Numérisation : Choix technologiques

Dole - 08/11/2007

Catherine Mocellin

Bibliothèque nationale de France
Département de la conservation
Service numérisation

Catherine.mocellin@bnf.fr

La numérisation :
quelques rappels

- Principes de la numérisation
  - Transformer la représentation analogique d'un document en représentation codée en mode binaire (0, 1)
  - À partir des capteurs de lumière du numériseur
- La cellule photoélectrique du numériseur analyse la lumière réfléchie par l'original : le courant varie en fonction de l'intensité du gris, pour devenir maximal pour le blanc
- Un rayon de lumière traduit ces deux valeurs :
  - 1 = passage du courant = lumière = blanc
  - 0 = pas de courant = pas de lumière = noir

La numérisation :

quelques rappels

- L’image numérique est composée d’une juxtaposition d’éléments d’images (pixels) dont la luminosité est quantifiée par une valeur numérique
- Le nombre de pixels (L*l) représentant la dimension physique de l’original détermine la **résolution** de l’image
  (points par pouce : ppi ; ou dots per inch : dpi)

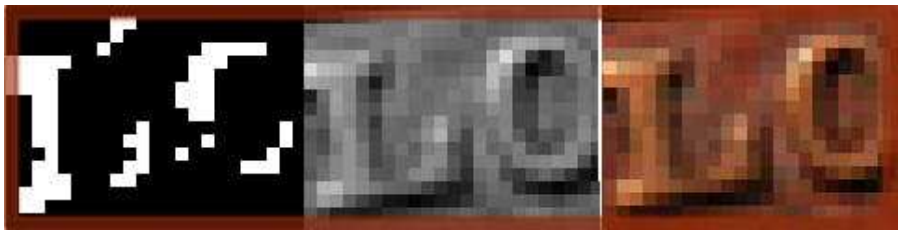


Image agrandie des pixels noirs et blancs, niveaux de gris et couleur

Résolution des images

Aspects des résolutions 72, 150, 300 dpi en tons continus (agrandissement 4x).

- 72 dpi pour la diffusion en ligne (Web)
- 150 dpi pour l'impression bureautique
- 300 dpi pour la substitution et l'imprimerie à l'échelle originale
- Les mêmes résolutions s'appliquent aux documents en niveaux de gris
- Plus la résolution choisie est haute, plus les temps de prise de vue peuvent être longs et plus le poids de fichier est important

Dynamique des couleurs

- Différents codages :
  - le mode *bitonal* (noir et blanc) : reproduction des textes pouvant être illustrés de gravures,
  - Le mode *niveau de gris* : photographies en noir et blanc,
  - Le mode *RVB* (rouge vert bleu) : reproduction de tous documents en couleurs.

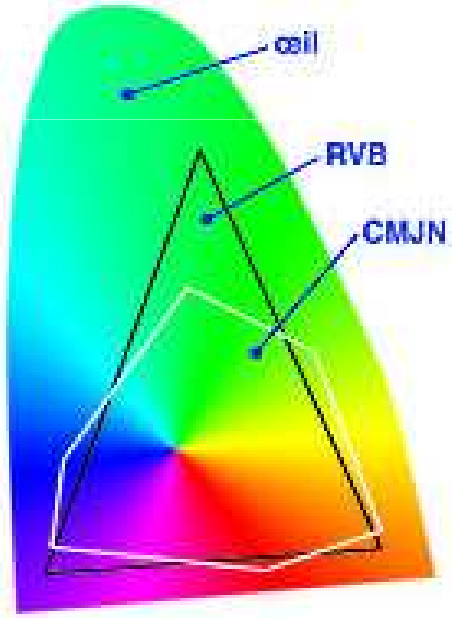
- L'encodage traduit en valeurs numériques la tonalité, la luminance et la saturation de la couleur perçue par l'œil

Espace de couleur

- Un espace colorimétrique regroupe toutes les couleurs visibles selon la technologie employée.

