhaite imposer une limite à 1.8 Go dans le poids des fichiers, mais elle ne dit pas ce qui a déterminé ce choix.

[p. 21] Dans **G**:`btv1b8600057b` je n'ai pas trouvé la double page de la règle 8.

[p. 49] La règle d'alignement unique sur le bord du cache ne suffit pas. Il faut deux numérisations, une par rapport au cache (éventuellement de résolution moindre) et une par rapport à la carte.

[p. 51] L'assemblage de vues est évoqué, mais pas les problèmes d'ajustements à l'assemblage. Ces problèmes sont fréquents et liés à la non-planéité ou aux déformations des originaux.

[p. 53] La règle d'alignement est aussi à revoir ici et deux numérisations peuvent s'avérer nécessaires.

4.4 Référentiel de numérisation des documents transparents [5]

[p. 10] Les résolutions suggérées (300 et 400 dpi) sont totalement aberrantes ; en effet, si un négatif 24×36 reproduit une feuille A4 que l'on numériserait à 600 dpi, alors, sachant que le facteur de réduction de A4 vers 24×36 est de l'ordre de 8, il faudrait en fait numériser le

5. Version du 8 novembre 2013.

support transparent non pas à 600 dpi, mais à huit fois plus, soit à environ 5000 dpi ; ceci suppose évidemment que le support transparent puisse stocker les détails de l'original, ce qui n'est pas le cas ; on estime habituellement que la résolution maximale d'un négatif 24×36 est de l'ordre de 3000 dpi ; le négatif contient moins d'informations que l'original, mais numériser à 300 ou 400 dpi serait bien en-deça des capacités du support transparent ; il se peut que l'information passée de l'original au transparent sous-utilise les capacités du transparent, par exemple en cas de bougé, ou d'une mauvaise optique, mais le plus simple est de toujours se fixer comme objectif de numériser les transparents reproduisant des originaux (ce que la BnF appelle les supports de substitution) à 3000 dpi environ.

[p. 11] Les raisonnements de la BnF s'appuient sur une notion de rapport de réduction, alors que la numérisation doit s'intéresser aux contenu informationnel des supports. Le fait que l'image représentée est réduite d'un facteur 3, 5 ou 10 est presque indifférent. Même pour un facteur 3 depuis un original A4, il serait bon de numériser en 1800 dpi, donc bien au-delà de ce que semble aujourd'hui suggérer la BnF.

[p. 12] La justification d'une résolution de 2500 dpi est totalement fantaisiste. Il n'y a pas de rapport entre la résolution d'un original (cette fois-ci) et d'un rapport d'agrandissement, quel qu'il soit. Ce que la BnF, et d'autres bibliothèques, doivent comprendre, c'est que chaque support a une capacité de stockage d'informations, et que cette capacité est en général limitée par d'autres facteurs, notamment l'optique et les mouvements. Pour un film argentique, la taille des grains détermine la résolution maximale. Pour d'autres types de supports, il y a d'autres critères. Il n'y a aucune raison d'utiliser la même résolution pour un autochrome, une diapositive 100 ISO, une diapositive 25 ISO, etc. Pour un négatif 24×36, la résolution optimale se situe aux alentours de 3000 dpi, et elle est très rarement atteinte. Il est certes fort probable que 2500 dpi suffisent la plupart du temps, mais mieux vaut faire des tests pour voir si les originaux ne contiendraient pas plus de détails.

On peut s'étonner du fait que la BnF propose des résolutions différentes pour des supports identiques :

1. 300 ou 400 dpi si le support représente une copie ;
2. 2500 dpi si le support représente un original.

C'est non seulement incohérent, mais en plus c'est oublier que les

diapositives, autochromes, etc., ne sont pas vraiment des originaux, mais des reproductions d'objets. Dans un cas l'objet est un livre, dans l'autre c'est un paysage, une scène, etc. La BnF ne devrait pas faire les distinctions qu'elle fait.

