Les documents numériques : numérisation et diffusion des documents

Jason Frodot

Département Sciences de l'Information et de la Documentation

Jason FRODOT

MASTER 1, MENTION ICCD
(Option : Sciences de l'Information et du Document)

MEMOIRE DE STAGE
Mission effectuée du 2 juin au 11 juillet 2014

À
L'ANRT (Atelier national de reproduction de thèses)
Villeneuve d'Ascq

LES DOCUMENTS NUMÉRIQUES : NUMéRISATION ET DIFFUSION DES DOCUMENTS

Sous la direction de :
Mr Joachim Schopfel (tuteur universitaire)
Mr Rachid Berbache (tuteur professionnel)

Soutenu le 11 septembre 2014 l'UFR DECCID-SID
Université Charles de Gaulle, Lille 3 (Campus Pont de Bois)
BP 60 149, 59 653 Villeneuve d'Ascq Cedex

Année Universitaire 2013/2014

Remerciements

Ce mémoire s'est fait dans le cadre de l'Anrt (Atelier national de reproduction de thèses), un service qui s'occupe du traitement des thèses et de leurs exportations vers différentes bibliothèques pour satisfaire d'éventuelles demandes.

Je tiens d'abord à remercier ma famille, qui m'a aidé à chercher un stage en documentation ou dans les archives, recherche que j'ai réalisé dans de nombreux endroits et pendant au moins 2 mois, et qui par la suite m'a encouragé durant mon stage.

Je remercie également le personnel de l'Anrt qui a accepté de m'engager en tant que stagiaire. Parmi eux, je tiens à remercier :

- Joachim Schopfel, qui a accepté de m'engager comme stagiaire et qui m'a conseillé des ouvrages dans le cadre de mon mémoire, à savoir « Manuel de numérisation » et « Manuel de constitution de bibliothèques numériques » (tous deux étant sous la direction de Thierry Claerr et Isabelle Westeel) ;

- Rachid Berbache, qui m'a fait visiter les lieux et montrer comment exporter des informations à partir d'un site ;

- Fredérique Dhooghe, qui me montra comment numériser, puis vérifier les thèses et enfin les exporter, et enfin qui m'a bien aidée face aux problèmes de numérisation ;

-Maryse Noncle qui avec Joachim Schopfel et Rachid Berbache, a accepté de m'engager ;

- enfin Ludovic Custodio, qui m'aida à résoudre les problèmes matériels et informatiques.

Résumé

Dans un contexte de développement accru de l'Internet et des demandes de la part des utilisateurs, la numérisation et la diffusion de l'information se doivent de répondre aux défis engendrés par ces développements. Pour cela, il convient au documentaliste et à l'archiviste de scanner l'information contenue dans le document papier à travers un scanner, utiliser des formats adaptés à la numérisation et à l'exportation des données, vérifier si les couleurs ne sont pas déséquilibrées au niveau de la disposition, ou si possible les supprimer si celles-ci devaient gêner la numérisation, faire attention à d'éventuelles erreurs de codage qui pourraient nuire à la lecture de l'information ainsi qu'à sa protection, et enfin à exporter ces informations à travers différents sites sous différents formats, cela bien entendu avec l'autorisation de l'auteur du texte contenants ces informations dans le cadre de la loi.
Par conséquent, la numérisation et la diffusion de l'information sont concernées par de multiples enjeux comme par exemple les enjeux stratégiques, juridiques et informatiques. Enjeux qui sont eux-mêmes liés à de multiples problèmes comme celle de la protection informatiques des données à travers le codage et le choix des formats, celle de la protection des droits d'auteur au moyen des lois les encadrant et enfin celle des liens par partenariat avec différents sites de diffusion de documents numérisés (comme par exemple la diffusion de thèses avec l'ABES (Agence bibliographique de l'enseignement supérieur)), ainsi que diverses bibliothèques (comme les bibliothèques universitaires) et centres de documentation .

Mots-clés :

Bibliothèques, codage, couleur, diffusion, documentation, droit d'auteur, format, information, Internet, numérisation

Tables des matières

Liste des abréviations

ABES : Agence bibliographique de l'enseignement supérieur
AIIM : Association for Information and Image Management
ANR : Agence national de la recherche
ANRT: Atelier national de reproduction de thèses
BMP : Bitmap
BnF : Bibliothèque nationale de France
CCD: Charge-Couple Device
CCITT : Comité consultatif international téléphonique et télégraphique
CFC : Centre français d'exploitation du droit de copie
CIE : « Commission Internationale de l'éclairage »
Cm : Centimètre
CMOS : Complementary Metal Oxide Semiconductor
CNRS: Centre national de la recherche scientifique
CPI : Code de la propriété intellectuelle
CSV :Comma-separated values
DADVSI : [Loi relative aux] droits d'auteur et droits voisins dans la société de l'information
DCT (de l'anglais*Discrete Cosine Transform*) : Transformée en cosinus discrète
Dpi (de l'anglais *Dot Per Inch*) : Point par pouce
DTD : Document Type Definition
EAD : Encoded Archival Description
EXIF: Exchangeable image file format
FAX G4 : Facsimile Group 4
Fk : foreign key
GIF (de l'anglais : *Graphic Interchange Format*): Format d'échange d'images
HDR: High Dynamic Range
HTML : Hypertext Markup Language
ICC: International Color Consortium
id : identifier
IPG: International Programmers Guild
IPTC: International Press Telecommunications Council
ISBN : International Standard Book Number
ISO (de l'anglais: *International Organization for Standardization):* Organisation internationale de normalisation
JBIG: Joint Bi-level Image experts Group
JFIF : JPEG File Interchange Format
JPEG : Joint Photographic Experts Group
Lab (ou l*a*b): « l » pour la clarté, « a » pour la gamme de 600 niveaux sur l'axe rouge-vert, « b » pour la gamme de 600 niveau sur l'axe jaune-bleu
LED: Light Emitting Diode
LZW: Lempel-Ziv-Welch
MARC : MAchine-Readable Cataloging

MESR: Ministère de l'Enseignement supérieur et de la Recherche
MLD : Modèle logique de données relationnelles
Mm : Millimètre
Mpix : Megapixels
Mo : Méga-octet
Nm : Nanomètre
NNT: Numéro national de thèse
NCSA (de l'anglais *National Center for Supercomputing Applications*) : Centre national pour les applications des super-ordinateurs
OCLC : Online Computer Library Center
OCR (de l'anglais *optical character recognition*) : Reconnaissance optique de caractères
PDF : Portable Document Format
Pk : primary key
PNG : Portable Network Graphics
Px : Pixel
RGB : Red, Green, Blue (voir signification pour RVB)
RLE : Run-Lenght-Encoding
RVB : Rouge, vert, bleu
STIC: Sciences et technologie de l'information et de la communication
TICE: Technologies de l'information et de la communication pour l'enseignement
TIFF : Tagged Image File Format
URL : Uniform Resource Locator
USB: Universal Serial Bus
UTF-8 : Universal Character Set Transformation Format - 8 bits
XML: Extensible Markup Language
XMP : Extensible Metadata Platform

Introduction

Lors de ma recherche de stage, qui a été très longue, il eut d'abord été difficile de savoir quel thème abordé, à savoir si celui-ci est quelque chose dans la documentation, dans les bibliothèques ou les archives, jusqu'à ce que l'on me propose un poste à l'ANRT, situé à l'université Lille 3 Charles-de-Gaulle. L'ANRT est un atelier crée en 1971 dans le but de reproduire les thèses de doctorats par impression, mise en microfiche (depuis 1983) et numérisation à la demande, et dont il existait avant la réunification le 1er janvier 2011 deux ateliers, selon la discipline (1):

- le premier rattaché à l'université Lille-III pour les lettres, sciences politiques, juridiques, et humaines et sociales ;
- le second rattaché à l'université Grenoble-II pour les sciences exactes, la médecine, la pharmacie, les sciences économiques et de gestion.

Lorsque j'effectuais mon stage, mon principal travail consistait d'abord à numériser des thèses et les formulaires qui en sont rattachés au moyen d'un scanner et d'un logiciel informatique (Capture Perfect), tout en tenant compte de la taille de ces thèses, du type d'impression initiale (recto, recto/verso ou en ignorant les pages blanches), du fait qu'ils sont en couleur ou non (dans le cas présent, il faut retirer les infiltrations et l'arrière-plan) et enfin de la présence ou de l'absence d'agrafes reliant certaines pages; puis à vérifier ces thèses sur un autre logiciel (PixEdit 7), dont les tâches consistaient à vérifier s'il ne manque pas une ou plusieurs pages sur une ou plusieurs thèses, puis à supprimer les couleurs de fond de certaines pages n'ayant pas pu être retirées (dans ce cas la suppression des pages de la thèse numérisée puis l'enregistrement séparé de la thèse corrigée s'avèrent nécessaires); et enfin à exporter les fichiers vers un autre poste ayant une autre tâche que celle de la numérisation des documents.
Par la suite, une autre tâche m'a été confiée, à savoir de mettre en place un document (par exemple sur Bloc-Notes) dans lequel on doit d'abord donner le numéro NNT (qui équivaut en fait à l'identifiant de la bibliothèque universitaire ciblée) puis, après l'ajout d'un point virgule, l'URL de la thèse ciblée. Le tout en tapant le numéro ANRT puis copiant les informations depuis un site (ici l'ANRT). Cela donne, pour la plupart des thèses, le résultat suivant avec l'exemple ci-dessous:

1986PA100194; http://www.diffusiontheses.fr/3837-thèse-de-larroque-michel.html

Ce document doit avoir pour titre le mois puis l'année de la liste des thèses ciblées (ex: Mars 2014).
De ces deux types de tâches, qui n'ont en apparence aucun point commun entre elles, elles ont pourtant comme lien le fait d'être inclus dans une chaîne dans laquelle s'entremêlent numérisation du document, partenariat avec diverses organisations comme les bibliothèques, les droits d'auteurs avec les autorisations des auteurs d'un document, choix du format et aspect

(1) http://fr.wikipedia.org/wiki/Atelier_national_de_reproduction_des_thèses

physique des documents. Une chaîne dans lequel semble inclus un contexte de dématérialisation du document, autrement dit du passage du format papier au format numérique, ainsi qu'une demande toujours plus croissante des demandes en matière d'information et de documents, auxquelles les bibliothèques, les centres de documentation et les archives se doivent de répondre. Or compte tenu de l'évolution croissante des documents, des informations, et des demandes toujours plus immédiates, ainsi que des difficultés de gérer cet ensemble aussi bien dans les domaines économiques que structurels et matériels, les organisations se doivent de réaliser des partenariats avec d'autres (par exemple le partenariat entre l'ANRT, l'ABES et le MESR) dans le but d'assurer plus efficacement la diffusion des documents. De ce fait, les tâches effectuées dans une organisation semblent en réalité sont pour chacunes d'entre elles des rouages dans un rouage, lui-même inclus dans un ensemble plus grand. Un ensemble qui peut être mis à mal si l'une des tâches venait à manquer ou si des problèmes venaient à apparaître dans l'une des tâches concernés. Il convient par conséquent d'établir la problématique suivante: en quoi les tâches de la documentation, en particulier la numérisation, sont-elles utiles au bon fonctionnement de cet ensemble ?

En effet, pour faire face à ce contexte d'augmentation du nombre d'informations, de documents et de demandes, les bibliothécaires, les documentalistes et les archivistes (alors métiers à documents) se doivent de répondre le plus rapidement possible et aux plus de demande possibles. Pour cela, ces métiers se doivent non seulement établir des partenariats entre organisations, mais aussi disposer du matériel adapté. Ainsi vient le scanner, qui numérise le document non numérique, et l'ordinateur, qui non seulement fait exécuter la tâche de numérisation, mais aussi adapte ce document selon le format adapté à ce dernier (ex: TIFF ou JPEG) tout en corrigeant d'éventuelles erreurs (comme les couleurs indésirables par exemple), le tout grâce à des logiciels de numérisation et de correction de documents. Or tout support qui se respecte souffre de divers inconvénients. En effet, si le document numérique se caractérise par son immédiateté et par son ubiquité, il peut également souffrir de problèmes propre à l'informatique comme les erreurs de codage ou des formats inadaptés, qui peuvent entraîner la perte des informations propre à un document ou l'illisibilité de ces derniers. D'autre part les techniques de numérisation ne sont pas uniformes en raison d'une part de la multiplicité des formes de document (comme les livres et les microfilms), et d'autre part parce que certains documents peuvent se révéler fragiles (ex: les livres rares en raison de leurs anciennetés dans la plupart des cas). Sans compter du fait que la numérisation est elle-même soumise à de nombreuses contraintes juridiques, notamment en ce qui concerne les droits d'auteurs et le délai de validité pour une reproduction. Ces inconvénients peuvent de ce fait nuire à la bonne diffusion des documents.

Diffusion des documents à laquelle la numérisation semble occuper, à travers ses avantages et ses inconvénients, une place centrale dans la diffusion des documents entre les organisations. Ce qui amène à la grande problématique: comment expliquer cette importance centrale qu'occupe la numérisation ?

Pour tenter de répondre à cette problématique, la démarche entreprise dans le stage effectué consiste à voir comment numériser une thèse à travers les paramètres et l'observation des problèmes eventuelles, à la corriger si besoin, à l'exporter sur d'autres postes, et enfin à récupérer les informations depuis un site. Le tout couplé à une observation de l'entreprise et à la récupération d'informations des pages de sites et des livres susceptibles d'aider dans la résolution de la problématique.

En ce qui concerne le plan, nous nous intéresserons d'abord sur le pourquoi de la numérisation sous ses différents aspects, puis à ses techniques, et enfin à la diffusion des documents pour finalement établir une conclusion.