- Ci-contre le gamut de couleurs avec deux espaces colorimétriques (l'ensemble des couleurs visibles ou reproductibles) :

1. œil : les nuances de couleur qu'un œil humain normal peut voir, assez proche de ce qu'on peut obtenir sur les meilleures photographies.
2. RVB : ce qu'un système d'écran (télévision et moniteur informatique) peut représenter,
3. CMJN : la quadrichromie, ce que l'imprimerie traditionnelle peut imprimer.



Rendu fidèle des couleurs

- La bonne reproduction des couleurs dépend d'un grand nombre de variables :
  - le niveau d'illumination au moment de la capture
  - les capacités du système de numérisation

- Utilisation d’un profil ICC pour décrire la manière dont un périphérique restitue les couleurs
- Utilisation de mires normalisées
- Étalonnage : consiste à définir un état colorimétrique de référence

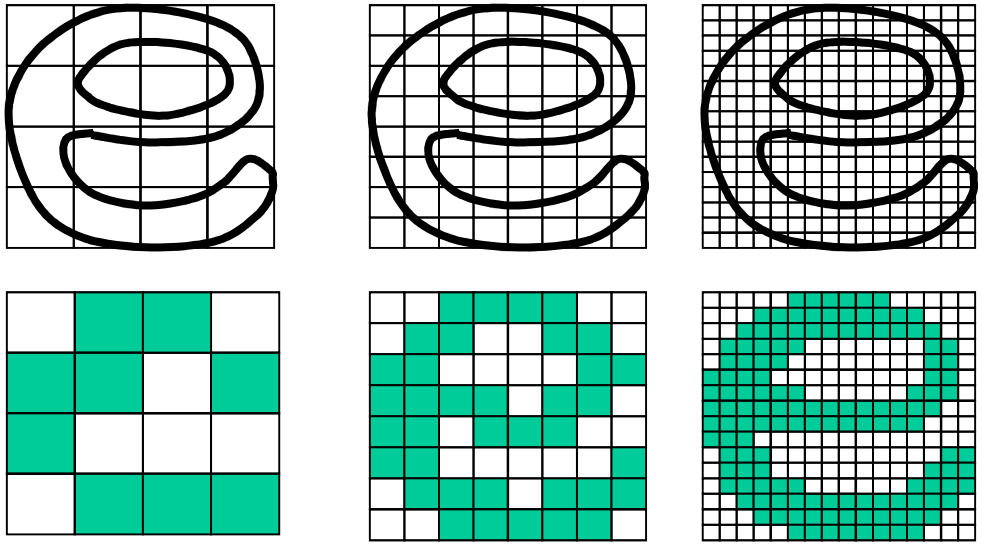
Numérisation en mode image

- Avantages :
  - Un fac-similé
  - Coût inférieur (moins de traitements)
- Inconvénients :
  - Pas de recherche plein texte, pas de manipulations du texte
  - Prévoir un format de diffusion pour limiter les temps de téléchargement
  - Avoir un système d'indexation et de recherche performant et riche

Numérisation en mode texte

- Avantages
  - Indexation et recherche plein texte, autres manipulations
  - Souplesse et portabilité
- Inconvénients
  - Lourdeur de réalisation, coût élevé (balisage…)
- Solutions intermédiaires…
  - Mode image avec points d'accès en mode texte
    - table des matières, index
  - Mode image avec OCRisation
    - Création de fichiers texte contenant les coordonnées des éléments
    - Affichage combiné (PDF multicouche par ex.)

