La numérisation

Médiaquitaine

6 octobre 2009

i.westeel@cr-npdc.fr

I. Les enjeux

Une question de visibilité

Enjeux

- Diffusion / Accessibilité / Visibilité pour le public (public distant)
- Valorisation des collections
- Préservation de documents originaux

- Mise en place de système de recherche documentaire performant / renouvellement de la recherche

- Projets séduisants et aidés
- Offre de nouveaux services

Quels documents ?

- Souvent priorité aux documents uniques, libres de droit et d'intérêt local.
  - Manuscrits : programmes nationaux
  - Presse locale : problèmes cumulés : complétude, format, collections reliées et volumétrie importante – indexation difficile
  - Fonds d'archives : nombreux exemples, volumétrie importante
  - Fonds contemporains : questions juridiques (ne sont pas insurmontables)

Réalisations nombreuses

- Patrimoine numérique (www.numerique.culture.fr) : catalogue des fonds numérisés du Ministère de la Culture et de la Communication
  - Plus de 500 institutions – 1500 collections environ
- Visibilité par le moteur de recherche « Collections » du portail Culture.fr et par le portail européen multilingue « Michael » (description des collections numérisées du patrimoine culturel européen)
- Numes 2009 – Ministère de l'Enseignement supérieur et de la Recherche

Mais…

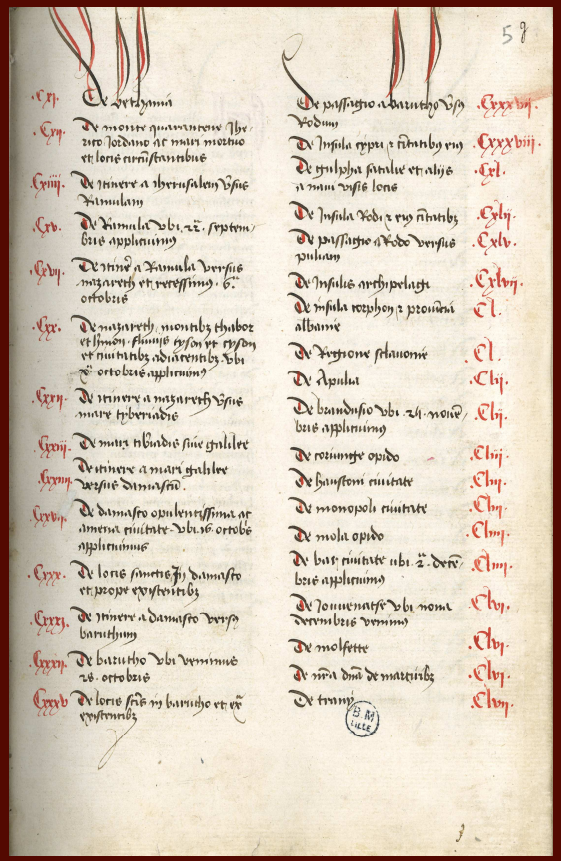
- 52 % des documents numérisés par les établissements culturels français ne sont pas en ligne
- Les internautes ne vont pas sur les sites des bibliothèques…
  - Près de 90 % des recherches passent par un moteur de recherche
  - Mai 2008 - Plus de 90 % des requêtes effectuées sur Internet en France passent par Google (Baromètre Xiti - Institut d'études sur le web)

