La numérisation des textes et des images : techniques et réalisations.
Univ. Lille 3. Janvier 2003

Le Conservatoire numérique des arts et métiers : historique du projet et organisation du site.

Pierre Cubaud, Geneviève Deblock
Avec la collaboration de Marie-Christine Radix
Conservatoire National des Arts et Métiers, Paris
{cubaud, deblock, radix}@cnam.fr

Introduction

Les bibliothèques numériques ont bénéficié ces dernières années du progrès constant des technologies de captation, stockage et transmission numérique ainsi que de la chute de leur coût [1]. Le développement du World-Wide-Web a également permis d'atteindre un public international considérable par le biais d'une interface standard et ergonomique. L'usage de ressources numériques est maintenant une réalité quotidienne pour de nombreux chercheurs : banques de données bibliographiques, mais aussi accès aux documents primaires que sont les revues électroniques, actes de conférences, etc.

En France, le projet de numérisation de la bibliothèque nationale (Gallica) a démontré que la diffusion sur Internet d'un vaste corpus d'ouvrages de caractère patrimonial numérisés en mode image pouvait rencontrer un lectorat très important. Le temps est donc venu pour les institutions publiques plus modestes d'entreprendre à leur tour un programme de numérisation et de diffusion des fonds dont elles sont dépositaires. C'est le cas de la Bibliothèque du Conservatoire National des Arts et Métiers (CNAM) qui dispose d'un fonds patrimonial de 60 000 ouvrages scientifiques et techniques.

Après un bref historique et la présentation des orientations retenues, nous aborderons plus particulièrement les options choisies pour le traitement des fac-similés, puis l'organisation du site. Nous décrivons en annexe quelques points pratiques concernant le nommage des fichiers images et les méta-données.

Le CNUM : historique du projet, contenu

L'idée première de la Bibliothèque du CNAM était de préserver et valoriser un fonds patrimonial très consulté, majoritairement composé d'ouvrages et de périodiques du 19^{e} siècle, et menacé par l'acidité des papiers qui le composent. Le microfilmage pratiqué jusqu'alors pouvait être avantageusement complété, puis remplacé par la numérisation. Le centre d'études et de recherches en informatique (CEDRIC) avait par ailleurs développé une expertise en matière de diffusion de documents numériques sur l'internet avec le développement de la base de textes classiques francophones ABU [6].

Le " Conservatoire numérique des arts et métiers " a démarré en janvier 1998 comme projet interne et autofinancé de trois services du CNAM : la bibliothèque centrale (à l'initiative du projet), le centre d'histoire des techniques (CDHT) et le CEDRIC. Des

experts d'autres institutions européennes y ont collaboré ponctuellement. L'intention du projet est d'extraire du fonds patrimonial une collection francophone représentative permettant d'aborder l'histoire des techniques dans toutes ses dimensions [2] :
- La constitution des savoirs, par le biais de monographies et d'articles de spécialistes en directions de leurs pairs,
- L'histoire des institutions : rapports d'activités, de commissions, de jurys, travaux des enseignants du CNAM,
- La médiation vers le public : ouvrages et revues d'enseignement populaire, de vulgarisation, récréations et romans scientifiques
Le site s'adresse aux enseignants et aux chercheurs, mais il est également un outil de vulgarisation.

Le service Web (http://cnum.cnam.fr) a été inauguré en janvier 2000, avec la mise en ligne d'une cinquantaine d'ouvrages relatifs à l'électricité au XVIIIe siècle [3]. Le calendrier des travaux est consultable sur le site, sous la rubrique "Information -> historique du projet -> archives". L'année 2001 a vu la numérisation d'un classique de l'histoire industrielle (les *Grandes Usines* de Julien Turgan, 1860-1885) et une série de monographies (16^{e}-19^{e} siècle) consacrées aux machines et à l'agriculture. Deux périodiques (les *Annales du CNAM* et *La Nature*) sont en cours de traitement en 2002, ainsi qu'un important dictionnaire (*Dictionnaire technologique*, 1822-1835). L'année 2003 sera consacrée essentiellement à la mise en ligne des rapports officiels des expositions nationales des "produits de l'industrie française" (1798-1849) et des expositions universelles tenues à Paris (1855-1900), dans lesquelles le CNAM fut très impliqué.