Chapitre 1 : Pourquoi numériser

Avant de s'intéresser aux pourquoi de la numérisation, il convient d'abord de définir ce qu'est un document numérique pour mieux les analyser.

1) Définition

Un document numérique est difficile à définir dans la mesure où il comporte plusieurs définitions, variables en fonction des sources (ci-dessous une liste non exhaustive):

- Selon la définition de la norme ISO 19005-1 : *« Représentation numérique, selon un format de page, d'une agrégation de données textuelles et graphiques, et de métadonnées utiles à l'identification, la compréhension, et le rendu des données, qui peut être reproduite sur un papier ou une microforme optique sans perte significative de son contenu en information. »*

- Roger T. Pédauque , du STIC CNRS et auteur de "Document: forme, signe et médium, les reformulations du numérique", en propose trois :

+ Définition 1: « *Un document numérique est un ensemble de données organisées selon une structure stable associée à des règles de mise en forme permettant une lisibilité partagée entre son concepteur et ses lecteurs.* »(1) [6]

+ Définition 2: « *Un document numérique est un texte dont les éléments sont potentiellement analysables par un système de connaissance en vue de son exploitation par un lecteur compétent.* » (1) [6]

+ Définition 3: *Un document numérique est la trace de relations sociales reconstruite par les dispositifs informatiques. (1)* [6]

- Selon Wikipédia, un document numérique est *« une forme de représentation de l'information consultable à l'écran d'un appareil électronique. L'affichage de ce type de document peut être apparenté soit au « document » même, ou soit à l'interface logicielle. Suivant l'intervention d'applications informatiques dans une partie de son contenu (bases de données, POO), les changements dans l'organisation logique de ses données peuvent être apportés [...]. »* (2)

---

(1) **PÉDAUQUE, Roger**. *Document: forme, signe et médium, les reformulations du numérique.* 8 juillet 2003. (consultation le 27 juillet 2012),<http://archivesic.ccsd.cnrs.fr/docs/00/06/21/99/PDF/sic_00000511.pdf>
(2) http://fr.wikipedia.org/wiki/Document_numérique

Quoi qu'il en soit, un document numérique est quelque chose qui se rapporte à la fois à l'informatique (avec les termes "données", "métadonnées" et "interface"), au social ("relation social") et à ce qu'il contient ("texte", "textuelle","structure" et "contenu"). Quelque chose qui partage avec le document non numérique (dont il faut noter également la difficulté de le définir) le fait qu'il dispose d'un contenant, d'un contenu, des informations, et le fait qu'il nécessite au moins un émetteur et un récepteur. Un document numérique peut prendre plusieurs catégories:

- les documents édités : livres électroniques, e-books, revues scientifiques électroniques; (1)[2]

- les documents non édités: archives ouvertes et institutionnelles, production pédagogique, jeux de données de recherches et données administratives. (1)[2]

Des catégories de documents qui sont, dans tous les cas et pour au moins une partie d'entre elles, des versions électroniques de documents sous formes papier (livres, thèses, feuilles) de plaques de verre, de microfilms ou de microfiches (2) (comme c'est le cas à l'ANRT), tous de forme, de taille, de résistance physique, de durée de conservation (en particulier les microfiches) et de coûts variables. Ces documents sont divisés selon les mêmes catégories que les documents numériques, à savoir :

- parmi les documents édités (1) [2]:

+ livres, périodiques et presse,
+ revues et ouvrages scientifiques et pédagogiques,
+ documents patrimoniaux (imprimés, estampes, cartes et plans imprimés, partitions, cartes postales, etc.);

- parmi les documents non édités (1) [2]:

+ documents patrimoniaux (archives, manuscrits, dessins, etc.),
+ documents iconographiques (négatifs, plaques de verre, Ektachromes, etc.),
+ archives institutionnelles et informations administratives,
+ archives de la recherche (archives, manuscrits, photographies, cartes et plans, etc.),
+ documents pédagogiques (cours polycopiés, préparations de cours, notes, etc.).

Ainsi à travres ces listes des documents édités et non édités, on peut en déduire plusieurs enjeux, à savoir d'abord les enjeux culturels, avec la diffusion des oeuvres d'art et d'esprit dans le cadre des patrimoines culturelles, les enjeux scientifiques, avec de nouveaux corpus de textes susceptibles d'être mise à jour, et les enjeux économiques, si l'on se réfère au fait qu'une

(1) **Sous la direction de Thierry Claerr et Isabelle Westeel.** *Manuel de la numérisation.* Editions du Cercle de la librairie. Paris: Electre, 2011. 320 pages. ISBN : 978-2-7654-0983-0. (Page 19)

(2) A noter que l'ANRT dispose de deux types de microfiches pour chaque thèse, de composition et de durée de conservation différentes

numérisation implique un budget de la part du service.

2) Le contexte

Contexte général

Lorsque la numérisation commença au début des années 1990 dans les bibliothèques et les archives (avec les exemples des Archives nationales et des archives départementales de la Mayenne en 1993), elle avait pour but de préserver le document, bien que la qualité des documents numérisés de l'époque (qui étaient d'abord réalisés sur des microfiches) n'était pas forcément bonne. Avec l'émergence de la mise en ligne, la qualité de l'image ainsi que les fonctionnalités de recherche s'améliorèrent, au point que le texte de l'oeuvre numérisé et l'exemplaire numérisé supplante désormais la consultation physique des documents, cela grâce à l'OCR et aux réseaux de télécommunication.
Ainsi dans les services, les opérations de numérisation, le nombre de documents à numériser et par conséquent le public commencèrent à s'accroître considérablement dès le début des années 2000, à un point tel que le public virtuel devenait l'un des enjeux majeurs des services des bibliothèques et des archives. C'est alors que depuis 2005, date à laquelle on assiste au lancement du programme de numérisation do moteur américain Google, que la décision d'augmenter la quantité de documents numériques mis à la disposition du public, bien que tous les documents ne sont pas numérisés car non inclus dans le cahier des charges (par exemple les formats non standard, les tableaux de chiffres et les dépliants). Ce cahier des charges fait des choix en fonction de ses propres critères, puis selon les exigences de conservation, ensuite d'après des critères intellectuelles et enfin dans le cadre de la chaîne de numérisation, où on sélectionne les documents à numériser selon leur état de conservation, leur caractère précieux et selon leur conformité aux critères de normalisation de traitement industriel. Chaîne de numérisation dans laquelle les chantiers de numérisation de masse, qui ont pour but principal d'adresser à une production industrielle un grand nombre d'ouvrages à numériser, ne peuvent être effectués que si l'institution dispose déjà d'un catalogue informatisé dans son fonds pour assurer la traçabilité des ouvrages entre l'institution et la chaîne de production.
Face à la masse des documents à numériser, les traitements et les indexations effectués manuellement deviennent de plus en plus inadaptés au contraire des traitements et des indexations effectués automatiquement, de manière à ce qu'il y ait un accès par le contenu et une diffusion qui exploite les bonnes méthodes de compression. Toutefois les traitements automatiques peuvent comporter des erreurs (ex : résolution trop faible) mais peuvent être évités lors de la rédaction du cahier des charges.
Le tout peut ainsi s'inscrire dans le cadre d'une révolution numérique, dans laquelle les technologies de numérisation et d'informatique amènent à de nouvelles pratiques et à de nouveaux usages, notamment en ce qui conçernent les documents récents. On peut citer par exemple la micro-informatique et les réseaux, comme Internet, qui se sont généralisés à un point tel que c'est la société qui a finit par les approprier, puis les faire évoluer, et cela de deux manières qui se complètent mutuellement:

- L'adaptation de la société avec l'environnement numérique, dans lequel non seulement les institutions culturelles proposent les documents numérisés, mais en plus conçoivent

des applications dont les lecteurs peuvent y participer à l'enrichissement des informations descriptives.
- L'adaptation du numérique en fonction des besoins, dans la mesure où l'usager peut créer sur la Toile un monde virtuel à sa convenance par l'intermédiaire d'outil permettant de personnaliser son environnement de consultation, tout en enrichissant les information à sa disposition.

Contexte en France

Si en France, on assiste à une évolution comparable aux autres pays européens en matière de numérisation, on assiste toutefois à l'émergence d'une perspective de constitution d'une bibliothèque numérique européenne, portée par le ministère de la Culture et de la Communication et mis en oeuvre par la BnF. Pour éviter la dispersion des initiatives de numérisation, on assiste à la mise en place au niveau national de catalogues collectifs des ressources numériques intégrées au portail européen Michael (portail multilingue qui rassemble les collections numérisées de différents pays européens). Ces catalogues sont au nombre de 2 :

- *Patrimoine numérique*

Mis en oeuvre par le ministère de la Culture et de la Communication (qui a son propre moteur de recherche (1)) et intégré en 2004 dans le projet européen Michael (2), il s'agit d'un catalogue national des fonds culturels numérisés qui recense les collections ayant fait l'objet d'une numérisation ainsi que les institutions qui y sont impliquées (cela qu'importe la source de finançement). En février 2011, il recense pas moins de 1714 collections et 595 institutions. Il valorise les initiatives de numérisation conduites par les institution culturelles (ex: musées et services d'archives) tout en offrant des modes de navigation pour accéder aux collections, comme l'accès géographique, le sujet, le type d'institution et le type de document.

- *NUMES* (3)

Développé par le MESR, géré actuellement par l'ABES et en cohérence avec le catalogue *Patrimoine numérique*, c'est un catalogue qui est destiné aux collections des institutions dépendant de la recherche et de l'enseignement supérieur. Il offre une visibilité sous l'angle national et international des activités de numérisation effectuées pour les besoins de l'enseignement universitaire, de la recherche et pour la préservation du patrimoine documentaire et scientifique.
En ce qui concerne la numérisation en elle-même, ses résultats dépendent des institutions qui le pratiquent. Dans le cas des bibliothèques des collectivités territoriales (qu'ils soient régionales, départementales ou communales), les collectivités participent à l'activité de numérisation des institutions culturelles, tout en coopérant avec les institutions majeurs comme la BnF, et cela jusqu'en avoir été promoteur. Ces opérations, même si elles ne sont financés que par les collectivités et sont modestes (hormis dans le cas de la Ville de Lyon, qui a effectué un

(1) http://collections.culture.fr
(2) http://www.numérique.culture.fr
(3) http://www.numes.fr

partenariat avec Google pour la numérisation et la diffusion de 500000 ouvrages anciens), sont approuvées et soutenues par le ministère de la Culture et de la Communication dans la mesure où elles font partie du Plan national de numérisation, qui à été lancé en 1996. Comparée à Gallica, qui fut inaugurée en 1997 et qui a effectuée une numérisation de masse d'un très grand nombre de documents (soit en février 2011 267540 ouvrages, 800000 fascicules de périodiques et 244000 images) en provenance de ses collections, des éditeurs associés ou des bibliothèques partenaires, les initiatives de ces collectivités sont modestes mais sont peu à peu intégrés dans le paysage numériques et répondent aux préoccupations des politiques culturelles des bibliothèques territoriales. Cela en mettant en place des bibliothèques de taille plus modeste grâce aux collectivités, aux professionnels et au projet culturel se rattachant, et on peut donner l'exemple de la Bibliothèque municipale de Baud qui numérise des cartes postales ayant pour thème la Bretagne et le monde maritime. Le résultat est que l'on dénombre 466 fonds numérisés par 175 bibliothèques (dont 135 municipales) (1), composés en partie de collections patrimoniales, de documents liées à l'histoire locale (comme les oeuvres d'écrivains se rattachant à une ville ou à une région) ou de documents rares, numérisés pour répondre à la diversité du public. Ces documents sont mise en ligne par des portails régionaux qui fédèrent parfois les collectivités locales.
Pour ce qui est du réseau des Archives de France, il englobe les Archives nationales (qui comprennent les Archives nationales, les Archives nationales d'Outre-mer et les Archives nationales du Monde du travail) et les archives territoriales (qui sont composées des archives départementales, des archives municipales et des archives régionales). Avec la décentralisation dans les années 1980, ces derniers sont intégrés les conseils généraux, les municipalités et les conseils régionaux, qui leurs procurent leurs moyens humains et financiers. Ainsi le nombre de projets de numérisation des documents d'archives et le nombre de fonds numérisés accessibles sur Internet augmentèrent, et on dénombre pas moins de 777 projets publiés par les services d'archives (dont 480 au niveau départemental) (1). Ces projets peuvent être divisés en plusieurs types:

- Les documents iconographiques, qui font parties des projets les plus nombreux (ex: estampes, affiches, cartes postales, etc.);
- Les plans cadastraux;
- L'état civil (ex: registres paroissiaux, registres d'état civil, tables décénnales, etc.), dont 2 à 3 millions de registres paroissiaux et d'état civil sont numérisés par départements;
- Les journaux locaux;
- Divers types de registres et de documents comme les recensements de population et les archives notariales.;
- Des documents "oraux" comme les enregistrements sonores et audiovisuels;
- Divers types de documents non classés (ex: les plans-terriers du 18 ème siècle de l'abbaye de Cluny et les cahiers des instituteurs pendant la Grande Guerre);

(1) Chiffres en provenance du catalogue *Patrimoine numérique*

A noter que les inventaires et les instruments de recherche, en particulier les répertoires imprimés et dactylographiés plus anciens et les plus rares, sont également numérisés. Pour cela, l'océrisation et l'encodage au format XML (suivant la DTD EAD) est nécéssaire pour la récupération d'un document électronique structuré, ce qui a d'ailleurs déjà été réalisé par les Archives nationales d'Outre-Mer et de certains services départementaux.
Dans le cas de l'enseignement supérieur et de la recherche (dont l'ANRT en est lié), les bibliothèques numériques sont peu nombreux et de taille modeste si l'on considère la taille des fonds patrimoniaux et leurs diversité, ce qui peut être cependant corrigé à terme par la réforme des universités, la multiplication des sources de financement et la recherche de notoriété. Par exemple on peut citer le cas de PôLiB, qui est la bibliothèque numérique patrimoniale des universités Lille 1, Lille 2 et Lille 3 qui l'ont reprises en 2008, alors qu'il s'agissait à l'origine d'un projet lancé en 2001 par le pôle universitaire Lille Nord-Pas-de-Calais. Elle a pour but de valoriser le patrimoine écrit des universités lilloises et les bibliothèques régionales partenaires dans le domaine de l'histoire des sciences.(1)
Ainsi la numérisation s'inscrit dans un contexte de multiplication des types de documents, du type de public, d'évolution de la technologie et de numérisation en masse depuis 2005. Ce qui implique par conséquent plusieurs aspects.