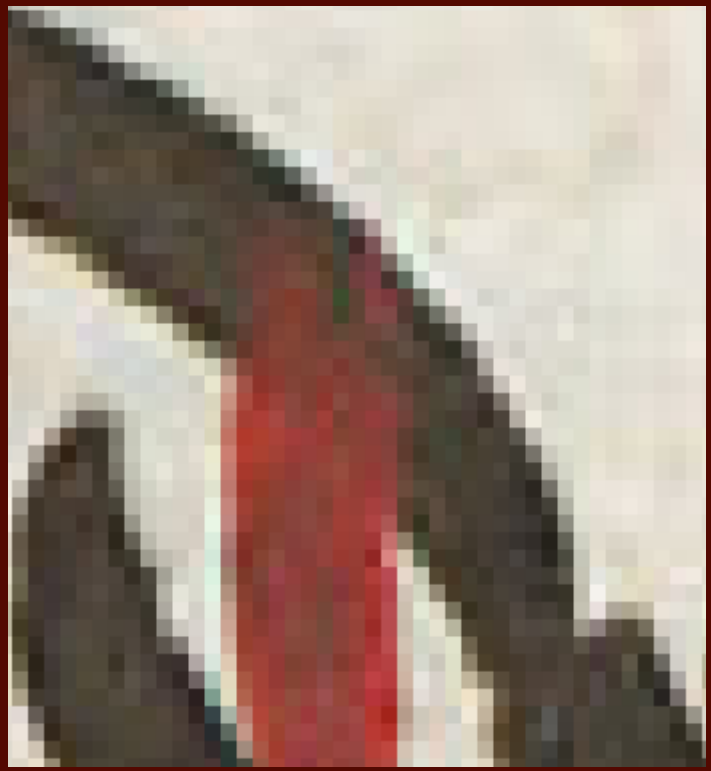
II. Les techniques ?

Un savant mélange

(bibliothécaires et informaticiens)

Des pixels






bit bande passante cadrage cédérom
compression conservation
consultation contrôle couleur dpi exif
fichier format image iptc jpeg logiciel
lzw matrice de pixels métadonnées niveaux de gris
noir et blanc nommage norme ocr
octet optique pdf pixel profil ICC
reconnaissance de formes redressement
résolution scanner texte tiff
traitement xmp

Qu’est ce qu’une image numérique ?

- Fichier informatique contenant (sous forme d’un code binaire) les données d’une image
  - Image analogique (argentique) / image numérique (informatique)
- Sa visualisation nécessite sa restitution sur un écran ou son impression sur une feuille de papier – nécessité d’un support

On distingue

- Image numérique native / image numérisée
- Image de pixels (= image bitmap) / image vectorielle (données décrites par des expressions mathématiques)

Bit

- Bit (binary digit) = la plus petite unité d'information manipulable par une machine informatique
- Cette information est représentée par des impulsions électriques : 2 états : « 1 » ou « 0 »
- 1 bit = 2 états (soit 1, soit 0)
- 2 bits = 2x2 = 4 états (00, 01, 10, 11)
- 3 bits = 2x2x2 = 8 états
- 8 bits = 2x2x2x2x2x2x2x2 = 256 états

Octet

- Octet (*byte* en anglais *by eight*) = unité d'information composée de 8 bits

- Souvent utilisé pour coder un caractère, une lettre…
  - A = 01000001
  - a = 01100001

Pixel

- Pixel (ou élément d'image - abréviation de l'anglais *Picture element*) = unité élémentaire de l'image numérique
- Une image bitmap est une matrice de pixels (2 dimensions : horizontale et verticale)


SUPRASCAN
6002 RGB
DELL

La résolution

- Résolution (spatiale) = résultat du passage du laser sur le document. Elle mesure le nombre de points sur un pouce : dpi (Dots Per Inch) ou ppp (Pixels Par Pouce). [1 Inch = 1 Pouce = 2,54 cm]. Par exemple 300 dpi
  - *D'une façon générale, plus le nombre de pixels sera élevé, plus on verra les détails de l'image mais plus le fichier sera lourd.*
  - *Ordre de grandeur : A4 = 300 dpi*

La dimension ou définition (en pixels)

- largeur et longueur de l'image (en pouces) multiplié par la résolution (en dpi) = définition (en pixels)

- Par exemple pour une photographie 13 x 18 cm numérisée en 400 dpi, les dimensions seront :
- en largeur : (18 : 2.54) x 400 = 2834 pixels
- en longueur : (13 : 2,54) x 400 = 2047 pixels
- L’image sera donc représentée par une matrice de pixels de 2834 x 2047 = 5 801 198 pixels.