Ces ouvrages ont été édités avant le 20^{e} siècle et sont libres de droits. Signalons cependant une incursion dans le domaine contemporain avec un ouvrage d'informatique numérisé en accord avec les auteurs et l'éditeur (*Systèmes d'exploitations des ordinateurs*, du groupe Crocus, Dunod, 1978). L'expérience sera renouvellée en 2003 (J.-P. Meinadier. *Structure et fonctionnement des ordinateurs*, Larousse, 1977).

En janvier 2002, le site web proposait 80 ouvrages, soit 25 000 pages, et il recevait une moyenne de 1 500 visites[*] par mois. Le corpus mis en ligne est actuellement de 224 volumes intégralement numérisés, soit 92 850 pages. La consultation du site est globalement en progression et atteint environ 4 000 visites mensuelles ces derniers mois (fig. 1). Le projet CNUM est maintenant reconnu par notre établissement. Il bénéficie d'une ligne budgétaire, et à reçu une aide au titre de la recherche.

Constitution du fonds numérisé

La diversité des titres retenus nous a obligés à aborder de front toute la complexité de l'édition scientifique ancienne : lourd appareil éditorial, nombreuses illustrations, souvent en planches hors texte, paginations multiples et parfois défectueuses, ouvrages en plusieurs volumes, recueils factices, etc. Il a donc fallu dans un premier temps élaborer un modèle descriptif valide pour tous les cas de figure rencontrés. Ce modèle a ensuite permis de définir la structure des bases de données utilisées pour la

[*] On considère ici les requêtes sur la page d'accueil du site.

production des fac-similés et leur mise en ligne (ces bases de données sont les fac-similés, les notices bibliographiques, les tables des matières et les tables des illustrations).

L'option de numériser les ouvrages en mode image a été choisie (coût, fidélité à l'original). De plus, ce format semble pouvoir bientôt devenir compatible avec la recherche en plein texte (OWR : Optical Word Recognition). La diffusion des textes est ainsi fidèle à la mise en page et au contenu, mais sans normes particulières de taille qui permettraient de travailler précisément sur le matériel typographique.

Les ouvrages de formats 12°, 8° et 4° sont retenus, à l'exclusion des folios, pour permettre une lisibilité des pages compatible avec la taille des écrans actuels. Les titres proviennent tous du fonds de la bibliothèque du CNAM et doivent être numérisés sans être déreliés. Les premiers lots ont été numérisés en noir et blanc avec une résolution de 400 PPI par des prestataires extérieurs. A partir de 2001, les ouvrages ont été numérisés en 300, puis 400 PPI avec 256 niveaux de gris. Cette numérisation est destinée à l'archivage, au format TIFF-g4 pour les noir et blancs et JPEG (sans perte) et PNG pour les niveaux de gris. Le traitement des fac-similés en vue de la diffusion sera abordé plus loin. La numérisation en couleur est encore coûteuse en France. Nous avons choisi de repousser à plus tard le traitement de titres contenant des illustrations en couleur. Les prestataires assurent le découpage et le redressement des images captées, ainsi que leur nommage. Les lots d'images sont remis sur CD-ROM, DVD-ROM ou DAT.

Les notices bibliographiques, les tables des matières et tables des illustrations sont systématiquement saisies (en interne au début pour les tables, puis par un prestataire extérieur). Aucun essai de reconnaissance automatique (OCR) n'a à ce jour été mené. Aux tables existantes sont ajoutées les mentions préliminaires et les appendices (dédicaces, préface du traducteur ou de l'éditeur, privilèges, errata, etc.). Les tables sont établies avec l'aide d'un historien expert si elles n'existent pas (Cf Végèce,). Ces ajouts aux textes originaux apparaissent entre crochets carrés. L'orthographe des textes est fidèlement transcrite. Cependant, la reproduction fidèle des tables est très difficile, car celles-ci présentent très souvent un nombre important de niveaux de hiérarchie, et des normes typographiques très diverses. Pour permettre une plus grande lisibilité, les longues tables des matières sont présentées selon deux niveaux de hiérarchie, sous forme abrégée, puis sous forme intégrale. Ce travail de présentation devra être affiné.