3) Ce qu'implique la numérisation

Pour qu'une numérisation se fasse correctement, il faut tenir compte de nombreux aspects qui l'entourent

Que faut-il numériser ?

Dans une numérisation, il faut aussi savoir ce que l'on veut numériser. En effet, il faut savoir quel type de document veut-on numériser. Autrement dit, les questions à se poser sont les suivantes: quel aspect du document doit être numérisé (document papier, document sonore....) ?, dans quel domaine (thèses de lettres, thèses de sciences....) ?, quel est le format initial du document à numériser (livres, thèses, brochures....) ?, dans quel thème (par exemple une région précise) ?, dans quel but (conservation, complétion d'une bibliothèque ou d'une collection....) ?, est-ce un texte, une image photographique ou bien sonore ?
Ainsi ce que l'on veut numériser détermine le matériel qui sera adapté à chaque type de document, comme les scanners à défilement pour les lots de feuilles volantes de thèses (nécessitant donc des moyens techniques), ainsi que le(s) format(s) de numérisation (TIFF ou JPEG), l'espace de numérisation, le nombre de documents à numériser, le personnel nécessaire (nécessitant donc des moyens humains) , le coût de la numérisation (nécessitant des moyens financiers), le temps nécéssaire à la numérisation, variable selon les documents et la méthode choisie (page par page ou par lot) et les partenariats avec différentes institutions (comme le partenariat ANRT-ABES) .

(1) http://polig.univ-lille3.fr

Les droits d'auteurs

Lorsque l'on parle de droit d'auteur, on parle de deux types de droits que peut disposer l'auteur d'une oeuvre: les droits patrimoniaux, qui sont des droits de reproduction et de communication au public dans lequel l'auteur peut autoriser ou interdire la reproduction et la communication de ses documents, et qui sont limités dans le temps; et les droits moraux, qui sont des droits qui permettent de préserver la paternité de l'auteur (qui a le droit de choisir l'anonymat ou un pseudonyme) tout en assurant l'intégrité de l'oeuvre, c'est à dire pas d'ajout, de suppression ou de modification de contenu sans l'autorisation de l'auteur.
Ainsi l'auteur a le droit de faire ce qu'il veut de son oeuvre, ce qui peut poser des questions conçernant les droits d'auteur et la propriété intellectuelle, à savoir: cette dernière est-elle pertinente ? Peut-on partager et protéger ? Si oui, que faut-il protéger exactement ? C'est pour cela que les licences Creatives Commons ((CC)) sont crées pour tenter de répondre à ces questions. Ils servent selon les cas à :

- protéger la paternité de l'oeuvre;
- empêcher d'éventuelles modifications effectuées sans l'autorisation de l'auteur
- empêcher la diffusion d'une oeuvre à des fins commerciales sans l'autorisation de l'auteur;
- partager une oeuvre à l'identique de l'original.

Notons aussi que l'article L.112-2 (1) du CPI donne une liste de création entrant dans la catégorie des oeuvres de l'esprit, comme les documents écrits (livres, brochures, et autres écrits littéraires, artistiques et scientifiques, comme par exemple les thèses), les photographies, les oeuvres d'art, les plans et cartes géographiques, etc (1). On peut inclure aussi comme oeuvres protégées les lettres missives, les bases de données (ce dernier pouvant bénéficier sous certaines conditions d'une protection "sui generis" , assuré par l'article L.342-1 du CPI) sur son contenu.), les index et les résumés.
Quelque soit la création celui qui peut autoriser la divulguation reste l'auteur (ou en cas de mort de ce dernier les ayant droits, ou à défaut le juge de tribunal de grande instance), sauf si celui-ci a cédé ses droits de reproduction et de représentation par exemple à un éditeur ou à une revue.
Toutefois on peut observer des cas où il existe plusieurs auteurs (comme c'est le cas de certaines thèses numérisées). Dans ce cas, il faut distinguer une oeuvre de collaboration, dans lequel plusieurs auteurs travaillent sur un projet commun et dont chacun d'entre eux peuvent bénéficier du droit d'auteur, d'une oeuvre collective dans laquelle l'oeuvre est crée sur l'initiative d'une personne physique ou morale (dont le nom est divulgué sur l'oeuvre), qui bénéficie ainsi du droit d'auteur dans son intégralité.
S'il existe bon nombre de situation dans lequel il faut tenir compte de l'autorisation de l'auteur (ce dont tient compte l'ANRT), il existe d'autres cas où l'autorisation de l'auteur n'est pas nécéssaire, parmi lesquelles :

(1) A noter que les oeuvres citées ci-dessus sont des exemples parmi une liste de 14 types d'oeuvres inclus dans l'article en question

- Le fait que les établissements dépositaire du droit légal peuvent reproduire les documents de quelque manière que ce soit dans des buts de collecte, de conservation et de consultation surplace, cela à des fins non commerciales, choses permises par les nouvelles dispositions de la loi DADVSI (article L.132-4 du Code du patrimoine (1)).
- Le fait que les bibliothèques, les musées et les services d'archives puissent reproduire des documents pour les mêmes raisons évoquées précédemment, et cela dans des buts non économiques et non commerciales, bien qu'il n'existe pas de définition précise d'une bibliothèque et que les critères et les contraintes sont parfois flous.
- Le fait qu'un document est reproduit pour répondre à des situations de handicap telles que le handicap physique, sensoriel, mental, cognitive et psychique, dans des formats et des supports adaptés. Pour cela, les établissements devront effectuer une demande de dépôt des fichiers numériques conçernant une oeuvre à la BnF ou au Centre national du livre, à condition que cela s'effectue deux ans après le dépôt légal de l'oeuvre imprimé et que cela ne porte pas préjudice aux intérêts légitimes de l'auteur.
- Les courtes citations, qui peuvent se faire sous certaines conditions : qu'elles aient un caractères critique, polémique, scientifique ou d'information; que le nom de l'auteur et la source du document soient indiqués; et qu'elles soient brèves.

L'analyse du public

Pour savoir comment numériser, on se doit de savoir d'abord quel public est ciblé, et par conséquent quels usages sont réalisés par ce public. Pour cela, Isabelle Westeel et Jean-François Moufflet en définissent trois types (2) [2]:
- Le grand public, qui est *"en quête, la plupart du temps, d'informations généralistes, de thématiques larges, pour avoir un aperçu, de façon pragmatique, de ce qui existe. Curieux, mais ne désirant pas s'embarasser de connaissance trop spécifiques ou techniques, il recherche avant tout la simplicité d'utilisation et d'accessibilité à l'information"* (2)[2]
- Le public "scolaire": dans le cas présent, la numérisation, qui est considéré comme un vecteur pédagogique très important, *"participe à l'ouverture des scolaires et des étudiants au monde extrascolaire, par l'accès aux multiples informations, dans un but d'apprentissage et de découverte, mais aussi d'approfondissement des connaissances, tout en restant ludiques et innovants. Dotée d'un potentiel pédagogique conséquent, elle participe également à un nouveau type de communication (forums, échanges entre écoles, relations entre élèves, partages entre enseignants) de nature collaborative."* (2)[2]

(1) "L'auteur ne peut interdire aux organismes dépositaires, pour l'application du présent titre:
1° la consultation de l'oeuvre sur place par des chercheurs dûment accrédités par chaque organisme dépositaire sur des postes individuels de consultation dont l'usage est exclusivement réservé à ces chercheurs;
2° la reproduction d'une oeuvre sur tout support et par tout procédé, lorsque cette reproduction est nécessaire à la collecte, à la conservation et à la consultation sur place dans les conditions prévues au 1°."
(2) Sous la direction de Thierry Claerr et Isabelle Westeel. *Manuel de la numérisation.* Editions du Cercle de la librairie. Paris: Electre, 2011. 320 pages. ISBN : 978-2-7654-0983-0. (Page 87)

- Le public spécialisé: ici, "*la numérisation est stratégique pour tout les domaines de la recherche*" (1)[2], dans la mesure où elle peut rendre acessible de l'information rare et précieuse. Selon les mêmes auteurs, "*elle facilite et rend plus efficace la recherche et l'exploration des liens par les bibliothécaire, les étudiants, les enseignants, les érudits, les chargés de cours à l'université, les chercheurs, etc. car elle permet d'étudier les documents et des travaux disparates dans des contextes nouveaux.*"(1) [2]. Il s'agit ici du public principalement conçerné par les universités, et donc du public ciblé par l'ANRT.

Le coût

Il s'agit de l'un des aspects des plus importants pour une numérisation. En effet, le coût d'une numérisation dépend du budget du service de numérisation. Le budget, qui représente environ environ deux tiers des coûts totaux, dans lequel il tenir en compte le calendrier budgétaire et les obligations en terme de paiements et de marchés. Le budget tient en compte :

- La préparation et la restauration des documents, dont il faut tenir compte à la fois de la volumétrie des documents, de l'état des documents (voir si le document n'est pas trop dégradé, que ce soit par la numérisation ou autre chose) et par conséquent de son état sanitaire. Si le document est trop dégradé, des opérations de restauration pourraient être effectuées. La vérification, la restauration et la préparation matérielle nécéssite une mobilisation d'équipes de personel, et donc un budget;
- Le signalement des documents par numérisation qui, associé à un catalogage précis des documents, est réalisé selon les normes en vigueur tout en permettant l'accès au document numérisé et l'interopérabilité;
- Le déplacement du document, qui peut se faire au frais du prestataire selon les termes du marché passé avec ce dernier. Si on veut réduire les frais, on peut soit assurer soi-même le transport, soit sous-traiter le document (par le commanditaire ou par le prestataire).
- La numérisation elle-même, qui dépend à la fois du temps passé à reproduire la page, du nombre de vues, et du coût de la page. Le tout dépend de la technique utilisée : pour la numérisation patrimoniale, qui consiste à scanner lentement et manuellement des documents précieux, fragiles ou abimés, le coût est de 0,60 à 3 euros la page; quant à la numérisation par page, qui dépend du nombre de pages de l'ouvrage relié et par conséquent du matériel, l'opération peut prendre plusieurs dizaines de minutes et coûter de 10 à 40 centimes la page; et dans le cas d'une numérisation de masse, qui s'effectue pour les documents à plat et des ouvrages massicotés par scanner bureautique, le coût est de 5 centimes par page. Le coût de la numérisation dépend aussi du passage du document numérisé à l'OCR: pour l'OCR brut, c'est 5 centimes par page; pour l'OCR acceptable, c'est 20 centimes la page et 15 euros le document; et pour le document HCR HQ, c'est 90 euros le document.

(1) **Sous la direction de Thierry Claerr et Isabelle Westeel.** *Manuel de la numérisation.* Editions du Cercle de la librairie. Paris: Electre, 2011. 320 pages. ISBN : 978-2-7654-0983-0. (Page 87)

- L'achat du matériel, qui inclut le scanner, diverses appareils de photographie numérique, les ordinateur, le disque dur, un écran de qualité professionnelle et des licences de logiciels de traitement et de retouches d'images (pour la vérification de la qualité et la gestion de la photographie numérique). Le coût sera élevé si ces matériaux sont achetés en grand nombre;
- Les supports de livraison: pour livrer des données au commanditaire, il faut utiliser des supports amovibles, comme par exemple les disques durs qui permettent une livraison et un chargement des données sur un système d'information de manière rapide. Quelque soit le support, le prestataire facturera son achat.
- La mise en ligne, qui en devient une étape presque obligée pour la diffusion, et dont les coûts liées à celles-ci doivent être prévue (pour l'élaboration du site Internet, l'achat et l'installation des logiciels);
- La conservation des données, qui dépend des supports amovibles et de la surveillance régulière des supports de stockage au moyen d'un contrôle par échantillon de la lisibilité des données et d'une planification des opération de migration (dans le cas de supports vieux ou de mauvaise qualité). Dans le cas des stockage sur des supports en ligne, des opérations de réplication sont nécessaires pour effectuer une copie de sécurité et sauvegarder des données;
- Les ressources humaines, qui consistent à recruter des vacataires pour travailler sur le conditionnement, la mise en ordre et l'indexation des documents. A noter que le coût humain représente environ la moitié du budget.