Exercice

- Une page en format A4 (21 x 29,7 cm) numérisée à la résolution 300 dpi
- Quelle dimension pour la matrice de pixels ?

Corrigé

- Considérons une page en format A4 (21 x 29,7 cm) numérisée à la résolution 300 dpi. Ses dimensions sont :
  - Largeur (21 : 2,54) x 300 = 2480 pixels
  - Longueur (29,7 : 2,54) x 300 = 3507 pixels
  - La page est donc représentée par un tableau (une matrice) de 2480 x 3507 = 8.697 millions de pixels

Profondeur de bits ou profondeur de couleur

- Nombre de bits utilisés pour représenter l'information de chaque couleur

- Trois modes
  - Noir et blanc (bitonal)
  - Niveaux de gris
  - Couleurs

Quatre grandes possibilités

Nombre de bits pour coder la profondeur du pixel	Description
1 bit (chaque pixel occupe un bit)	Noir et blanc ou image binaire ou image au trait. Image dont chaque pixel ne peut avoir que 2 valeurs : blanc ou noir (0 ou 1)
8 bits = 1 octet	Niveaux de gris. Les 8 bits génèrent 256 valeurs par pixel : du noir pur (0) au blanc maximal (255)

24 bits = 3 octets	RVB (3 couches de couleur). Chaque couleur est codée sur 1 octet = 8 bits Chaque pixel est codé sur 3 octets donc 24 bits. Rouge = 0 à 255 Vert = 0 à 255 Bleu = 0 à 255 Donc 256 x 256 x 256 = 16 millions de couleurs

32 bits = 4 octets	CMJN (4 couches de couleur) Chaque couleur est codée sur 1 octet = 8 bits Chaque pixel est codé sur 4 octets donc 32 bits. Cyan = 0 à 255 Magenta = 0 à 255 Jaune = 0 à 255 Noir = 0 à 255 Donc 256 x 256 x 256 x 256 = plus de 4 milliards de couleurs différentes

Taille de fichier ou poids d'une image

- La taille du fichier est obtenue en multipliant la définition de l'image (longueur x largeur en pixels) par la profondeur de bit.
  - *On divise cette quantité par 8 pour un résultat en octet.*

- 1 Kilo-octet (Ko) = 1024 octets
- 1 Mégaoctet (Mo) = 1024 Ko
- 1 Gigaoctet (Go) = 1024 Mo
- 1 Teraoctet (To) = 1024 Go

- Par exemple pour une photographie 13 x 18 cm numérisée en 400 dpi, les dimensions seront :
- en largeur : (18 : 2.54) x 400 = 2834 pixels
- en longueur : (13 : 2,54) x 400 = 2047 pixels
- L’image sera donc représentée par une matrice de pixels de 2834 x 2047 = 5 801 198 pixels.

- Reprenons l’exemple de la photographie numérisée. Le poids de l’image dépend alors du mode de numérisation utilisée :
- noir et blanc : pixel codé sur un bit (noir ou blanc)
- niveaux de gris : pixel codé sur un octet soit huit bits, soit 256 niveaux de gris (du blanc au noir)
- couleur : pixel codé sur 3 niveaux d’octets (soit 24 bits) correspondant aux trois couleur RVB (Rouge, Vert, Bleu). Chacune de ces trois couleurs est codée sur 8 bits, soit 256 niveaux.
-

- Noir et blanc : 5 801 198 / 8 = 725 149 octets = 708 Ko
- Niveaux de gris : 5 801 198 = 5 801 198 octets = 5.532 Mo
- Couleur : 5 801 198 x 3 = 17 403 594 octets = 16.597 Mo

Choix

- Type de document (texte, image, carte…)
- Taille du document et taille des caractères
- Exploitation finale de l’image numérisée

Mode texte / mode image

- Mode image = codification du document en une matrice de pixels. Le document est une image.