Sur cette même page, l'exception présentée en 4.2.2 est aberrante. Les supports présentés (diapositive, $4'' \times 5''$, etc.) n'étant pas des agrandissements, rien ne justifie de les numériser à des résolutions différentes. Ils doivent tous être numérisés à 2500 ou 3000 dpi. Une fois de plus, c'est la focalisation sur la présentation qui conduit la BnF à faire les mauvais choix.

[p. 26] Je suis allé voir **G**:`btv1b9064681q` et la première image s'avère être de taille 1536×1050, ce qui, en supposant l'original en 24×36, donne une résolution d'environ 1100 dpi, ce qui n'est certainement pas suffisant et semble contredire la résolution annoncée page 12. D'un autre côté, il est possible que les diapositives originales soient suffisamment mauvaises pour que 1100 dpi suffisent.

4.5 Référentiel OCR [6]

Ce référentiel concerne essentiellement la production d'ALTO, un format simple que je découvre ici, et qui n'appelle pas de remarques particulières, sinon que certaines idées de ce format pourraient être intégrées dans les structures de tables, comme j'y reviendrai plus loin.

Sinon, je profite de ce référentiel pour signaler que certains passages en majuscules gagneraient à être accentués, par exemple p. 37, mais aussi ailleurs, et dans les autres référentiels.

4.6 Référentiel d'enrichissement des métadonnées [7]

[p. 3] « A l'inverse des microformes, [la numérisation] n'est pas autonome et nécessite des métadonnées ». Au contraire, je ne pense pas que l'on puisse dire que la microforme est autonome. Contient-elle tout ce que l'on aimerait qu'elle contienne ? Et si oui, alors la numérisation pure et simple d'une microforme est nécessairement autonome, ce qui semble être une contradiction. On n'aurait alors pas besoin de lui adjoindre des métadonnées. Il faut donc faire d'une part la part entre les informations présentes dans le document (microforme ou autre) que l'on numérise, et celles qui ne s'y trouvent

6. Version du 7 novembre 2013.
7. Version du 7 novembre 2013.

pas, d'autre part distinguer la forme de ces informations, qui peut être plus ou moins pratique à utiliser.

[p. 9] L'exemple du journal « La lanterne » est quasi illisible, même en soit-disante haute-résolution : **G**:`bpt6k75202819/f1.highres`

4.7 Référentiel traitement des tables [8]

[p. 8] On apprend ici, de manière un peu surprenante, que la BnF souhaite limiter la production d'index et de tables des matières complexes. La motivation de cette affirmation n'est pas très claire, mais peut-être que la BnF souhaite ne produire ces éléments que s'ils apportent une plus-value nette. Avec une table des matières sous forme image, on pourra toujours se débrouiller. Cela dit, personnellement, je me sers assez peu des tables formées d'hyperliens, dans le document ou dans des menus déroulants. Je continue d'avoir une approche « papier », même avec le numérique, et je me limite souvent à faire des recherches dans le texte pour certains mots précis.

[p. 19] Je suis un peu surpris que la seconde table ne soit pas une sous-table de la première, mais c'est apparemment un choix de `tdmnum` et il est vrai que les `<div ...>` permettent en général de déduire cette structuration, laquelle d'autre part n'est peut-être pas indispendable à reproduire pour la navigation. En d'autres termes, on n'a pas absolument besoin de savoir que la section « Des diverses sociétés et de leurs règles » est une section du chapitre III pour pouvoir naviguer.

[p. 23] Je m'étonne que la préface n'ait pas été mise dans un `<div1>` car au fond elle est au même niveau que le chapitre 1.

[p. 33] Pourquoi les `FOREIGN` ne comportent-t-ils pas la partie identifiant le document (i.e. le `OOOOOOO`) ?

[p. 35] Pourquoi dans cet exemple n'utilise-t-on pas une balise `<div>` pour mettre le « I. Entrée en matière etc. » plus bas que la « Première partie » ?