Le tout dépend du respect du Code des marchés publics (décret n°2006-975 du 1er août 2006), dont l'application permet "le respect des principes de libertés d'accès à la commande publique, d'égalité de traitement des candidats et de transparence des procédures."(1). Ce qui sous-entend au moins trois obligations : définir un besoin, respecter des obligations de publicité et de mise en concurrence et choisir l'offre la plus avantageuse d'un point de vue économique. A noter qu'il existe des seuils de marché dans lequel sont définis les modalités de publicité et de mise en concurrence. On peut remarquer que plus le montant du marché est élevé, plus la publicité et la publication dans un certain type de presse deviennent obligatoires et plus la mise en concurrence sera obligatoire et fera l'objet d'une procédure formalisée. (2)[2]

La tenue d'un cahier des charges

Pour assurer le bon déroulement de la numérisation, une bonne tenue d'un cahier des charges sera nécéssaire. Ce cahier des charges est plus précisément "un cahier des clauses techniques particulières", qui est inclus dans un dossier de marché public et qui dispose de clauses ayant pour but de déterminer la préparation, l'exécution de la numérisation et la diffusion des documents. On établira ici une liste non exhaustive de ces clauses (3) [2]:

---

(1) Article 1 du Code des marchés

(2) **Sous la direction de Thierry Claerr et Isabelle Westeel.** *Manuel de la numérisation.* Editions du Cercle de la librairie. Paris: Electre, 2011. 320 pages. ISBN : 978-2-7654-0983-0. (Tableaux des pages 112-113)

(3) **Sous la direction de Thierry Claerr et Isabelle Westeel.** *Manuel de la numérisation.* Editions du Cercle de la librairie. Paris: Electre, 2011. 320 pages. ISBN : 978-2-7654-0983-0. (Tableaux pages 127 à 134)

- Le contexte et les objectifs du marché:

Il s'agit de déterminer la nature et l'objectif du marché, à savoir :
+ l'identité et les missions du commanditaire;
+ les types de documents concernés;
+ le résultat attendu, l'exploitation prévue et les usages envisagés;
+ le périmètre de la prestation;
+ les relations avec d'autres prestataires pour d'éventuelles tâches différentes;

- Le résumé des prestations à réaliser:

Il s'agit d'énumérer les futures tâches réalisées par le prestataire par ordre chronologique, ou sinon indiquer les prestations couvertes par un autre prestataire.

- La présentation des documents à numériser:

C'est la description détaillée des documents qui feront l'objet d'une numérisation.

- La décomposition du marché en sous-ensemble:

Il s'agit de préciser le découpage d'une prestation en unités, donnant lieu à des livraisons par le prestataire et correspondant à des envois échelonnés de documents ou à des tâches distincts à effectuer.

- Le déroulement et les délais d'exécution :

C'est la description de l'enchaînement chronologique de l'envoi et des tâches, ainsi que des délais de réalisation et de contrôle prévus pour chacune des tâches.

- Les conditions générales d'exécution:

Il faut préciser les règles de base qui seront suivies durant la prestation et qui guideront la relation prestataire/commanditaire.

- La mise à disposition des documents:

Description de la préparation du document (ex: préparation du format) et de la forme du bordereau les accompagnant (informations, format)

- La prise en charge, le transport et le stockage des documents:

Au niveau de ces aspects, il faut définir les rôles et les responsabilités du prestataire et du commanditaire, ainsi que les contraintes de chacune des tâches.

- Les caractéristiques des fichiers à fournir:

Il s'agit ici de détailler les caractéristiques techniques des fichiers images (ex: résolution, colorimétrie....), préciser le format texte désiré après océrisation, expliciter les règles de nommage de fichier et préciser le format des métadonnées s'y associant.

- Le droit d'auteur :

C'est la précision selon laquelle le prestataire ne dispose pas de droit d'auteur sur les images numériques produites.

4) Les objectifs

Avant de définir la façon de numériser, il convient d'abord d'en connaître les objectifs poursuivis, qui doivent d'abord répondre à certains critères. Pour cela, il existe un moyen mnémotechnique pour connaître ces critères, à savoir l'acronyme SMART (intelligent en anglais), dont les significations sont :

- « S » pour Spécifique (*Specific*)
- « M » pour Mesurable (*Measurable*)
- « A » pour Atteignable (*Achievable*)
- « R » pour Réaliste (*Realistic*)
- « T » pour Temporellement défini (*Time-bound*)

Autrement dit les objectifs poursuivis, ainsi que chacun de ces critères se doivent d'être le plus proche de la réalité possible, qu'il soit matériellement, économiquement, temporellement et humainement. Isabelle Westeel (archiviste paléographe, conservateur des bibliothèques, titulaire d'un DEA d'Histoire et un des directeurs de l'ouvrage « Manuel de la numérisation ») et Jean-François Moufflet (archiviste paléographe, conservateur du patrimoine et doctorant en histoire médiévale) proposent par ailleurs des listes d'objectifs selon deux ordres (chacun d'entre eux concernant un ordre et directement issue de l'ouvrage « Manuel de la numérisation ») :

Au niveau patrimonial :

Ces objectifs sont :

- favoriser l'accès aux documents anciens et fragiles. Pour cela cet objectif doit se faire par la reconstitution virtuelle des collections, la comparaison des exemplaires géographiquement distants, la valorisation et la facilitation de l'accès à l'information par l'offre de nouveaux modes de consultation pour le public ;

- faire connaître des documents non communicables ou n'ayant jamais été communiqués en raison de leur états physique (fragilité des pages de documents, documents tendant à se dégrader lors de certaines conditions définis comme la température et l'humidité, en particulier dans le cas des archives). Dans ce cas, les documents apparaissent comme des substituts pour pallier ces défauts ;

- faire connaître des documents dont l'accès au contenu informatif est difficile, comme par exemple les manuscrits enluminés et les marques de lecture. Pour cela, la numérisation permet de faire ressortir des détails jusqu'alors inaccessible à l'œil humain ;

- préserver le contenu informatif des documents condamnés à une dégradation irrémédiable du fait de la fragilité du support ;

- transférer dans un format numérique des supports qui sont en danger, empruntés ou malmenés pour mieux les préserver et réduire leur consultation, comme c'est le cas des plaques de verre, des films sur Celluloïd ou de la presse ;

- valoriser le patrimoine local, régional, national et international. Dans ce cas, chaque institution doit s'interroger sur ce qui peut être facteur d'identité et de communauté à travers les documents conservés. En effet, les documents requiérant le plus grand intérêt ne sont pas forcément les plus esthétiques, les plus prestigieux ou les plus anciens.

A travers ces objectifs au niveau du patrimoine, on peut remarquer que la numérisation peut se faire dans un but de conservation et de communication des documents, et dans une moindre mesure un but politique et culturel.

Au niveau documentaire et scientifique :

Les principaux objectifs au niveau de ces points de vue sont :

- la diffusion des documents à un grand nombre de personnes grâce aux nouveaux modes de consultation et de communication ;

- la participation à la dynamique actuelle de la dissémination de l'information, le soutien à la création de contenus et de produits multimédias, tout en favorisant la maîtrise des technologies nouvelles de l'information ;

- la mise en œuvre des outils et produits numériques dans le cadre des TICE (Technologies de l'Information et de la Communication pour l'Education) ;

- l'inscription du projet de numérisation dans la politique culturelle du territoire, en partenariat avec d'autres structures de conservation, et sur des thématiques partagées ;

- offrir au public intéressé la possibilité de consulter une riche collection régionale ou locale ;

- la permission de l'accès à des collections thématiques numérisées en lien avec la politique documentaire de l'établissement ;

- la participation à des programmes de coopération et de concertation et donc la numérisation par exemple d'imprimés complémentaires du programme de numérisation de masse mené par la BnF ou des programmes menés par les autres établissements culturels (archives, bibliothèques ou musées). Dans le cas des bibliothèques universitaires, les documents numérisés sont pour la plupart des thèses de doctorats ou des mémoires de Master ;

- l'insertion dans des programmes coopératifs nationaux et internationaux, comme le projet européen Europeana Regia.

(Ce dernier est un programme de numérisation lancé en 2010 ayant pour but de reconstituer une bibliothèque virtuelle de documents royaux européens du Moyen-âge à la Renaissance. Elle associe pour cela cinq grandes bibliothèques européennes) ;

- permettre à l'établissement de participer à des projets de recherche, comme les projets Corpus lancés par l'Agence national de la recherche (ANR) ;

- la mise en ligne de la production imprimée des chercheurs ;

- la mise en ligne des fonds d'archives d'une grande richesse pour un usage pédagogique ou un usage de recherche, personnelle ou scientifique ;

- veiller à répondre aux besoins des chercheurs en corpus, en reconstitution virtuelle de collections ou de bibliothèques dispersées ;

- enfin, la mise en place d'un projet pilote pour tester les performances techniques de telles technologies, en particulier les projets innovants.

Ici les buts induits par la liste d'objectifs au niveau documentaire et scientifique concernent d'une part les buts scientifiques, et d'autre part dans des buts techniques, culturels, scientifiques et de coopération entre les différentes organisations.
Une fois les objectifs ciblés et le cahier des charges réalisé, il s'agit maintenant de passer à la numérisation.

CHAPITRE 2 : La numérisation

Durant le stage effectué, la principale activité consistait à numériser des thèses à l'aide d'un logiciel (CapturePerfect 3.0) et d'un scanner relié à celui-ci, après quoi il fallait ensuite éventuellement retirer les couleurs des pages de certaines thèses, vérifier si des pages ne sont pas manquantes et finalement exporter ces thèses. Pour comprendre comment numériser un document, nous nous intéresserons d'abord à certains de ses aspects.

1) L'image numérique

Une image numérique dispose de nombreuses représentations, par exemple des matrices de pixels pour le mode « point », dans lequel nous nous intéresserons vu qu'il s'agit de la seule représentation en numérique, ou même des représentations tel un ensemble de vecteurs ou un ensemble géométrique. Dans la première représentation on inclut non seulement la forme physique du document, mais aussi chacun des pixels, les contrastes lumineux, les couleurs, la dimension et le codage. Toutefois, il convient d'abord d'analyser les différents types d'images numériques.

1.1) Les types d'images numériques

L'image matricielle

Appelée aussi « image bitmap », l'image matricielle désigne une image constituée d'une sorte de tableau (appelé ici matrice) dans laquelle chaque case, appelée ici point, possède sa propre couleur (voir illustration 1 de l'annexe). Par conséquent, ce type d'image est une juxtaposition de points formant un tout qui se doit d'être le plus cohérent possible (tel une image). Dans le cadre de la numérisation, une image matricielle se fait dans deux espaces :

- le premier est l'espace spatial dans lequel l'image est numérisée en suivant le principe d'échantillonnage, dans lequel l'axe des abscisses et celui des ordonnées se rejoignent pour déterminer l'emplacement de l'échantillon (appelé ici pixel) dans l'image. Le nombre de ces échantillons déterminera la définition de l'image.
- Le second est l'espace des couleurs dont la quantification implique que « les différentes valeurs de luminosité que peut prendre un pixel sont numérisés pour représenter sa couleur et son intensité ». Sa précision implique de la justesse de la profondeur de l'image, qui dépend du nombre de bits dans lesquels la luminosité est codée

Par conséquent, la qualité de l'image dépend à la fois du nombre de pixels et de la quantité d'information contenue dans ces pixels. En ce qui concerne la définition, on peut remarquer que plus le nombre de pixels est élevé, plus le nombre de détails fins visibles le sera également, et que par conséquent plus la qualité de l'image sera bonne (ex : une image de 1280x1024 px sera de meilleur qualité qu'une image de 640x480 px).Or il arrive parfois que l'image perde de sa qualité

lors de la compression des données, destinées à réduire la taille des images stockées sur un disque, en raison du fait que lors de la compression des données, des informations peuvent se perdre car étant non compatibles avec la technique de compression choisie (on parle alors de technique de compression destructives).
Des informations qui peuvent aussi concerner la couleur du pixel, qui deviendra de ce fait différente de celui de l'image d'origine. Ainsi quant on parle de couleur, on parle aussi de « palette », dans laquelle chaque pixel est associé au rang qu'occupe sa couleur (Ex : les images GIF, qui dispose d'une tablette de 256 couleurs). Dans le cas des images délivrées par des logiciels de dessins, elles disposent lors de la compression d'un triplet RVB, dont le canal est codé sur 8 bits, ayant 256 niveaux.
Pour stocker et échanger ce type d'image, il est nécessaire de compresser et de stocker une image selon un format graphique, sous formats BMP, GIF, TIFF, PNG et JPEG.
Il s'agit du type d'image numérique utilisé lors de la numérisation des thèses et de la décoloration des pages couleurs durant le stage, plus précisément lorsque l'on zoome une page couleur d'une thèse.

L'image vectorielle

Il s'agit d'une *« image numérique composée d'objets géométriques individuels (segments de droite, polygones, arcs de cercle, etc) définis chacun par divers attributs de forme, de position, de couleurs, etc »* Chaque objet géométrique est représenté par une formule mathématique (par exemple une courbe est représenté par plusieurs point et une équation), ce qui permet à celui-ci de subir des transformation sans perte d'information (cela contrairement à l'image matricielle) et avec l'utilisation d'un faible nombre d'information grâce à l'ordinateur qui recalcule les données selon l'échelle demandée. De plus, l'image vectorielle est fortement liée à sa représentation, qui induit que plusieurs couches de dessin peuvent être superposées avec chaque point qui peut écraser un autre point. (Voir illustration 2 de l'annexe)
Ce qui permet donc d'une part d'avoir un fichier très peu volumineux tout en ayant des tracés fin et précis, mais qui d'autre part a l'inconvénient de s'appliquer uniquement aux formes simples, ce qui fait qu'une image réaliste telle une photo ne peut être rendue vectoriellement.

1.2) L'échantillonnage

Avant de parler d'échantillonnage, il convient de définir la définition d'une image. Une définition d'image correspond « au nombre de cellules disponibles pour la représenter dans la matrice de pixels », s'exprimant le plus souvent en millions de pixels et de mégapixels. La définition s'applique au niveau des écrans de télévision et des appareils photos numériques.
A cette notion peut s'adjoindre celle de la résolution d'échantillonnage, dont les mesures se font en points/cm et points/pouce (ex : pour 300 dpi, il faut 300 pixels représenter 2,54 cm en hauteur et en largeur). Une notion qui n'est pas applicable sur un support analogique (comme les originaux ou un support photographique), ni sur une image de synthèse en mode point crée sur ordinateur (comme les images de jeux vidéo), mais dans le cadre de processus de numérisation et d'impression sur un support réel (ex : photo, pages de livres ...).