Impact de la résolution pour la conversion OCR



- Fonctionnement
  - Segmentation : découpage de la page et des blocs de texte en « boîtes »
  - Reconnaissance : au sein de chaque boîte, reconnaissance des caractères

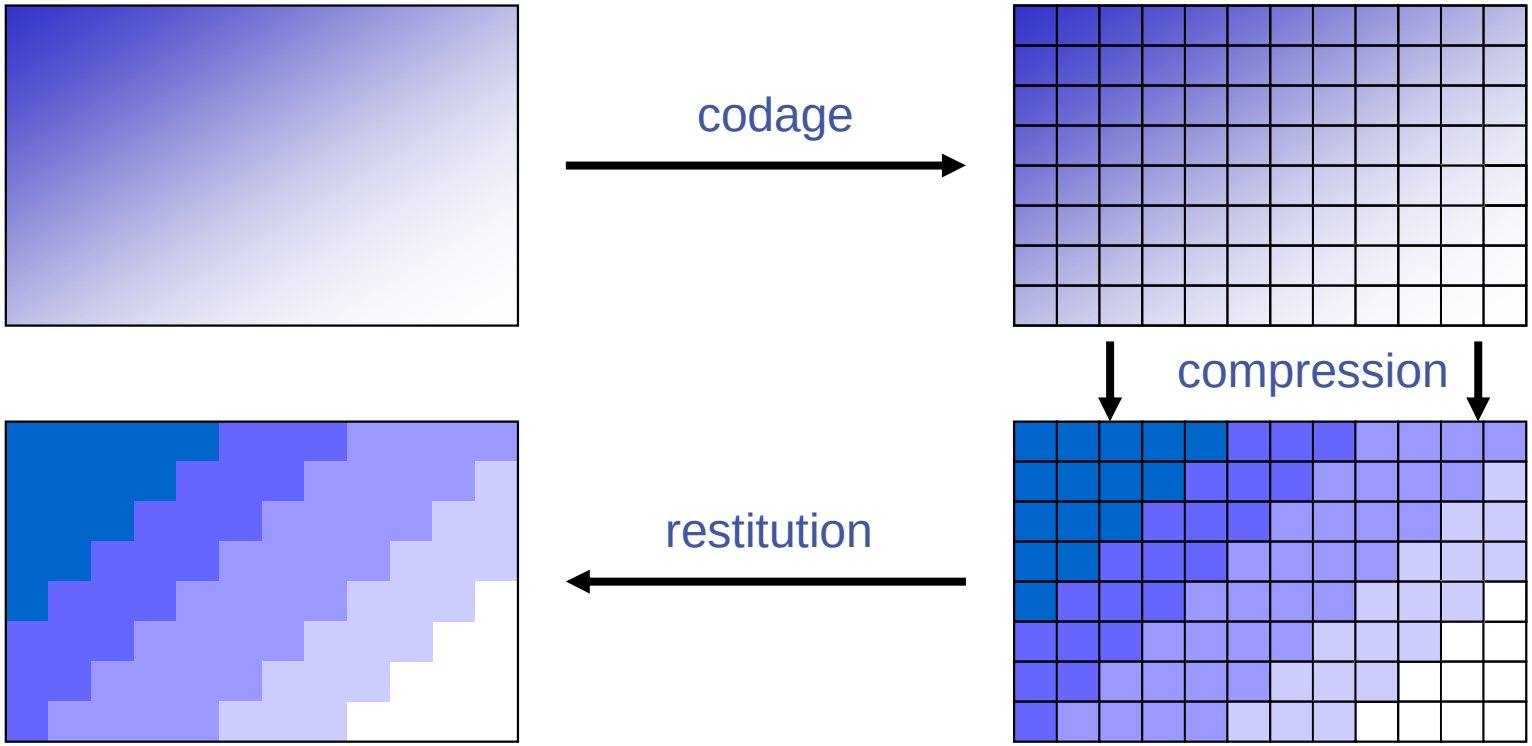
Obstacles pour l'OCR

- Qualité d'impression
  - Courbures et inclinaisons de lignes, caractères déformés : segmentation difficile
  - L'impression de l'original doit être très régulière et propre (pas de taches, etc)
- Polices de caractères
  - Trop resserrées, trop irrégulières (caractères trop gras, trop grands) : risques de confusion entre les caractères
  - Caractères non latins sont mal reconnus (grec, fraktur, gothique, manuscrits…)
- Structure du texte
  - Structure en colonne type presse : nécessité de définir un ordre
  - Éléments non textuels (graphiques, illustrations…)