- Mode texte
  - Saisie directe (ex Traitement de texte)
  - Techniques OCR reconnaissance optique de caractères

Le format TIFF = Format de conservation

- TIFF = Tagged Image File Format
- Format matriciel
- Algorithmes de compression (LZW, CCITT Groupe 6…)
- Appartient à Adobe, spécifications publiques

Le format JPEG = Format de diffusion

- JFIF = JPEG File Interchange Format
- Algorithme de compression JPEG créé par le Joint Photographic Experts Group ISO/CCITT en 1999 (ISO 10918-1)
- Spécifications publiques

Le format PDF

- PDF = Portable Document Format
- Représentation du document
- Développé par Adobe ; spécifications publiques

- Existence du format PDF - Océrisé

- Maîtrise de techniques pointues et en constante évolution
- Prise en compte du facteur temps

- Pour des opérations menées en interne ou confiées à un prestataire extérieur

Donc…
de l'informatique et des informaticiens

mais aussi…

… des techniques plus «traditionnelles»

- Observation et préparation des documents
  - Caractéristiques physiques des documents
  - Classement, récolement, comptage
  - Déreliage si besoin, restauration, conditionnement
  - Tableaux de récolement
- Etablissement des métadonnées
  - Traitement documentaire, description
  - Nouveaux formats : XML – Dublin Core, EAD, TEI…
  - Rigueur « informatique »

Définition

- Une métadonnée est une donnée qui définit et décrit une autre donnée
  - ISO/IEC 11179-3 : Metadata = data that defines and describes other data
- « de l'information structurée qui décrit, explique, localise la ressource et en facilite la recherche, l'usage et la gestion »
  - National Information Standards Organisation (NISO), Understanding Metadata, 2004,

- les métadonnées descriptives ou informations d'identification : titre, auteur, mots-clés…
  - Accès – indexation / Exemplaire – identifiant pérenne URL
- les métadonnées techniques ou métadonnées de structure c'est-à-dire les données sur la version du document : date, format… ainsi que les liens vers les ressources apparentées (source ou en relation) (fichiers et relations)
- les métadonnées administratives : propriété intellectuelle, droits d'accès, informations sur la préservation (historique des modifications) et l'archivage pérenne de la ressource

- Gestion des documents
  - Au niveau le plus bas des fichiers numériques
  - Au niveau le plus haut : la collection dans son ensemble
  - **Granularité**

La granularité de l'information

- Cataloguer au plus près de la ressource ? (indexation la plus précise possible)