La résolution peut s'exprimer linéaire et au « carré ». Dans le dernier cas, si on utilise le double du nombre de carreaux au mètre pour le carrelage par rapport à ce qui était prévu, le devis du carreleur sera au quadruple et non au double, ce qui implique par conséquent qu'il faudra 4 fois plus d'espaces dans un disque.
Le tableau 1 (voir annexes) indique la définition de l'image obtenue lors de la numérisation d'un support en fonction du format de ce dernier ainsi que de la résolution d'échantillonnage adoptée.

1.3) Codage

Pour la représentation des couleurs, il existe quatre options disponibles correspondant à des approximations de la réalité, au prix d'une réduction du poids de l'image, le tout dépendant de la nature du support :

Les images en couleur vraie :

Les images en couleur vraie s'appuient essentiellement sur toutes les couleurs visibles par l'œil humain, c'est-à-dire toutes les couleurs synthétisées par la combinaison des trois lumières monochromatiques rouge, vert et bleu (RVB, ou RGB en anglais) qui sont des couleurs primaires. Couleurs primaires qui seront les bases de la représentation informatique des couleurs lumineuses par trois valeurs numériques, codées pour chacun d'entre eux par un nombre entier, cela au moyen de la combinaison des canaux R, V et B.
Pour le codage, il faut un certain nombre de bits appelé profondeur (« *bit depth* »), qui sera pour chaque canal de 8 à 16 bits (les conversions numériques des valeurs analogiques mesurées par les capteurs des scanners s'effectuant sur 12 ou 14 bits). Soit un codage de la couleur sur 24 ou 48 bits, ce qui permet de représenter au minimum 16 millions de nuances de couleurs (dans le cas des canaux à 8 bits) et par conséquent une perte de précision par rapport aux valeurs mesurées par les capteurs ; et avec 16 bits par canal 200000 milliards de nuances (pourtant insuffisant pour l'imagerie de synthèse par les images « HDR », qui nécessite au minimum 32 bits par canal).
Cependant la capacité de discrimination des teintes par l'œil humain n'est pas uniforme dans l'espace de couleur RVB, ce qui engendre deux inconvénients : le premier est que cela engendre un certain « gaspillage » dans la mesure où on ne peut tirer un avantage d'un nombre de bits élevé que dans certaines régions de la gamme où la capacité discriminatoire est élevée ; le second est que même si des variations des valeurs R, V et B sont identiques, cela n'entraînera pas forcément les mêmes écarts des couleurs perçues dans les régions de l'espace des couleurs. Raisons pour lesquelles d'autres espaces de couleurs perceptuellement uniforme, comme l'espace CIE LAB ont été définis puis utilisés pour la mesure des écarts de couleurs pour se rapprocher le plus possible de la vision humaine.
A noter que cette option est utilisée par l'ANRT dans le cadre de la numérisation et du codage, cela contrairement aux autres options de la représentation des couleurs, que nous présenterons malgré tout dans la mesure où les logiciels de numérisation de l'ANRT proposent quand même ces représentations.

Images en couleur palettisée

Dans cette option, on utilise un nombre réduit de couleurs qui constitueront une « palette » (à peu près comme la palette d'un peintre), pour approximer les couleurs des pixels afin de réduire le poids des images (et donc la taille du fichier), tout en les conservant en couleurs. Dans ce cas, c'est la position sur la tablette et non la couleur qui sera associée à la valeur associée à chaque pixel et par conséquent, le nombre de bits utilisé le codage influe sur le nombre de teintes (ex : pour 8 bits, on a une palette de 256 teintes), ce qui permet entre autre de réduire le poids de l'image en couleur vraie, celui-ci hors compression. Il existe deux manière d'approximer les couleurs réelles d'une image, que l'on peut réaliser grâce à des algorithmes :

- soit en fixant une palette de taille et de contenu donné, puis en recherchant pour le pixel ciblé la couleur de la palette qui se rapproche le plus de la valeur du pixel ;
- soit en fixant la taille maximale de la palette avant de rechercher le contenu qui fournira la meilleure approximation de l'ensemble des couleurs. Dans ce dernier cas les algorithmes seront plus complexes que le cas précédent.

Images en niveau de gris

Or on peut aussi ignorer la couleur des supports numérisés pour mettre les image en noir, blanc et gris (comme les anciennes photos) et donc réduire davantage le poids du fichier en utilisant la notion de luminance. Celle-ci pondère l'intensité lumineuse contenue par la sensibilité chromatique de la vision humaine, et en traduisant en des nuances de gris (de la plus claire en se rapprochant du blanc, à la plus foncée en se rapprochant du noir). Pour cela, une formule relie la luminance (Y) aux valeurs R, V et B pour se rapprocher le plus possible à la sensibilité humaine aux couleurs de base :

$$Y = 0{,}212\ R + 0{,}715\ V + 0{,}072\ B$$

Comme on peut le remarquer sur la formule, les coefficients les plus forts se situent notamment au niveau du vert et c'est d'ailleurs pour cette raison que les capteurs des appareils photos numériques contiennent majoritairement des photosites verts. Quoi qu'il en soit, la somme des coefficients est de 1 et l'effet produit est que la valeur de la luminance varie selon le même intervalle que les valeurs R, V et B. Par conséquent, on aura plusieurs niveaux de gris, qui peuvent s'exprimer sur 8 bits (pour 256 niveaux de gris) ou 16 bits (pour 65536 niveaux de gris), bien que l'œil humain ne puisse distinguer qu'environ 200 nuances de gris, en grande partie des tons sombres. C'est d'ailleurs pour cela que par rapport à un gris de référence, un gris d'une luminance inférieure à 18% sera considérée comme très sombre, ce qui peut poser problème au niveau de la résolution. Pourtant on peut le résoudre de deux manières :

- en adoptant un codage sur 16 bits, ce qui aura pour conséquence le fait que la majorité des couleurs claires ne sera plus distinguable par l'œil humain (d'où une dimension de « gaspillage de bits »), bien qu'en contrepartie les informations ne seront pas perdues au niveau des tons sombres. Ce codage permettra d'effectuer plusieurs manipulations logicielles de la courbe réponse tonale sans risquer l'accumulation des erreurs d'arrondis, ainsi que nu-

mériser pour la préservation d'originaux photographiques, en particulier ceux en noir et blanc.
- en adoptant un codage de 8 bits tout en mettant en place une pondération non linéaire des valeurs de luminance simulant la sensibilité de la vision humaine.

Images bitonales

Dans cette option, l'information minimale par pixel est une information « tout ou rien », que l'on peut traduire par « noir ou blanc ». Ici l'image ne nécessite qu'un seul bit par pixel. Sa production implique un traitement numérique d'une image (en gris ou couleur vraie) par « binarisation » (ou seuillage) au niveau du scanner ou lors du post-traitement. Si par exemple une image est en niveau de gris, la binarisation fera en sorte que les pixels de luminance seront codés en blanc ou en noir.
Il existe plusieurs méthodes de seuillage « adaptatives » qui peuvent faire varier le seuil d'une page à l'autre en fonction des caractéristiques locales de contraste. Ces méthodes se retrouvent surtout dans des logiciels de post-traitement, plus précis que les logiciels de scanner.

1.4) Les méthodes de compression

Lorsque l'on parle de compression d'image, on parle non seulement de coûts de stockage, mais aussi des résolutions de capteurs, des projets de numérisation, des mathématiques appliquées et de l'algorithmique. Dans ces deux derniers domaines, les méthodes ont progressées depuis 50 ans grâce aux télécommunications et sont désormais transposées de manière complexe dans le domaine de l'image.
On observe au moins 2 types de compression d'images.

Les compressions sans perte

Dans cette méthode, qui est par définition réversible, on utilise des algorithmes de codage optimal. Dans ce type de codage, on considère qu'une page A (ici en tant que page blanche) a un poids inférieur à celui d'une page B (qui comporte ici 5 lignes), elle-même inférieure à celui d'une page C (qui aura ici 40 pages), et cela après compression.
Ces algorithmes peuvent aussi bien s'appliquer à un fichier texte (ex : codage de Huffman, Run-Lengh-Encoding (RLE)) qu'à un fichier image (ex : RLE, LZW, Deflate). Dans ce dernier cas, il existe des méthodes qui s'appliquent essentiellement pour un certains type d'image comme par exemple les méthodes FAX G4 et JBIG 2 pour les images bitonales. Quoi qu'il en soit, leur principal point commun est la recherche d'élimination de la redondance, qui s'exprime de différentes manières :

- La RLE, approche la plus simple, consiste à factoriser la description des lignes de pixels identiques. Par exemple on écrit « blanc (100) » au lieu d'écrire 100 fois « blanc ».

- La méthode FAX 64 consiste à coder chaque ligne de pixels de manière différentielle par rapport à la ligne suivante. Cependant cette méthode a pour principal inconvénient le fait qu'en présence de pixels isolés sur le fond de page, elle perd de son efficacité. Ce qui fait que cette méthode sera peu à peu remplacée par la méthode JBIG, qui offre un taux de compression supérieur.
- Quant à la méthode JBIG2, qui remplacera celle de JBIG, on intègre une « heuristique de segmentation de page en zones de texte, zones tramées et « autres » ». Pour cela, un dictionnaire empirique de « formes générique » modélisera les zones de redondance dans les zones de texte et déterminera qu'en cas de nouvelles formes isolées si on doit encoder par différence avec une forme déjà référencée dans le dictionnaire empirique, ou si la forme générique est nouvelle, le tout sans avoir à associer les formes dans un alphabet existant. Le logiciel PDF Compressor, dont le but est de recompresser les documents PDF images, est une implémentation de JBIG2 associée à JPEG 2000.

Les compressions à perte

Dans le cas où toutes les ressources du codage optimal sont épuisées pour compresser une image et réduire son poids, on peut toutefois abandonner une partie de l'information contenu dans l'image, à condition que cette perte ne soit pas visible à l'œil humain, alors imparfait, ou alors de manière minime. Il existe ainsi des méthodes de compression à perte conçues pour les images numériques, tout en s'inspirant des constats sur l'œil humain qui sont :

- La sensibilité aux contrastes diminue avec la fréquence spatiale des variations, jusqu'à ce qu'on ne distingue plus qu'un gris uniforme : la limite peut être de 7 paires de ligne par millimètre pour une acuité visuelle de 10/10 pour une distance de vision de 25 cm.
- Avec la fréquence spatiale, la sensibilité aux variations de teinte diminue.

Pour cela on effectue une analyse fréquentielle des fluctuations spatiales de la luminance et de la teinte, tout en les négligeant dès un certain seuil déterminé par des coefficients d'atténuation croissant avec la fréquence. Ainsi plus les seuils de filtrage sont abaissés, plus le poids décroît avec cependant en contrepartie une perte d'information de plus en plus perceptible, en raison d'une part de la compression de la gamme de tons de 16 à 8 bits par canal, et d'autre part, que les altérations et les pertes de l'information interviennent à tous les niveaux. Ces pertes interviennent que ce soit au niveau du capteur à travers la discrétisation du signal, au niveau de l'électronique à travers la conversion analogique/digitale du signal, au niveau du processus de quantification associé et au niveau des traitements d'images intégrées au scanner à travers le dématriçage. Ce qui donne à cette méthode un qualificatif de « méthode destructive », en particulier chez les non-spécialistes.

2) Les formats de fichiers

Il existe de nombreux formats dans lesquels la numérisation peut être effectuée. Lors du stage effectué à l'ANRT, les formats utilisés principalement sont les formats TIFF, JPEG et PDF, dont

les deux premiers sont d'une part les objets d'un enregistrement direct lors de la numérisation de la page, d'autre part utilisés pour la sauvegarde en cas de rectification de formats de fichiers sur le logiciel PixEdit7 (comme la vérification de pages manquantes et la suppression des couleurs de certaines pages), nécessitant d'abord pour cela la suppression du fichier contenant ces pages depuis le logiciel CapturePerfect. Ces formats s'avèrent être les formats les plus utilisés dans le cadre de la numérisation. Il convient donc d'en faire une présentation de ces formats.

Le format TIFF

Il s'agit du format le plus ancien, dans la mesure où il a été développé en 1986 par la société Aldus. Il est très utilisé pour les documents mono-pages, les photothèques, ainsi que pour la sauvegarde des archives parce d'une part, il conserve la gamme complète des couleurs, cela peu importe le mode colorimétrique et le niveau d'encodage (1 à 14 bits par couche), et d'autre part parce qu'il s'agit du seul format multipage, vu qu'il peut contenir une séquence de page, bien que la taille limite pour un fichier soit de 4 Go.
Il dispose d'un en-tête riche comportant des métadonnées (comme EXIF, IPTC et XMP) repérés par des « tags » qui peuvent être extensibles, et peut supporter les images bitonales, les niveaux de gris 8 ou 16 bits, la palette de 2 à 256, et la couleur RVB 24 ou 48 bits. Selon le type de codage, les méthodes de compression sont les suivantes :

- Absence de compression
- Pour les images bitonales, utilisation des moteurs CCITT groupe 3 (G3) et CCITT groupe 4 (G4)
- Pour toutes les images, compression des fichiers par l'algorithme "LZW" ou "ZIP" (ou Deflate) sans dégradation de la qualité des couleurs ni des niveaux de gris, ainsi que les méthodes packbits (RLE) ou, dans le cas des compression avec perte, l'algorithme JPEG (niveau paramétrable au niveau de la qualité).