La compression des images

- Elle permet de réduire la taille des fichiers en supprimant des pixels ou des couleurs
  - Compression sans perte
  - Compression avec perte
- **En aucun cas les formats engendrant des pertes de données irréversibles ne doivent être utilisés pour la sauvegarde des images**

Compression JPEG



Paramétrage du taux de qualité (0 à 100%)

Numérisation : formats

- TIFF (Tagged Image File Format)
  - Format permettant de documenter les images (tags) : dimensions, nombre de couleurs, matériel utilisé, données d'indexation (cote, copyright…)
  - Permet de stocker des images de taille importante sans déperdition de qualité et indépendamment des plateformes et des périphériques
  - Permet l'usage de plusieurs espaces de couleur
  - Ex. : Centres de service de conservation de l'OCLC

- JPEG 2000
  - Compression de meilleure qualité que le JPEG
  - Dégradation « sélective » de certaines zones moins stratégiques de l'image
  - Nécessite des capacités de stockage moins importantes
  - Ex : BN Norvège : format d'archivage à long terme

Numérisation et conservation

- Une numérisation de qualité contribue à la conservation à long terme des documents :
  - Non communication des originaux : la version numérique doit être suffisante pour satisfaire 99 % des besoins des lecteurs
  - Remplacer des originaux manquants et/ou permettre des sorties COM (films, fac-similés)
  - Préservation : il est plus aisé de conserver les fichiers numériques lorsqu'ils sont riches et bien documentés
- Document numérique « de qualité » :
  - Choix optimal de la résolution (ni trop haute, ni trop basse)
  - Indexation et documentation de chaque image, du document numérique, du procédé de numérisation
  - Fidélité rigoureuse à l'original et qualité de la prise de vue
  - Non compression, ou compression réversible
  - Format(s) le(s) plus ouvert(s) possible(s)

Choix technologiques : paramètres

- Une bonne numérisation est une adéquation entre :
  - Les objectifs du projet (numérisation « de masse », numérisation sélective…)
  - Les moyens financiers
  - Les caractéristiques physiques du document/du fonds
  - Les capacités du matériel et le système de numérisation
  - La fidélité de l'image numérisée à l'original
  - Les paramètres de numérisation : couleur ? Résolution ? Format de sortie ? Format d'affichage ?
  - L'usage du fichier numérisé : conservation pure ? Diffusion à titre gratuit, payant ? De consultation ? Quels services associés ?
  - La portabilité en réseau (temps d'affichage, etc)
  - Les moyens de conservation à long terme
  - Les moyens de signalement pour la recherche

Numérisation : caractéristiques physiques

- Support :
  - Nature et fragilité : papier, film, vélin, …
  - Aspect : opacité, couleur
  - Présentation : dimensions, reliure, feuillets montés sur onglet, dépliants et paperoles, etc
  - Échelle d'agrandissement éventuelle
- Contenu :
  - Type : photo, texte, gravure, graphiques, cartes, etc
  - Qualité : graphie, contraste, taches éventuelles
  - Mode d'obtention : imprimé, manuscrit, dessin, etc
- Plus un document/un fonds est hétérogène plus la numérisation est complexe et lourde car nécessite des réglages particuliers

Numérisation BnF : profondeur d'acquisition

- Choix de la profondeur d'acquisition (nombres de bits utilisés pour représenter chaque pixel) :
  - Noir et blanc (2 tons) : imprimés, dessins au trait, graphiques
  - Niveaux de gris (256 tons) : gravures et photos noir et blanc, certains imprimés (peu contrastés ou tachés)
  - Couleur (16,7 millions de tons) : documents en couleur ou contenant des pages en couleur
- De la profondeur d'acquisition dépend le poids du fichier et donc les moyens à prévoir pour gérer la portabilité en réseau (stockage, affichage) :