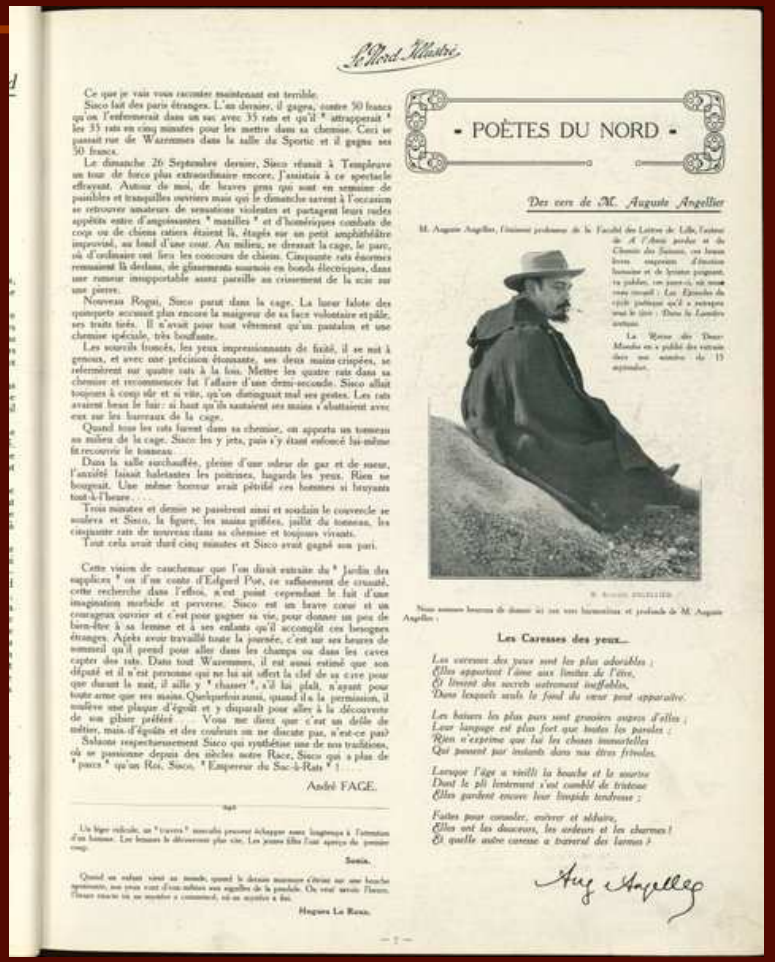
- Faire des choix
- Moyens nécessaires

Collections diverses – systèmes de navigation difficiles à mettre en place

- Description à la pièce et pour un ensemble
- Navigation par dossier et par image
- Paramétrage des moteurs de recherche

Granularité : page d'une revue

- Le Nord illustré : bi-mensuel d'actualité régionale – Lille : Nuez, 1909-1914
  - 1ère année, n°1, 15 octobre 1909
    - p. 7 Poètes du Nord : Des vers de M. Auguste Angellier (une photographie)
      - Un poème : Les Caresses des yeux…



Le Nord Illustré

- POÈTES DU NORD -

Des vers de M. Auguste Angellier

Les Caresses des yeux...

Les caresses des yeux sont les plus adorables ;
Elles appartient l'âme aux limites de l'être,
Et livrent des secrets notrement ineffables,
Dans lesquels seuls le fond du cœur peut apparaître.

Les baisers les plus purs sont grossiers auprès d'elles ;
Leur langage est plus fort que toutes les paroles ;
Rien n'exprime que lui les choses immortelles
Que pressent par instants dans nos âmes frivoles.

Lorsque l'âge a vieilli la bouche et le sourire
Dont le pli lentement s'est comblé de tristesse
Elles gardent encore leur limpide tendresse ;

Faites pour consoler, enivrer et séduire,
Elles ont les douceurs, les ardeurs et les charmes !
Et quelle autre caresse a traversé des larmes ?

Aug Angellier

Dublin Core

- 1995 à Dublin (Ohio) [siège d'OCLC]
- Format de description minimale
- Catalogage simplifié
- 15 zones : facultatives, répétables, dans n'importe quel ordre

- Février 2003. Norme ISO 15836

15 éléments en 3 groupes

- Contenu
- Propriété intellectuelle
- Instanciation (version de la ressource)
  - Ex. instanciation en HTML, en XML, sur papier…

Contenu	<title> <subject> <description> <type> <source> <relation> <coverage>
Propriété intellectuelle	<creator> <publisher> <contributor> <rights>
Instanciation ou version	<date> <format> <identifier> <language>

Un projet de numérisation…

- Sélection du corpus et des contenus à numériser
- Analyse juridique
- Validation
- Planification / Année budgétaire / Code des marchés publics
- Catalogage et description
- Récolement / Constitution des lots / Bordereaux d'envoi
- Numérisation
- Suivi des prestations
- Contrôle-qualité
- Retour des documents en magasin
- Migration des données dans le système de stockage (archivage ?)
- Mise en ligne
- Accompagnement éditorial
- Communication…

Limites et difficultés

- Le projet n’est pas inscrit dans le projet d’établissement
- Le projet n’est pas inscrit dans le schéma directeur
- Le projet n’a pas sa place dans l’organigramme de l’établissement (mise en place d’une équipe)
- La première place revient à la technique et non aux contenus

Echec…

- Offre en conseils limitée
- Offre logicielle complexe et limitée
- Travail sur les recherches fédérées, les portails
- Interactions avec le public ?