L'existence et l'emplacement des tables des matières, des tables des illustrations et des index sont indiqués sur des tableaux descriptifs des ouvrages, joints au cahier des charges (cf. annexe 2). La mise en correspondance des informations issues des tables avec celles issues des fac-similés de pages est effectuée par programme au CEDRIC. C'est cette étape qui révèle le plus d'erreurs. La cause en est certainement le fait qu'elles soient saisies par deux groupes de travail distincts et changeants. Un cahier des charges a été établi pour y remédier : il s'affine au fil du temps. Les erreurs concernant les fac-similés eux-mêmes sont plus rares (fichier corrompu, lacune, mauvais découpage, par ex.) du fait du contrôle qualité effectué chez les prestataires. Un autre facteur souvent méconnu est le taux d'erreurs des tables originales : leur correction nécessite là encore l'intervention d'un expert. Malgré ces réserves, nous pensons que le "zéro-défaut" est un objectif à notre portée du fait de la petitesse des lots numérisés chaque année.

Traitement des fac-similés

Nous avons pris le parti de dissocier totalement le format des fac-similés diffusés de celui des numérisations. La faiblesse et surtout le caractère aléatoire du débit de l'Internet "au large" impose des contraintes assez strictes sur la taille des fichiers diffusés. Par ailleurs, la plupart des navigateurs Web ne gèrent par défaut que trois formats de fichier : GIF, PNG et JFIF (JPEG). Aucun des trois n'est véritablement adapté au type d'images que nous diffusons [4]. En particulier, la compression JPEG, a été conçue pour les tons continus et la couleur, et non pour des fac-similés de texte imprimé et de gravures au trait. Nous avons donc dégradé les images diffusées à 100 PPI (à peu près la résolution d'un écran) et limité le nombre de niveaux de gris. Pour le périodique *La Nature*, la dégradation a été limitée à 133 PPI, ce qui permet une meilleure lisibilité des illustrations et de leurs légendes. Ce choix devrait d'ailleurs être généralisé à l'ensemble des ouvrages édités au XIXe siècle (voir *"Les grandes usines")*.

La figure 2 récapitule les résultats d'un test effectué en 1999 sur un premier lot de 11087 images de doubles pages [5]. On y compare les tailles des fichiers d'origine (TIFF-g4 en 400 PPI) avec celles de fichiers dégradés en 100 PPI aux formats GIF et PNG en 16 niveaux de gris. On remarquera la forte variabilité des tailles de fichiers TIFF-g4, qui traduit l'inadéquation de ce type de compression pour les fac-similés de gravures. Les deux pics d'effectifs pour les formats GIF et PNG correspondent en revanche clairement aux deux type d'images traitées : fac-similés de texte et planches gravées. Le format GIF a finalement été préféré au PNG car il est le seul à permettre l'encodage des 8 niveaux de gris sur 3 bits. Les images résultantes ont ainsi une taille moyenne de l'ordre de 100 Ko, soit 10 à 20 fois moins que les originaux, mais sont parfois aux limites de l'acceptable en termes de lisibilité. Toujours pour des raisons de coût, ces traitements d'adaptation sont menés au CEDRIC. Il est évident que ces choix techniques devront être remis en cause dans le futur, selon l'ampleur des progrès en matière de débit de transmission et de capacité de stockage.

Organisation du site

Un site Web de bibliothèque numérisée a une structure assez spécifique, avec peu de profondeur et une hyper-textualisation relativement faible. L'immense majorité de ses pages sont des fac-similés des pages des ouvrages papier, et, du point de vue ergonomique, la navigation dans ces dernières doit être étudiée très finement. Les autres rubriques du site sont plus conventionnelles et incluent : une rubrique de présentation, une rubrique d'aide, un moteur de recherche et, bien sûr, un catalogue alphabétique des textes diffusés.

A la demande des historiens participant au Conservatoire numérique, une rubrique d"outils" a été ajoutée à cet ensemble : à l'occasion des travaux de sélection des titres sur les différents thèmes, sont élaborés introductions thématiques, bibliographies, biographies, renvois sur d'autres sites. Leur consultation est un complément d'information pour les lecteurs qui le désirent. Ces différentes rubriques sont accessibles à partir d'une barre d'outils placée à gauche du fac-similé.

L'usager accède aux ouvrages à partir du catalogue, par une page descriptive incluant la notice bibliographique et une table abrégée. Cette page est construite une fois pour toute lors de la mise à jour du serveur, et peut être référencée par des sites externes (moteurs de recherche, en particulier). Il est également possible d'accéder aux ouvrages par l'intermédiaire du moteur de recherche. Dans la dernière version du CNUM, cette recherche peut se restreindre à un seul ouvrage isolé (périodique ou monographie en plusieurs volumes). Une amélioration doit cependant être apportée à ce moteur de recherche : l'appauvrissement des textes de requête (actuellement, les signes diacritiques sont pris en compte dans les requêtes).