A 4, 300 dpi	Mo	Ko
Noir et blanc	1.04 Mo	1 062 Ko
Niv. gris	8.30 Mo	8 497 Ko
Couleur	21.89 Mo	24 490 Ko

Numérisation BnF : résolution

- Choix de la résolution :
  - Gamme possible de 120 à 1200 dpi
  - Pour la conservation :
    - Documents < A6 : 300 dpi
    - Documents > A6, ou manuscrits aux petites enluminures : 600 dpi
- La résolution doit être pertinente et adaptée au type de document
- La résolution impacte le poids des fichiers

A 4	300 DPI		600 DPI	
	Mo	Ko	Mo	Ko
Noir et blanc	1.04 Mo	1 062 Ko	4.15 Mo	4 248 Ko
Niv. gris	8.30 Mo	8 497 Ko	33.19 Mo	33 987 Ko
Couleur	21.89 Mo	24 490 Ko	99.57 Mo	101 961 Ko

Numérisation BnF : prises de vue

- Fidélité à l'original lors de la prise de vue
  - Reproduire au plus près l'original sans l'améliorer
  - Pas de retouche en post-production
  - Réglages optimisés lors de la prise de vue (éclairage, contraste, marges, etc)
  - Une page / image, dans son intégralité, sans vue de détail
  - Insertion de fonds de couleur neutre pour les projets iconographie
- Traitements post-numérisation autorisés sur les images
  - Recadrage
  - Détourage jusqu'au bord extérieur des pages
  - Redressement
  - Remise dans l'ordre des images
- Utilisation de mires