En passant, je serais intéressé de savoir si la BnF a des projets de réalisation de tables plus fines, permettant de superposer des liens aux images, comme *Google books* le fait sur certains ouvrages. `tdmnum` pourrait permettre cela, si `tdmnum` était étendu avec les coordonnées des liens, tout comme ALTO intègre les coordonnées des mots. On se rapprocherait alors d'un PDF hypertexte.

[p. 37] Le codage ne correspond pas à la table illustrée, mais à celui de **G**:`bpt6k1015246/f128.image`

8. Version du 17 octobre 2013.

[p. 39] La résolution à laquelle **G**:`bpt6k6346000w/f615.image` est donnée dans le référentiel est insuffisante.

[p. 46] Pourquoi « alphabeticque » a-t-il été translitéré en « alphabétique » alors qu'« estre » n'a pas été translitéré en « être » ?

[p. 48] Si 5..8 devient 5,8, comment distingue-t-on les titres renvoyant aux pages 5 à 8 et ceux qui renvoient aux pages 5 et 8 (cas qui pourrait exister) ?

[p. 49] L'extrait contenant une formule mathématique est montré avec une résolution à pleurer ! Il faudrait au minimum que la BnF choisisse des résolutions qui ne soient pas ridicules et ici on devine à peine que l'indice est un θ.

Cela dit, cet exemple est intéressant, parce qu'en allant voir l'image,

G:`bpt6k6438845t/f425.highres`

on se rend compte que la plus haute résolution est celle-là ! Ça, ce n'est pas très bon !

[p. 53] Personnellement, j'aurais tout mis sous « Litérature », car il y a vraiment deux niveaux et il suffit de regarder la page complète pour s'en convaincre : **G**:`bpt6k5475170f/f287.image`

[p. 74-75] Il y a une certaine confusion dans le document, puisque page 17 on nous dit que les tables sont saisies avec `<table>`, `<row>` et `<cell>`, page 25 on nous dit que les index sont saisis avec `<list>` et `<item>`, mais ici les tables des matières et des illustrations sont saisies comme s'il s'agissait d'index. Est-ce voulu ou est-ce une erreur ? Dans la mesure où l'ordre est celui des pages, je ne pense pas que l'on puisse parler d'un index déguisé. Le codage me paraît donc incorrect.

[p. 76-77] On a bien un index, donc saisi avec `<list>` et `<item>`.

[p. 88] **G**:`bpt6k6202330g/f643.image` utilise non pas l'alphabet grec, mais l'alphabet copte qui est apparenté au grec. L'exemple me semble donc mal choisi.

4.8 Référentiel ePub 3 [9]

[p. 10] Il est intéressant de lire que la BnF considère que ceux qui lisent les documents patrimoniaux dans le contexte de Gallica sont plus tolérants aux coquilles que ceux qui lisent ces documents en dehors de ce contexte, c'est-à-dire qui sont davantage habitués à des ouvrages très récents sur livre numérique, et qui ne sont donc pas passés par la phase d'OCR.

9. Version du 29 octobre 2013.

[p. 12] Je ne vois pas pourquoi on exclurait les monographies en plusieurs volumes pour la production d'ePub3 ; pour les autres types de documents évoqués, je comprends.
[p. 13] Il est intéressant de lire que la BnF estime que la consultation web se prête mieux à un travail de recherche que le livre numérique.
[p. 17] Les formules données sur cette page (« cardinal … ») font penser à l'époque reculée où les mathématiques ne connaissaient pas encore les symboles, cela ne facilite pas la lecture !
[p. 39] Il n'y a pas que des formules de mathématiques et de chimie…
[p. 61] (en bas et ailleurs) Indiquer que c'est du MARC.
[p. 75] « le format JPEG offre le meilleur ratio qualité/compression » contredit ce qui est dit ailleurs ; (et p. 77 on mentionne à nouveau le TIFF et le JPEG2000)
[p. 77] Les images qui sont référencées aux formats TIFF ou JPEG2000 pourront-elles être lues par toutes les liseuses ?
[p. 78] « UD » et « UC » sont des acronymes non explicités.
[p. 80] À titre personnel, j'utilise depuis 15 ans la qualité 75% qui me semble suffisante.