Conditions pour réussir une bibliothèque supplémentaire ?

- Projet inscrit dans la politique d'établissement et dans le schéma directeur
- Budget de fonctionnement – projets en mouvement
- Travail sur les données et les métadonnées – respect des standards et des normes – responsabilité du chef de projet
- Descriptions échangeables avec d'autres établissements : OAI, documenter les projets
- Perspective : élaborer des projets à court, moyen et long terme
- Mettre en ligne – navigations à inventer

- Un gros chantier de formation
- Un apprentissage long
- Des contacts nécessaires
- Veille, curiosité et « aimer le web »…

La mise en ligne

- Un “logiciel de gestion de documents électroniques”
  - Stocker l'information
  - Rechercher l'information
  - Gérer l'information
  - Diffuser l'information

La mise en ligne - exemples

- Gallica (BnF)
- Europeana
- Google Book Search

- Offres « portail »

- Logiciels Photothèques
  - Algoba – Orphea Studio
    - Musée du quai Branly, Bibliothèque nationale du Québec
  - Orkis – Ajaris Pro
    - Marie-Claire, Agence l'Illustration
  - Armadillo
    - Bibliothèque numérique des Champs libres - Rennes
  - Phraseanet IV
    - Libris Lille 3
  - Contentdm
    - SICD Universités de Strasbourg

- Solutions propriétaires
  - Arkhenum
    - Musée Jules Verne Nantes ; Bibliothèque d’Abbeville Fonds Macqueron ; Bibliothèque d’Orléans

- Briques logicielles libres
  - W3line
    - Bibliothèque numérique de Roubaix ; Normannia

- Logiciels libres
  - Eprints (GPL)
  - DSpace
  - HAL
  - Greenstone (GPL)
    - Paperspast (journaux et périodiques de Nouvelle-Zélande 1840-1915)
    - Bibliothèque de Bourg en Bresse (http://www.bourgendoc.fr)
  - SDX / Pleade (version 3)

- Mise en ligne des inventaires d'archives
  - Pleade
    - CHAN
    - Bibliothèque municipale de Lyon

- Cas particuliers
  - Cnum
  - Internum
  - BIUM
  - BVH

Les outils « Turning the pages »

- British Library Turning the pages
  - http://www.bl.uk/onlinegallery/ttp/ttpbooks.html
  - Alice au pays des merveilles
- Issuu (http://issuu.com)
  - Bibliothèque de Caluire et Cuire
- Utilisation de Flash et PDF

Flickr

Aller vers les internautes, utiliser leurs « outils »

- 16 janvier 2008 Library of Congress sur Flickr
  - Photos d'actualité des années 1910, Photos en couleurs des années 1930-1940
  - 20 mars 2008 : 50 photos supplémentaires
  - 68 notices bibliographiques modifiées

Aller vers les internautes, utiliser leurs « outils »

- Février 2007 Projet PhotosNormandie
  - Conseil régional de Basse Normandie – site Archives Normandie 1939-1945
  - Bataille de Normandie (juin-août 1944)
  - 2763 photos
  - 3439 légendes complétées
  - Documentation collaborative

Conclusion : “Une bibliothèque numérique qui tourne bien”

Valoriser les contenus

Améliorer l'accès

Fournir de nouveaux outils

Une navigation fluide, efficace et sans défaut (fonctionnalités de base)

\+ des fonctionnalités simples qui pourront être enrichies

\+ un système ouvert, durable, évolutif et des métadonnées bien réfléchies (le chef de projet est responsable des métadonnées)

« Les bibliothèques numériques sont des organisations qui offrent des ressources, y compris en personnel, pour sélectionner, structurer, offrir un accès intellectuel, distribuer et conserver l'intégrité de documents sous une forme numérique »

IFLA. *Digital libraries : definition, issues and challenges*, mars 1998