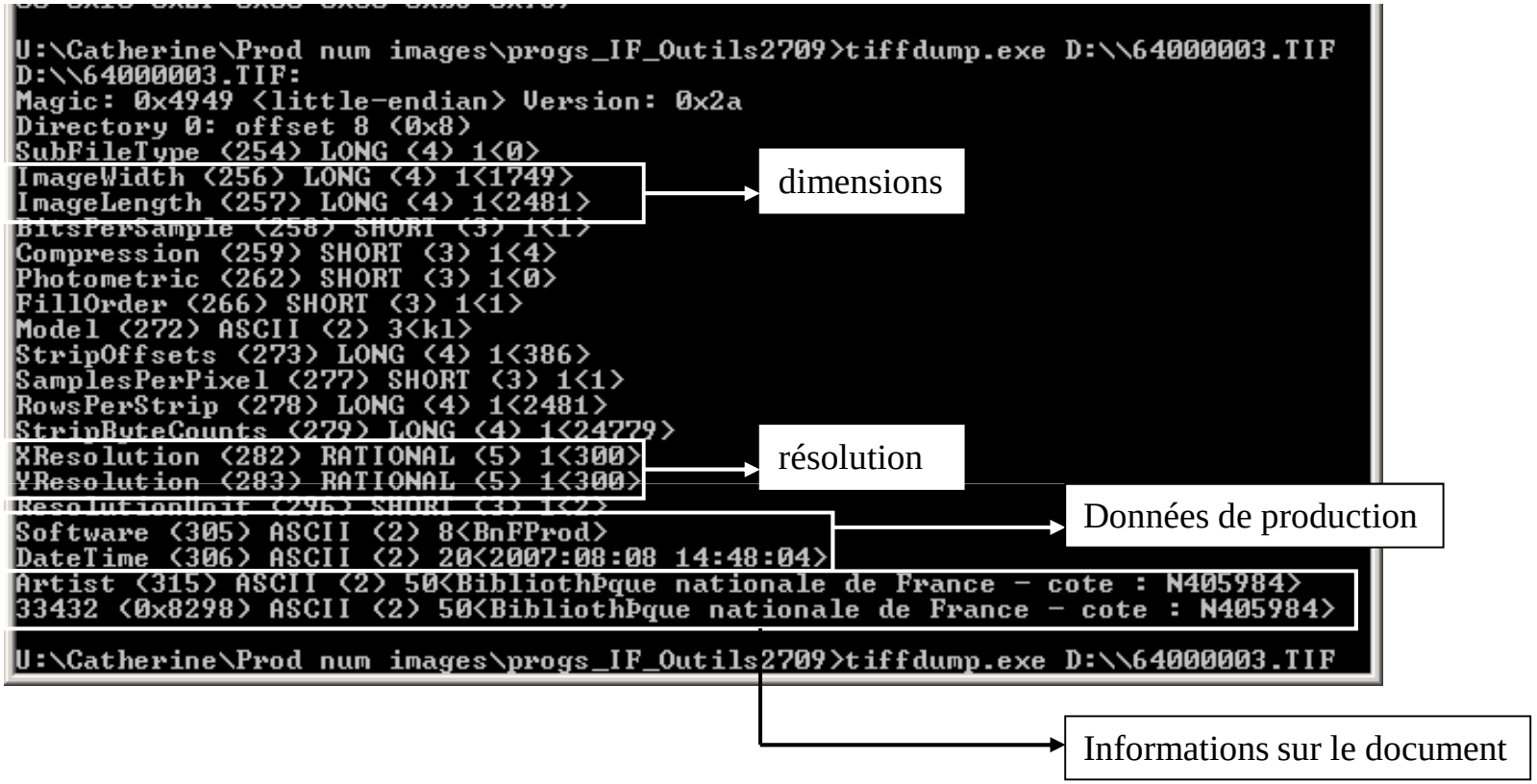
Numérisation BnF : formats

- Distinction format d'archivage – format de diffusion
  - Assurer l'indépendance du système de conservation par rapport aux outils et standards de consultation
  - Contraintes d'accès (temps d'affichage, droits, etc)
  - Assurer de bonnes conditions de consultation du document numérique
- Formats de diffusion
  - Imprimés : PDF – PDF multicouches pour l'ocr
  - Images fixes : JPEG

Numérisation BnF : formats

- Master destiné à l'archivage :
  - en TIFF v6 non compressé
    - Documents intégralement en couleur
    - Documents intégralement en niveau de gris (presse, essentiellement)
  - compression en CCITT/UIT-T GIV pour les imprimés en noir et blanc seulement
  - OCR : format ALTO
- En-tête des fichiers renseignée par des métadonnées de préservation :
  - Informations techniques sur l'image
  - Informations de production
  - Informations administratives

Métadonnées d'une image TIF



Informations techniques et de production propres à l'image

Numérisation : indexation du document numérique

- Fichier XML de métadonnées « refnum »
  - « Bibliographie » : données descriptives
  - « Production » : liées à la production en tant que telle (date de numérisation, nombre d'images, historique, n° de support...)
  - Données de structure : organisation des données produites
    - Correspondance entre le n° image et le n° logique de la page
    - Données particulières à certaines images

Numérisation : indexation du document numérique

- Typage des pages
  - Mise en valeur d'accès spécifiques au document : tables des matières, index, page de titre, de couverture
- Nécessite la définition de règles de typage
  - En fonction de la complexité de la pagination / foliotation
  - En fonction de l'utilisation du document par un lecteur, à court ou long terme

Conclusion

- Les technologies employées doivent être les plus libres et pérennes possibles
- Choisir des formats standardisés et stables dans le temps
- Les choix technologiques dépendent du type de projet et du fonds
- Ils doivent prendre en compte tous les aspects de la gestion d'un document :
  - Architecture informatique de la bibliothèque
  - Destination des documents (publics, droits d'accès aux versions numériques, modalités de consultation et de recherche, etc)
  - Capacités de stockage à long terme
  - Évolution des technologies et des outils de consultation
- Ce sont des problématiques identiques aux bibliothèques traditionnelles