Pour des collections numériques durables

Les formats de numérisation

Alain Boucher

Directeur de la numérisation

Bibliothèque et Archives nationales du Québec

Stage pratique sur la numérisation

Dakar, 24 janvier – 1er février 2011

Au sommaire

- Les normes dans le monde des technologies de l'information
- La technologie de la numérisation (un simple rappel des éléments essentiels)
- Les formats de fichiers pour la conservation à long terme et pour la diffusion sur Internet

Les normes dans le monde des technologies de l'information

- Dans bien des domaines (génie, comptabilité, bibliothéconomie, construction, etc.), les normes s'établissent par consensus de spécialistes et s'appliquent universellement.
- Dans le monde des technologies de l'information, ce sont fréquemment les lois du marché qui fixent les normes. Les normes qui s'imposent sont celles qui connaissent un succès commercial.
- Contourner une norme et en établir une nouvelle signifie souvent la réussite économique d'un produit et surtout d'une entreprise.

La technologie de la numérisation

- Un objet numérique (texte, image, son) est une reproduction du réel qui peut être traitée et exploitée par les moyens informatiques.

- Comme tous les autres aspects de l'informatique, c'est un domaine qui évolue constamment.

- Dans les institutions de mémoire (bibliothèques, archives, musées), il faut évidemment s'en tenir aux «valeurs sûres», qui offrent les meilleurs garanties de pérennité.

- «Si c'est nouveau, c'est mieux» ? Pas toujours!

La numérisation:
une question de conservation et de diffusion

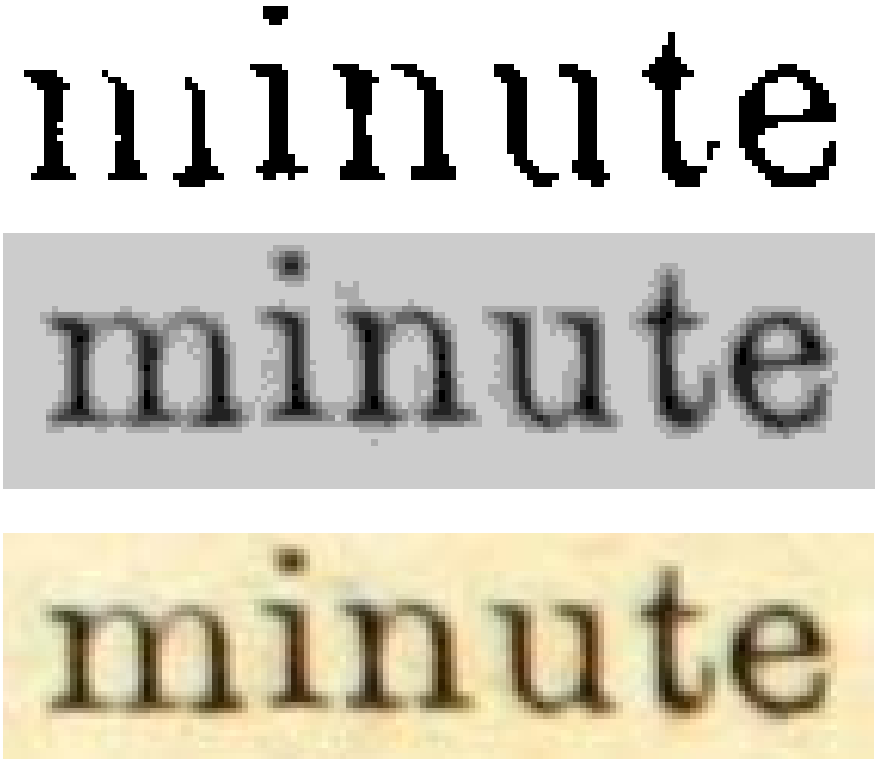
- En matière de numérisation, il y a deux préoccupations à prendre en compte :
  - Créer des objets numériques de la meilleure qualité possible et les archiver dans une perspective de long terme.
  - Mettre ces objets à la disposition des utilisateurs, en leur offrant ce qui est approprié dans l'état courant des technologies de diffusion sur Internet ou sur d'autres supports.

Une richesse qui se paie…

- Plus l'information qu'on veut conserver est riche, plus la taille du fichier résultant est considérable:
  - 1 bit = noir et blanc (bitonal)
  - 8 bits = 256 couleurs ou tonalités de gris
  - 24 bits = 16 millions de couleurs

Une richesse qui se paie…

Un simple mot sur papier jauni…



Une richesse qui se paie…

- La définition (résolution) des images se mesure en points par pouce :

  72 ppp = qualité minimale (écran d'ordinateur)

  300 ppp = documents textuels courant

  600 et 1200 ppp = images de haute qualité

  2400 ppp et + = négatifs 35 mm, diapositives, etc.

Le mieux est-il l'ennemi du bien ?

- Pour les documents de type image, la résolution à 300 ppp est généralement suffisante pour l'archivage et la reproduction « à l'identique » (mêmes dimensions que l'original).
- On adopte une résolution supérieure si on envisage la reproduction agrandie des documents (par ex., photographies, cartes postales).
- Adopter des normes très élevées peut entraîner des coûts importants sans avantages réels (numériser à 600 ppp donne des fichiers quatre fois plus gros qu'à 300 ppp).

Les formats les plus courants

- Le format TIFF non compressé est la norme la plus généralement adoptée pour la conservation à long terme.
- Le format JPEG est le format le plus courant pour la diffusion des images de type photographique.
- Pour les documents textuels en mode image ou image et texte (avec possibilité de recherche), le format PDF s'est largement imposé pour la diffusion.

Les formats les plus courants

- Le format PDF/A (sous-ensemble du format PDF, approuvé par l'ISO) est aussi admis comme format de conservation.

- Le format PDF au complet est devenu lui aussi une norme ISO au cours des derniers mois.

Un paysage qui évolue…

- De nouveaux formats de grand intérêt voient le jour continuellement :
  - JPEG2000
  - MPEG-21

- Leur adoption est freinée par la nécessité d'employer de nouveaux logiciels pour les exploiter ou d'adapter les logiciels existants.

Le cas des documents bureautiques

- Au-delà de la numérisation des documents en mode image, se pose la question de l'archivage et de la diffusion des documents numériques créés avec les logiciels de bureautique ou autres (PAO, etc.).

- Le langage de balisage XML est devenu la norme pour l'archivage des documents de bureautique. Microsoft l'a adopté comme standard avec *Office 2007.*

Le cas des documents bureautiques

- Reste que la normalisation n'est pas encore bien fixée:
  - ODF (Open Document Format), approuvé par l'ISO
  - Open XML de Microsoft
- Pour la diffusion, le format PDF (image sur texte) reste la solution la plus avantageuse.

Formats ouverts et formats propriétaires

- Le débat au sujet des formats « libres » ou « ouverts » par opposition aux formats « propriétaires » n'est pas vraiment fructueux.

- Un format « propriétaire » très largement répandu est préférable à un format « ouvert » qui n'a qu'une diffusion confidentielle.

Finalement, qu'en penser ?

- En matière de formats et de normes, les choix sont multiples et ils ne sont pas appelés à diminuer.
- L'attitude à prendre: il vaut mieux se tromper avec tout le monde que d'être tout seul à avoir raison (adopter les solutions les plus largement répandues).
- Une bonne dose d'«humilité technologique» s'impose:
  - Les décisions prises aujourd'hui n'ont pas valeur éternelle
  - Il y aura tôt ou tard nécessité d'investir dans le changement