La navigation dans les fac-similés se fait par un "feuilletage" séquentiel à partir d'une barre d'outils qui s'ouvre sur la droite de la page. L'utilisateur peut aussi accéder au numéro de page de son choix, naviguer exclusivement dans les illustrations, utiliser le zoom. La mise en regard de plusieurs pages est une fonction standard des navigateurs Web : elle trouve son plein intérêt ici (fig. 3). Le téléchargement intégral d'un volume (au format PDF) est actuellement en cours d'étude.

Les traitements de mise à jour du site (traitement des lots d'images, mise en forme des bases textuelles, création des pages HTML) s'effectuent sur un poste de travail de type macintosh G4 en Perl sous MacOSX. Le site lui-même est hébergé sur un micro-ordinateur de type PC sous Linux : une configuration jugée suffisante pour la diffusion de quelques milliers de volumes et quelques centaines de visites par jour. La charge imposée au serveur est en pratique assez faible et sa disponibilité très bonne (en regard aux incidents réseau, notamment).

Conclusion

Nous voudrions souligner pour conclure l'importance d'un partenariat entre bibliothécaires, informaticiens et historiens des techniques. Celui-ci aide en effet à aborder les problèmes techniques tant bibliothéconomiques qu'informatiques, en travaillant au service de textes dont l'organisation et le contenu sont maîtrisés.

On peut constater par ailleurs l'intérêt d'une coordination entre sites, telle que celle-ci, mais aussi telle que les portails (Ministère de la culture, signets de la BNF, CCRTI, catalogue critique des ressources textuelles sur Internet. CNRS). Il est en effet important de rassembler et de faire connaître les sites institutionnels existants.

Références bibliographiques

[1] LESK, M., *Practical digital libraries - Books, bytes & bucks*, Morgan Kaufmann, 1996
[2] DEBLOCK, G., ROZET, B., CUBAUD, P., " Le Conservatoire numérique des arts et métiers : une création partenariale ". *Bull. des Bibliothèques de France,* vol. 46 (4), 2001
[3] BLONDEL, C. " L'électricité et le magnétisme au XVIIIe siècle à travers la bibliothèque virtuelle du CNUM ", *Annales historiques de la révolution française*, 2000, n°2

[4] SALOMON, D. *Data compression, the complete reference*, Springer Verl., 1997
[5] CUBAUD, P., TOPOL, A., " A WWW-based digital library for antiquarian collections ". *Rapport de recherche du CEDRIC RR99-09*. Mars 1999
[6] CUBAUD, P., GIRARD, D., " ABU : une bibliothèque numérique et son public ". *Documents numériques,* vol. 2(3-4), 1998

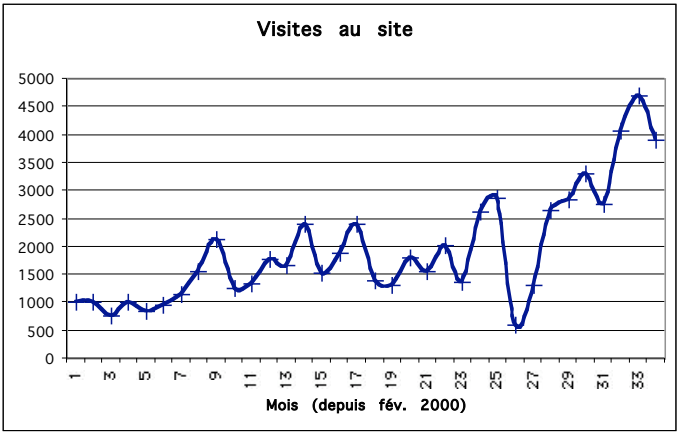

Figure 1. Nombre de visites mensuelles au site CNUM

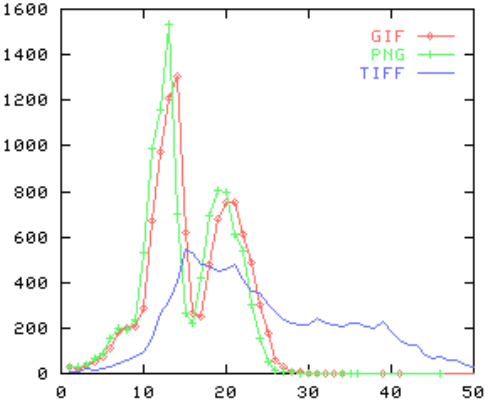

Figure 2. Histogramme des tailles de fichiers (unité = 10 Ko)