Les versions récentes du format TIFF, (la dernière étant la version TIFF 6.0, datant de 1992) ont une option pages multiples exploitable par les systèmes de reproduction, photocopieurs numériques, mais pas par les applications courantes.
Ce format est applicable pour les images en noir et blanc, en particulier lorsqu'il s'agit de pages contenant uniquement du texte.

Le format PDF

Crée en 1993 par la société Adobe en tant que format ouvert et ayant pour dernière version la version 1.7 (2006), le format PDF a été par la suite adopté en tant que norme ISO (la dernière version ayant pour référence la référence ISO 32000-1 :2008). Il est utilisé pour les documents multi-pages ou nécessitant une protection contre certaines utilisations, et a été conçu à l'origine comme un format de description de pages, ces dernières étant à la fois destinées à l'impression et à être affichable à l'écran pour des contrôles préalables. Il ne s'agit pas à proprement parler d'un format de compression mais un format de diffusion qui est de plus en plus utilisé par les utilisateurs.

Bien que n'étant pas à proprement parler un format d'images mais un format de documents, le format PDF peut toutefois contenir des images avec ou sans pertes de données, selon le moteur de compression qui sera choisi dans la boîte de dialogue. Il est capable comme le TIFF d'enregistrer les modes colorimétriques bitonal, niveau de gris et couleurs, dans les encodages supportés par le TIFF, tout comme des images vectorielles au format .eps (*encapsulated Postscript*) issues des logiciels de création vectorielle ou des images multicouches (bichromies) pour la prépresse, ainsi que des pages composées contenant du texte, des images pixellisées et des images vectorielles.
Toutefois il ne pourra être intégrée à un document électronique que si l'application hôtesse est munie des traducteurs (filtres d'importation) adéquats. Les applications bureautiques ne peuvent intégrer ces fichiers sans une conversion préalable dans une application de traitement d'image pixellisée, car ils sont dépourvus de fonctions de traitement des objets vectoriels PostScript, alors que ce dernier est justement la base du format dans la mesure où le PDF est une amélioration de PostScript sur plusieurs points (notamment au niveau de la compression des images dans laquelle PostScript a un mode de codage plus limité que PDF et ne dispose pas de police pour les formats ni de structure indexée du fichier permettant un accès direct aux pages.
Par la suite se créent les documents « PDF image », qui sont des pages dans lesquelles il n'y a qu'une image plein format avec en second plan « un texte caché », et qui sont devenus plus commode à diffuser les documents avec l'OCR. Ce dernier est une technique basée sur un procédé optique « qui permet à un système informatique de lire et de stocker de façon automatique du texte dactylographié, imprimé ou manuscrit sans que l'on ait à retaper ce dernier »

Le format JPEG

Crée en 1991 par une trentaine d'experts, qui se sont inspirés des premiers travaux de recherche mené entre 1980 et 1990 sur la compression spectrale des images, le format **JPEG** est destiné à la diffusion d'images à l'aide un taux de compression variable, le plus souvent avec perte. Il a été conçu pour introduire des altérations minimales lors de la compression d'images présentant des variations de tons continues. Toutefois, en cas de forte compression, des artéfacts apparaissent en bordure des contrastes, preuve que son adaptation à la compression d'images ayant des contrastes nets n'est pas facile, d'autant plus difficile que le format présente des inconvénients, à savoir :

- Il est difficile de choisir la valeur optimale à appliquer dans la mesure où il n'existe pas de relation simple entre l'indice qualité et le taux de compression obtenu. En effet, les niveaux de qualité «maximale» n'altéreront pas visiblement la qualité de l'image, cependant les nuances de couleurs seront réduites, ce qui peut être gênant pour le traitement ultérieur de l'image en retouche ou interprétation artistique (travail des couleurs, ou agrandissement par rééchantillonnage). Quant aux niveaux de qualité «supérieure», ils commencent à altérer l'image si on fait un agrandissement: il apparaît

(1) définition du site http://www.futura-sciences.com/magazines/hightech/infos/dico/d/informatique-ocr-3953/)

alors des «plaques» de même couleur très gênantes dans les dégradés (ciels) et dans les reflets sur les objets lisses. Les niveaux de qualité «moyenne» accentuent les défauts précédents et font apparaître du «sable» à la frange des zones de contraste. Enfin, les niveaux de qualité «basse» font apparaître un genre de tissage de panier en accentuant les défauts cités précédemment.

- Il n'existe pas de définition officielle du taux de qualité JPEG dans la mesure où la norme JPEG ne fournit aucune indication à ce sujet, ce qui fait que chaque logiciel peut utiliser à sa convenance sa propre échelle. Toutefois, l'IPG a tenté de développer une implémentation JPEG de référence dans laquelle est proposée une définition arbitraire de la qualité : une valeur comprise entre 0 et 100 influant sur le calcul des 2x64 coefficients d'atténuation appliqués selon les fréquences.

Quoi qu'il en soit, la compression JPEG repose sur trois étapes :

- La première consiste en un changement d'espace de couleur depuis l'espace RVB vers l'espace YCbCr en modélisant une couleurs par la combinaison des trois composantes, cela au moyen de la séparation de la luminance (Y, représentant l'intensité lumineuse perçue) et de la chrominance (Cb et Cr, représentant la teinte selon sa position selon un axe rouge/vert et un axe jaune/bleu). Le tout pour préparer l'analyse fréquentielle.
- L'analyse fréquentielle, qui consiste à filtrer les signaux de chrominance ainsi que le signal de luminance pour ensuite filtrer progressivement les hautes fréquences spatiales. Pour cela, le choix des coefficients de filtrage aura une influence sur la quantité d'information que l'on choisira d'ignorer, ainsi que sur le taux de compression et la qualité résultante. Cette analyse nécessite donc une transformation numérique appelée DCT qui traitera l'image par bloc de 8x8 pixels, transformation qui permettra par la suite de voir si des artefacts sont visibles à fort taux de compression.
- Enfin, une compression de l'information résiduelle au moyen d'un algorithme qui permet une compression progressive en 10 à 12 niveaux selon les logiciels, avec une dégradation invisible (niveau 12) à importante (niveau 0). A noter également que ce format ne supporte pas le mode bitonal (bitmap).

D'après Anne Debant (Directrice du CRBM depuis 2011) et Patrick Perrot (rédacteur et infographiste en entreprise) (1)[5], il existe des extensions (plug-ins) qui peuvent être installées dans des logiciels de retouche d'image rendant lisse et sans «sable» ni « paniers» les compressions JPEG, dans une taille inférieure à la compression classique (par exemple: Boxtop pro JPEG de Boxtopsoft). Leur fenêtre de dialogue montre en détail la texture de l'image et l'utilisateur peut pousser son taux de compression en fonction du résultat. Cette méthode peut être appliquée pour la conservation de grands fonds iconographiques pour gagner un espace disque conséquent par rapport au TIFF ou à une compression à taux fixe prédéterminé. Cependant cette solution ne peut être mise en œuvre que par un opérateur très averti et distinguant correctement les couleurs.

(1) http://www.piaf-archives.org/espace-formation/course/view.php?id=11

Ce format concerne les pages couleurs ou grises des documents, en particulier lorsqu'il s'agit d'illustration pour illustrer par exemple des idées. Toutefois dans le cas des pages de couleur grise, la numérisation tend à les noircir, rendant illisible les information de la page en question si cette dernière contient du texte.

3) Comment acquérir l'image numérique

Comme expliqué précédemment, la tâche principale effectuée lors du stage était de numériser les thèses. Or il faut noter qu'avant la numérisation, on enlevait d'abord les reliures de ces dernières (on appelle cette opération « massicotage »), probablement pour numériser plus efficacement chacune des pages de la thèse, qui peuvent être soit uniquement en recto, soit uniquement en verso, soit les deux à la fois. Toutefois, certaines pages sont entièrement blanches, ce qui peut poser problème au niveau de la numérotation où une partie d'entre elles peuvent être incluses dans le nombre de pages, et cela même si le numéro de page n'est pas inclus dans les pages en question, d'où des hésitations pour les inclure réellement dans la thèse. L'autre raison est que le scanner utilisé est un scanner à « défilement », qui dispose d'un rouleau d'entraînement permettant de numériser un certain nombre de pages à la minute, cela selon un procédé alliant l'optique à l'électronique. Ce type de scannage nécessite pour cela que les pages soient détachées et disposées en lot.
Ainsi à travers cet exemple, on peut remarquer non seulement qu'il existe divers types de support qui peuvent être numérisés, mais aussi qu'il existe plusieurs types d'appareil de numérisation, et donc diverses méthodes de numérisation. Quelque soit le type de numérisation, celui-ci s'effectue toujours selon quatre aspects que l'on présentera ci-dessous Même si lors du stage, un seul type de document, un seul type de scanner et une seule méthode de numérisation a été utilisée, nous nous intéresserons parallèlement aux autres documents, scanners et méthodes. Toutefois étant donné que les supports sont extrêmement variés, nous nous intéresserons plus particulièrement aux documents papier.

3.1) L'aspect mécanique

Dans cet aspect, il existe deux sous-aspects qui permettent de numériser efficacement le document physique :

Le transport du document

Il s'agit ici de présenter au moins une face au numériseur en faisant défiler des feuilles volantes ou en tournant chacune des pages reliées. La manière de transporter les documents dépend donc du fait que les pages soient reliées ou non.
Dans le cas des documents non reliés (comme les pages d'une thèse, des formulaires ou des livres massicotés), le scanner prévu pour cela est un scanner à défilement (voir illustration 4 de l'annexe). Ici les feuilles sont préparées et déposées par lot dans un bac d'alimentation, qui peut souvent accepter des lots de plusieurs centaines de pages. Le scannage des pages dispose d'un système d'entraînement de papier dont les plus grandes cadences sont similaires à celles des photocopieurs, et dont la productivité peut atteindre plus de 100 pages à la minute. Toutefois, il

faut veiller à ce qu'il n'y ait ni agrafe ni trace de colle sur les bords de page (chacune d'entre elles nuisant à l'entraînement des feuilles), ni de dommage sur les papiers lors du cheminement, ni de traces noires sur la surface de la feuille (nuisant au traitement de l'image et à la lecture de l'information si ces pages contiennent du texte).
Quant aux documents reliés (comme les livres), il existe plusieurs manières de les faire « transporter » :

- face en dessous dans le cas des scanners dits « à plat » (ou « *flatbed* »). Pour cela, la main humaine reste nécessaire en faisant tourner manuellement chaque page avant de le tourner au numériseur ;
- face au dessus pour les scanners dits « à livre ouvert », dont le principe est que l'ouvrage est numérisé à partir d'appareils spécialisés permettant de tourner les pages sans avoir à les retourner (on appelle cela des « tourne-pages », qui nécessite une certaine robotisation)

Toutefois il n'est pas toujours facile de numériser des documents reliés. En effet, les livres ne s'ouvrent pas toujours complètement à plat, en raison du fait que la largeur peut varier d'une page à l'autre. Or une image ne peut être rendue numériquement de manière optimale que si la largeur de chacune des pages est à égale distance de mise au point. Pour remédier à ce problème, il existe au moins deux solutions, qui se font principalement par des numériseurs à plateau de grand format (s'appliquant ici pour des documents de format A1 ou A0) :

- La première consiste à utiliser « un berceau », qui est constitué de deux plateaux qui peuvent être réglables au niveau de l'écartement et de la hauteur pour compenser la variation de hauteur entre la première et la dernière page. Ces pages peuvent être présentées soit en mode « double page » (dans ce cas la page est découpée en deux fichiers distincts), soit en mode « page » (c'est-à-dire une image pour chacune des pages), qui est ici le mode le plus utilisé actuellement A noter que dans le cas des numériseurs avec deux capteurs matriciels, on peut créer une image par capteur avec une zone de recouvrement entre les deux.
- La seconde consiste à utiliser un support à ouverture réduite, de telle sorte que la (les) page(s) est (sont) présentée(s) de manières différentes, que ce soit en mettant à plat une seul page, ce qui a pour inconvénient de diminuer la productivité de moitié, ou en mettant en « V » à 120 degrés les deux pages (deux capteurs sont dans ce cas positionnés horizontalement dans chaque partie du plateau. Cette méthode ne s'applique que dans le cas des livres à reliures fragiles, ceux ne s'ouvrant pas à plat et ceux dont l'image peut être améliorée par la suite si la courbure des informations s'y trouvant est réduite.

Le maintien du papier

Pour s'assurer d'une prise de vue correcte du document papier, il faut aussi s'assurer que le document soit maintenu bien en place, de manière orthogonale. Pour cela les numériseurs disposent de leurs propres manières de les maintenir.

- Pour les scanners à défilement, chaque feuille du lot est maintenue par des guides, après avoir été passé éventuellement sur des taqueuses (système de vibration permettant aux feuilles d'être bien calées de manière orthogonale). Si des feuilles sont prises en doubles, des capteurs à ultrason inclus dans les numériseurs permettent de les détecter.
- En ce qui concerne les scanners « à plat » (voir illustation 5 de l'annexe), le couvercle du numériseur maintient la feuille plaquée contre la vitre dans le but de positionner les bords de feuille bien droit.
- Dans le cas des scanners « à livre ouvert » (voir illustration 3 de l'annexe), il s'agit ici d'empêcher le livre de se refermer et d'aplatir le papier pour éviter que celui-ci fasse un coude avec la reliure et se soulève dans les angles extérieurs. Pour cela, plusieurs méthodes sont possibles en fonction des cas :

+ Les doigts de l'opérateur, qui consiste à passer une main sur la zone de reliure pour l'aplanir et à tirer sur les bords de page avec le pouce pour maintenir les pages ouverts, même si les numériseurs de livres sont équipés d'une pédale de déclenchement pour prendre des vues sans lâcher la page. Cependant cette méthode a pour inconvénient d'une part les traces de doigts sur la feuille (d'où la nécessité des gants blancs ou à défaut, des programmes d'effacement de doigts) et d'autre part le fait que les surfaces de pages ne sont pas toujours planes (ex : plis de journaux)
+ L'application d'une vitre avec de la pression sur deux pages ouverts d'un livre pour les aplatir. Cette méthode a cependant plusieurs inconvénient : la première est que cette technique endommage les reliures, la deuxième est qu'elle prend beaucoup de temps et d'énergie (pour les numériseurs non motorisés) au niveau de la préparation (soulever et abaisser la vitre), et la troisième est que la vitre peut écraser la reliure d'en haut, ce qui peut entraîner parfois des pertes d'informations.
+ Des solutions artisanales comme les fils de pêche, les baguettes de bois recouvertes de papier.

- Pour ce qui est des scanners à plateau, le document est placé de façon à ce que les pages soient soit bien planes, soit inclinées selon un angle de 120 degrés (voir précédemment sur le transport des papiers). Une telle méthode nécessite parfois de maintenir avec les doigts (avec les inconvénients déjà évoqués) et pour principal inconvénient d'endommager la reliure du livre.

Quelque soit le numériseur en question, des caches sont utilisés en fonction de la présentation de l'information numérisée. Ils ont pour but de masquer l'information étant autour de la page (ex : pages du dessous, épaisseur du volume visible par le scanner...), ainsi que de conserver l'apparence du livre. Pour cela, les caches en noir et blanc auront pour rôle de réduire la transparence du papier, tout en étant homogène dans le cadre d'un projet. Dans le cas du cache noir, il est plus facile à détourer mais engendrent des irrégularités et des trous, dont la correction peut engendrer une perte de temps, et donc de productivité.

3.2) L'aspect optique

Avant de s'intéresser à cet aspect, il faut savoir que ce n'est pas tant la page à proprement parler qui est numérisée mais des rayons lumineux, principaux constituant du signal analogique. Par conséquent, c'est aux constituants du signal auxquels nous nous intéresserons.

Lumière, objet et couleur

La lumière est un phénomène ayant deux natures : corpusculaires, autrement dit concernant les photons, qui constituent l'information de base enregistrée par le numériseur en venant frapper les capteurs ; et ondulatoire dans la mesure où elle est composée de fréquences, mises en évidence au moyen d'une source lumineuse et d'un prisme et dont on ne peut apercevoir uniquement que des ondes entre 400 et 700 nm (voir tableau 2 dans les annexes). Ces ondes influencent la coloration de l'image dans la mesure où la couleur est déterminée par sa longueur. Autrement dit, plus elle est élevée, plus elle tend au rouge. Par exemple, un objet de couleur rouge réfléchit les longueurs les plus élevées et absorbe les plus basses, ce qui est l'inverse pour un objet de couleur violet. Toutefois il arrive que certaines pages d'un document soient déjà en couleur et le restent même si on essaie de retirer les couleurs lors de la numérisation. Dans le cas des défauts au niveau des rendus des couleurs, on peut créer un profil ICC qui permettra de différencier les caractéristiques d'un document original et la perception du système de numérisation, ou bien en pratiquant une distorsion inverse, bien que le principal inconvénient de ce dernier est que cela aura pour effet de modifier en grande partie les informations colorimétrique A noter que la couleur d'une image est influencée par l'éclairage, qui peut l'éclaircir ou au contraire la rendra plus sombre.

L'éclairage

Qu'il soit à tubes fluorescent, à LED, ou ambiant, sa configuration est essentielle dans la mesure où la lumière est le vecteur principal de l'information, car elle participe à la reproduction de la couleur.
Pour cela, il existe des dispositions de la source lumineuse variable selon les modèles, comme une source zénithale (qui est la meilleure pour l'uniformité des pages) ou une ou deux sources disposées en biais (idéales pour rendre compte les légers reliefs et éviter les problèmes de brillance), dans le but d'illuminer uniformément les pages et réduire les risques de reflets (comme au niveau des dorures), déjà réduits grâce un traitement antireflet. A noter que l'éclairage ambiant est actuellement de moins en moins utilisé dans la mesure où il s'agit d'un éclairage instable qui dispose à la fois d'une faible profondeur de champ et d'un trop fort niveau de bruit, et donc entraîne une mauvaise qualité de la reproduction numérique.
Toutefois, l'éclairage restera imparfait quelque soit le type utilisé en raison du fait que la décomposition d'une source de lumière montrera qu'une page n'est pas totalement blanche. En effet, le spectre d'une lumière (ensemble des longueurs d'onde) indiquera des longueurs d'onde très élevées et d'autres au contraire très basses, ce qui fait que ces derniers ne pourront être réfléchis par l'objet à numériser, et cela aura pour effet de déformer la couleur de celui-ci.
Ces types d'éclairage guideront la lumière vers le système optique avant d'atteindre le capteur.

Le chemin optique

Pour cela, les rayons lumineux sont concentrés sur la page puis renvoyés sur la surface la plus petite du capteur. Ce qui nécessite donc des objectifs, influençant la qualité des images numériques et dont on en distingue deux types, ayant chacun un coût différent (bien que les hauts de gamme soient les plus chers) :

- Les objectifs à focales fixes, qui peuvent être par exemple à 35 mm, 50 mm ou 60 mm. Ce sont des objectifs n'ayant qu'une seule dimension de numérisation possible, ce qui peut poser problème si on a affaire à des documents d'une dimension inférieure ou supérieure au format fixé par l'objectif, car cela nécessite par la suite des corrections des résolutions qui peuvent coûter assez chers, et au final diminuer la production des documents. Ce genre de problème peut être résolu par les objectifs à focales variables.
- Les objectifs à focales variables, qui peuvent être par exemple à 28-70 mm, 28-105 mm ou 35 mm. Ces objectifs ont pour principe d'utiliser des zooms pour ajuster le compromis cadrage/résolution. Par exemple si le zoom est élevé, la résolution de l'image en sera tout autant, étant donné qu'une portion très petite de la surface de l'image sera captée par le même nombre de cellules. Si ce type d'objectif est moins coûteux et permet d'améliorer le nombre de documents produits, il est toutefois de moins bonne qualité comparé à un objectif à focales fixes.

Il existe sur certains appareils des ouvertures appelées diaphragmes qui permettent de faire passer le flux lumineux et dont on peut régler pour en faire passer une certaine quantité. Par exemple si le diaphragme est très ouvert, la quantité de lumière reçue sera accrue, le temps d'exposition en sera réduit, tout autant cependant que la profondeur de champ. Ainsi, quelque soit l'objectif utilisé, celui-ci présentera toujours des inconvénients que l'on énumérera ci-dessous :

- Le vignettage :

Très présent sur les objectifs à focales variables et sur les grands angles, c'est un phénomène qui peut se traduire par un assombrissement de l'image à mesure que l'on s'éloigne de son centre à cause d'un mouvement des lentilles. On peut atténuer ou compenser ce défaut de diverses manières : une distance focale plus élevée avec une fermeture du diaphragme, les traitement d'image avec par exemple le logiciel Photoshop ou les scanners professionnels, ou des mires ayant des patchs de niveaux de gris pour mesurer le phénomène.

- Les aberrations géométriques :

Ces aberrations peuvent être en barillet (avec un centre bombé) ou en croissant (avec un centre reculé), que l'on peut aisément remarquer sur un quadrillage. Dans le cas du premier, cela concerne les focales courts et dans le cas du second, les focales les plus longues. Ces défauts peuvent être évités ou corrigés en évitant d'utiliser les focales extrêmes et en utilisant les logiciels de retouche d'images (comme Photoshop) ou des logiciels dédiés aux photographies.

- L'aberration sphérique

Il s'agit d'une distorsion de l'image caractérisée par une perte de netteté lorsque les bords de la lentille et les rayons passant à son centre ne convergent pas sur le même plan, cela plus

particulièrement si l'ouverture du diaphragme est très ouverte. Ce qui a pour conséquence d'un effet de flou, limité cependant par l'utilisation combinée de différentes lentilles.

- Les aberrations chromatiques

Elles ont la forme d'irisation sur les contours présentant des contrastes avec un fond de page clair. Elles sont causées par la variation de l'indice de réfraction des longueurs d'ondes, qui ne sont pas déviées de la même façon lors du passage sur les lentilles (les plus faibles étant les plus facilement déviées), dont les verres peuvent s'avérer dispersifs.
Ces aberrations peuvent aussi être causées par un décalage des couleurs intervenant lors d'un mouvement des documents ou de vibrations pendant l'acquisition de l'image. Dans le cas du premier, on peut apercevoir des franges de couleurs sur les bordures de certains motifs, notamment au niveau des caractères et des traits et sur les lignes verticales couvrant la hauteur du document.
Quoi qu'il en soit, les aberrations chromatiques ne sont visibles que lorsque l'on zoome l'image à 50% ou 100%. Lors du stage effectué, ces aberrations ne sont pas retirées car on se devait de supprimer uniquement les couleurs des fonds de page, ce qui tendrait à indiquer que ces aberrations n'ont aucune ou peu d'incidence sur la qualité de l'image.

- Les anneaux de Newton

Ce sont des aberrations optiques intervenant lors de la numérisation de transparents prenant la forme d'anneaux concentriques, dues aux interférences entre les ondes lumineuses réfléchies et les ondes incidentes qui peuvent s'additionner ou se soustraire en fonction de la différence de phase induite par la couche d'air (raison pour laquelle les transparents ne sont jamais numérisés à l'ANRT) ; ainsi qu'aux projections coniques entraînant des déformations de l'image en trapèze et aux réflexions internes à l'objectif qui rendront l'image floue.
Cette aberration peut être corrigée au moyen de plusieurs verres spécifiques et des sprays qui peuvent éviter les couches d'air.

3.3) L'aspect électronique

Dans cet aspect, il est ici question de capteurs composés de photosites, dont le but est de réagir aux photons du signal lumineux. Le processus est que lorsque l'image entre en contact avec le capteur, les photons concentrent les électrons au niveau des photosites atteints. On distingue plusieurs technologies de capteurs, parmi lesquelles :

- Les CCD : il s'agit d'un type de capteur qui délivre un signal analogique sous forme de tension électrique et dont la surface est sensible à la lumière. Pour être quantifiés, ils transfèrent leurs charges à des composants électroniques externes. Les prises de vue pour ce type de capteurs sont très longues mais conservent un niveau de bruit moindre.
- Les CMOS : ce type de capteur dispose de transistors et de composants qui rendent l'appareil autonome. S'ils ne sont pas aussi photosensibles que les CCD, ils permettent en revanche des prises de vue rapides.

Quelque soit le capteur en question, il existera toujours des facteurs qui affecteront l'information délivrée, comme l'agitation thermique, qui aura pour conséquence que les pixels ne seront pas tous noirs parce que l'information des signaux en est perturbé ; la transmission d'une partie d'une charge d'un photosite à un autre ; et la non-reconnaissance d'une couleur par un photosite. Cela n'empêche pas d'effectuer plusieurs types d'acquisitions en fonction des cas :

- L'acquisition linéaire :

Pour effectuer cette acquisition, les capteurs doivent impérativement avoir la forme d'une tablette disposant d'un triple alignement de cellules photosensibles (chacun des alignements filtrant une couleur primaire (R, V, B)), impliquant de ce fait que la numérisation doit se faire avec un déplacement de la caméra au dessus du document, ou le déplacement du document. Cela en faisant défiler l'un des deux en longueur ou en largeur par un moteur à pas fixe. Ce qui permet de ne limiter la définition de l'image de sortie qu'à la taille du document. Définition qui peut être augmenté si on ralentit le moteur et augmente le nombre de lignes contenant les informations.
Si le capteur et le document sont solidaire en formant la caméra, il existe toutefois deux manières de scanner un document : soit en déplaçant la caméra entière ou le document entier, nécessitant pour cela des développements mécaniques lourds ; soit en déplaçant seulement le capteur, présentant l'inconvénient de récupérer les déformation de l'image. Ce type d'application s'applique principalement dans les scanners à défilement.

- L'acquisition matricielle

Dans ce type d'acquisition, les capteurs forment une matrice de cellules photosensibles (que l'on appellera « Matrice de Bayer », qui a la forme d'une mosaïque) et ont une définition en hauteur et en largeur, ce qui permet une acquisition plus rapide de l'image au prix cependant d'une définition plus limitée. Elle est plus limitée car elle dépend de la taille des photosites composant les cellules photosensibles, qui peuvent varier de 4 à 9 microns. Or avec la miniaturisation et l'augmentation du nombre de mégapixels, les dimensions des photosites sont réduites, réduisant ainsi la sensibilité des cellules à la lumière et augmentant le bruit, entraînant une diminution de la qualité de l'image.
Dans le cadre des couleurs, le capteur dédie à chaque photosite l'une des trois couleurs primaires, dont deux fois plus de vert que de bleu et de rouge en raison des contraintes de traitement de signal qui imposent un motif répétable de 2x2 cellules, et à cause de la sensibilité de l'œil humain qui est plus élevée que la longueur d'onde moyenne.

Quelque soit le type de capteur en question, il doit tenir compte de plusieurs facteurs pour bien numériser l'image :

- La définition :

Ce facteur dépend du nombre de photosites qui composent le facteur, et par conséquent de la résolution maximale d'échantillonnage proposée par le numériseur. Pour cela, il faut vérifier que la définition du capteur correspond à la combinaison « taille maximale x résolution maximale »

- La sensibilité

Il s'agit du taux de conversion de photons incidents en électrons, étant donné que seule une partie de la surface d'un photosite est sensible aux photons. Si la taille du photosite est réduite, sa partie

active est plus étroite et le capteur perd en sensibilité. Ce qui serait l'inverse si ce photosite est plus élevé.