Figure 3. Session de travail

Annexe I - Différentes phases d'une campagne de numérisation pour le CNUM:

1- Constitution d'une instance de validation scientifique
2- Recherche et choix des documents à numériser
3- Définition et règlement des problèmes juridiques posés
4- Evaluation des tâches et des coûts de numérisation
5- Préparation des documents et du train de numérisation

6- Comptage des pages, des illustrations, des différentes tables (description très précise). Localisation des manques, et recherche d'autres exemplaires pour compléter
7- Etablissement des notices bibliographiques (catalogage, dépouillement très fin de chaque numéro de périodique ; indexation RAMEAU)
8- Elaboration d'un cahier des charges, comportant en particulier les recommandations correspondant aux spécificités rencontrées lors de la préparation du train de numérisation
9- Choix de numérisation pour les différentes parties (mode, format), comptage
10- Prise de contact avec les éventuels partenaires, définition des instructions de traitement des images, et établissement d'une demande de devis.
11- Choix du prestataire

12- Saisie et balisage des tables des auteurs et des légendes des illustrations (CNAM ou prestataire extérieur)
13- Numérisation du texte en mode image (prestataire extérieur)
14- Traitements informatiques : conception et réalisation des interfaces de consultation
15- Production de métadonnées

16- Traitement des images (dégradation, etc)
17- Mise à jour des pages statiques (accueil, catalogue,…)
18- Relecture
19- Validation du prototype
20- Dépôt sur le serveur d'exploitation

Annexe 2 - Numérisation : nommage des fichiers

Deux sortes de fichiers sont préparés : les lots de fac-similés d'une part, les métadonnées d'autre part. L'unité de base est le volume. La gestion de ces fichiers s'effectue à partir d'un "identifiant", qui correspond à la cote de l'ouvrage, en capitales, sans espace :

> Exemples :8XAE11 pour 8° Xae 11, 4XAE593.3 pour le 3e volume de 4° Xae 593, 4XAE48_2 pour le deuxième ouvrage d'un recueil factice, etc.

1) Fac-similés.

Les fichiers sont groupés au sein d'un répertoire nommé par l'identifiant du volume. Chaque fichier image du volume est nommé selon la syntaxe suivante : N.T.X.ext

N : numéro d'ordre du fichier	T : type de l'image	X : numéro de page ou de planche	ext : extension
0001 à 9999 La 1e image du volume est 0001.	V = page de titre (en principe, 1 seule fois) T = texte P = planche hors texte B = revers blanc	Non numérotée : 0 Caractères romains : r1 pour I, r2 pour II, etc. Pagination multiple	tif jpg gif png

La numérotation est réinitialisée à chaque volume.	D = page double, ne pouvant pas être coupée. Ce cas entraine une rupture dans la numérotation. C = pl. à pivoter en format paysage L = signalement d'une lacune	: 2x1, 2x2 … 3x1, 3x2, etc. ou bien 2xr1 ; 2xr2… Ouvrages foliotés : 1r ; 1v ; 2r ; 2v… ou bien r1r ; r1v	

2) Métadonnées. Les fichiers sont nommés, en accord avec le répertoire des fichiers images correspondants :

Identifiant.desc : pour les notices bibliographiques
Identifiant.tmi : pour les tables des matières
Identifiant.tpi : pour les tables des illustrations

Exemples : 4XAE69.1.desc 4XAE69.1.tmi 4XAE69.1.tpi

ID#8SAR12
AU#LACEPEDE. Bernard-Germain-Etienne de La Ville, comte de
AS#***
TI#Essai sur l'éctricité naturelle et artificielle / par M. le Comte de La Cepède
ED#Paris : de l'Imprimerie de Monsieur, 1781
DE#2 tomes en 1 vol. (vi-375-[1] + [4]-389-[3] p.) ; 8°
CO#CNAM 8° SAR 12
SU#Electricité*histoire*18e siècle
TO#2
NI#381 394
PP#0 0
PT#4 3
DP#0 0
NU#12/1998