- La dynamique

Elle est liée à la capacité de stockage des électrons des photosites, dont les plus petites peuvent en être saturées si l'exposition est trop longue ou si la lumière est trop intense. Avec le minimum d'électrons nécessaire à la représentation de l'information la plus foncée, elle peut faire l'objet de la différence entre elle-même et ce minimum, déterminant ainsi la plage dynamique d'un capteur.

- Le bruit

Il en existe plusieurs origines : dans le cas du bruit « thermique », qui est lié au phénomène de « courant d'obscurité », la température élevé fait que des électrons peuvent se déplacer de manière aléatoire au niveau du capteur, créant des charges transformées en signal par l'électronique du scanner, cela indépendamment de la luminosité. Dans le cas des couleurs sombres, la pureté est perdue en raison des pixels parasites, qui prendront beaucoup d'importance dans le signal.
Il peut exister d'autres causes comme la persistances des charges acquises précédemment dans les photosites (phénomène de rémanence) ou la perte des charges lors du déchargement des photosites (effet de traîne).

3.4) L'aspect informatique

Dans cet aspect, il s'agit de finaliser la numérisation en le convertissant dans un format image, apporter des traitements informatiques et copier sous forme de fichier.
Pour cela, il faut un numériseur couplé à un PC, ce dernier disposant à la fois d'un système d'exploitation, d'un disque dur, de ports de communication, d'un écran, d'un clavier et diverses dispositifs de commande ; et dans lequel on installe un logiciel de réglage (ex : PixEdit7) qui permettent à la fois de régler le numériseur, de déclencher la numérisation et de vérifier le résultat. Ce logiciel dispose la notion de lot, dans lequel une série d'images se crée à partir à partir des paramètres enregistrés.
Ces paramètres concernent la résolution d'échantillonnage, le mode colorimétrique, le mode de compression et le format de fichier, auxquels on peut ajouter la détection des doubles ou des agrafes (pour les scanners à défilement), l'association d'un profil ICC, les applications de post-traitement, etc. Ces paramètres peuvent varier d'un lot à un autre, par exemple le changement de résolution, le passage d'un scanner recto à un scanner recto/verso pour les feuilles volantes.
Chaque image produit un fichier qui sera nommé à partir d'un identifiant. Par exemple dans le cadre du stage, on identifiait un fichier image de la manière suivante : 0000xxxx (ou parfois 0001xxxx, étant donné que le logiciel utilisé pose des problèmes dans la mesure où les configurations ne permettent pas une nouvelle numérisation sans écraser les fichiers précédents).

Une fois les fichiers réalisés, l'ensemble est défini selon le fichier en tant que TIFF ou JPEG, puis inclus dans un dossier, lui-même inclus dans un répertoire voir un ensemble de répertoires. Une fois le document numérisé, des post-traitements sont nécessaires pour la bonne lecture de

l'image. C'est le cas par exemple, lors du stage effectué, des retraits de fonds de couleurs (pour retirer le poids de l'image) et la rotation (qui est simple si elle est réalisée sous un angle de 90°, 180° ou 270°, dans le cas contraire il faut réaliser des calculs de pixels au moyen d'algorithmes, qui ne sont pas fiables à 100%, pour éviter des anomalies au niveaux des pixels). Une fois le post-traitement réalisé, il s'agit désormais de passer à la diffusion du document.

Chapitre 3 : De la diffusion des documents numériques

Lors du stage effectué, l'une des tâches à effectuer était de mettre en place un document (par exemple sur Bloc-Notes) dans lequel on doit d'abord donner le numéro NNT (identifiant de la bibliothèque universitaire ciblée) puis, après l'ajout d'un point virgule, l'URL de la thèse ciblée. Le tout en tapant le numéro ANRT puis copiant les informations depuis un site (ici l'ANRT), ce qui donne le résultats comme indiqué dans l'introduction. Ainsi la diffusion est quelque chose qui se rapporte à l'ordinateur, et plus précisément à Internet, au support numérique et aux sites Internet proposant des documents à diffuser.

1) De l'utilité des sites Internet pour la diffusion

Pour tenir compte de l'utilité des sites Internet, il faut tenir compte des usages que font les utilisateurs de ceux-ci. En effet, depuis le début des années 2000, et plus précisément depuis 2005 avec l'arrivée de Google, les recherches s'effectuent de plus en plus sur des sites Internet ainsi que sur divers réseaux, au point d'utiliser dans certains cas des terminaux mobiles, tels les smartphones, des tablettes ou des liseuses. Toutefois, certaines études tendent à mettre à mal l'utilisation des sites web: une téléphonie mobile qui tend à concurencer la téléphonie fixe en terme d'usage (1) , une augmentation des usages des smartphones dans le cadre de l'utilisation d'Internet (7% en 2007 contre 21% en juin 2012) (2) et des téléchargements des applications mobiles de plus en plus présents au dépens des sites web.
Cela signifie-t-il pour autant que le site web est condamné ? Ce serait cependant oublier que malgré sa grande mobilité et le fait qu'il permet de naviguer autant qu'un ordinateur, le support mobile présente l'inconvénient de ne pas toujours lire certains documents (comme les documents PDF dans le cas des tablettes Android), du fait que la capacité de mémoire et de stockage d'une tablette ou d'un smartphone restera toujours nettement inférieure à celle d'un ordinateur, ainsi que du fait que certains formats resteront incompatibles pour les supports mobiles (ex: le format PDF), et enfin le fait que certains supports mobiles nécessitent une connexion wifi, qui peut influer directement sur la navigation d'Internet et sur les certificats de sécurité. Ce sont des inconvénients dont ne dispose pas ou peu l'ordinateur qui, en dépit de son immobilité, reste utile pour stocker un grand nombre d'informations, en particulier si ce sont des commandes d'ouvrages ou de thèses dans les magasins de sites Internet, qui nécessitent souvent beaucoup d'octet.
Quoi qu'il en soit, les sites web connaissent tel essor qu'il en vient à tantôt concurrencer les recherches "traditionnelles" (les recherches par les livres), tantôt compléter ce dernier type de recherche pour mieux l'approfondir et le compléter (3)[5]. Ce qui n'empêche pas une partie de la population de "*concentrer ses pratiques culturelle sur une consommation technologique*

(1) http://www.credoc.fr/pdf/Sou/Credoc_Diffusiondes TIC_2012.pdf
(2) http://www.atinternet.fr/actualites/baisse-du-trafic-moyen-des-sites-web-en-france-en-2012-explosion-des-applis/
(3) http://lafrabrique.bnsa.aquitaine.fr/wp-content/uploads/2012/12/Pratiques_culturelles_Aquitaine_2012.pdf

exacerbée" (1) En ce qui concerne le patrimoine et l'aspect culturel, une étude du Crédoc tend à montrer que les visiteurs de musées, d'exposition et de monument sont de plus en plus attirés par les produits culturels comportant une innovation technologique (54% de satisfaits) (2)[3].
Ainsi on peut remarquer le fort attrait d'une partie de la population pour ce qui est technologique et numérique. D'un autre coté, il ne faut pas oublier que le fait de faire une commande sur Internet a de nombreux avantages, à savoir le fait que cela ne nécessite plus de se déplacer vers le service documentaire en question (on peut parler en quelque sorte d'une certaine "réduction de distance" entre l'utilisateur et le service), puis que la commande sur Internet prend autant moins de temps à envoyer que la réponse qui s'ensuit; et cela contrairement au courrier papier, dont l'envoi vers le destinataire et la réponse de ce dernier peuvent prendre plusieurs semaines voir plus étant donné le fait que le personnel du service dispose déjà de nombreuses commandes et de données à traiter puis à envoyer les réponses et les commandes, et enfin à numériser de nouveaux documents. Enfin le dernier avantage pour un site Internet proposant des documents est que le site peut proposer un même document pour plusieurs commanditaires. En effet, contrairement à un magasin non numérique qui propose des produits en quantité limité, il peut proposer des documents en quantité illimité, du fait que le document numérisé est un document qui passe au moyen des canaux numériques vers les ordinateurs, ce qui fait que le document numérique commandé peut être reproduit et commandé plusieurs fois en même temps (à condition que le support du destinataire ait suffisament d'espace). De ce fait, on peut parler d'ubiquité.
Une ubiquité qui répond au concept de "numérisation de masse", et dont le site peut être inclus sur un moteur de recherche comme Google (qui en est le plus important), à condition que le site en question respecte les obligations et les principes de ces moteurs, comme le principe d'hébergement ou la confidentialité. Une fois ces principes respectés, le moteur de recherche peut permettre la mise en place du site en question, qui peut donc inclure des documents numériques dans son magasin pour mieux les diffuser.

2) Comment faire une mise en ligne ?

Avant de faire une mise en ligne, il faut savoir que les diffusions des documents se font dans la plupart des cas au format PDF, ou plus précisément au format PDF Image, dans la mesure où il s'agit du format présentant le moins de perte de données. Format auquel l'ANRT utilise dans le cadre de la diffusion de leur thèses. Cette mise en ligne se fait en plusieurs étapes :

Définition des métadonnées

Avant de diffuser un document, il faut tout d'abord établir des métadonnées, afin de définir les caractéristiques du document. Dans le cas de l'ANRT, les métadonnées se construisent à partir des informations récoltées dans les formulaires de thèses (ou à défaut dans les thèses elles-mêmes). Ici, on crée un fichier .CSV dans lequel on construit un tableau dont on tient en

(1) **Sous la direction de Thierry Claerr et Isabelle Westeel.** *Manuel de constitution d'une bibliothèque numérique.* Editions du Cercle de la librairie. Paris: Electre, 2013. 408 pages. ISBN : 978-2-7654-1413-1. (page 45)
(2) http://www.credoc.fr/pdf/Rapp/R281.pdf

compte le titre de la thèse, le nom et le prénom de son auteur, le nombre de pages, le numéro d'identification de la thèse (n° NNT), le format et la date de la soutenance de la thèse. Ces métadonnées sont séparées par des séparateurs de champs sous forme de point-virgules (;), comme montré dans l'introduction et par des séparateurs de texte sous forme de guillemet (" "). Pour l'inclure ces métadonnées dans le site, on se connecte d'abord dans le panneau d'administration pour ensuite aller dans l'interface du site, dans lequel on choisit l'outil (ex : dossier), le fichier crée pour le charger ensuite, sélectionner les entités (ici les séparateurs de champs et les séparateurs de texte) et enfin choisir le fichier .CSV en question pour le charger. Le chargement du fichier aura pour conséquence de rendre les métadonnées visibles sur le site.

Encodage et structuration

Entretemps, les métadonnées, une fois définies, sont encodées selon un certain type de codage et structurées selon un format bien défini. Dans le cas de l'ANRT, l'encodage se fait en UTF-8, en référence aux 8 bits traversé dans chaque canal définissant une couleur. En revanche la structuration des métadonnées dans ce service est inconnu, vu que le service en question n'est pas une bibliothèque. Nous présenterons néanmoins deux structurations pour cerner les enjeux :

- **Dublin Core**

Crée en 1995 dans cadre de la conférence de Dublin (Ohio) et sous l'initiative de l'OCLC et de la NCSA, elle a pour but de palier les défauts des balises HTML qui sont :

- le lait qu'elles soient peu nombreuses ;
- Captives du Web pour l'affichage et les documents web ;
- Peu structurées car à plat et détachées de la logique du document ;
- Peu utilisées pour les détournements;

Ainsi ce format répond à ces problèmes en étant plus structuré, plus propices aux recherches et aux signalements de ressources et transversales aux autres formats. Il dispose de 15 éléments descriptifs, répétables et facultatifs :

Élément	Élément (anglais)	Commentaire
Titre	*Title*	Titre principal du document
Créateur	*Creator*	Nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document
Sujet	*Subject*	Mots-clefs, phrases de résumé, ou codes de classement
Description	*Description*	Résumé, table des matières, ou texte libre

Élément	*Élément (anglais)*	Commentaire
Editeur	*Editor*	Nom de la personne, de l'organisation ou du service à l'origine de la publication du document
Contributeur	*Contributor*	Nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document
Date	*Date*	Date d'un évènement dans le cycle de vie du document
Type de ressource	*Type*	Genre du contenu
Format	*Format*	Type MIME, ou format physique du document
Identifiant de la ressource	*Identifier*	Identificateur non ambigu, dont il est recommandé d'utiliser un système de référencement précis
Source	*Source*	Ressource dont dérive le document
Langue	*Language*	Langue du document
Relation	*Relation*	Lien avec d'autres ressources
Couverture	*Coverage*	Couverture spatiale (, pays, régions, noms de lieux) ou temporelle
Droit	*Rights*	Droits de propriété intellectuelle, Copyright, droits de propriété divers

(1)

A noter également qu'il existe une extension du Dublin Core appelé "Dublin Core qualifié", dans lequel certains éléments disposent eux-mêmes de sous-ensembles, comme par exemple le titre, la description, la couverture ou encore les relations. A noter qu'il s'agit du format le moins étendu (et donc le moins précis) en terme d'éléments.

- **Le format UNIMARC**

---

(1) http://fr.wikipedia.org/wiki/Dublin_Core

Ce format est un des nombreux dérivés du format MARC (qui fut crée en 1964 dans le cadre du catalogage des documents de la bibliothèque du Congrès américain), dont il en est le principal pivot entre tous les types, dans la mesure où il peut s'adapter à toutes les langues et à tous les encodages.
Il dispose de 10 blocs numérotés de 0 à 9, dans lequel on compte au moins jusqu'à 100 zones. Ces zones sont divisées en sous-zones (dont on peut le délimiter grâce au symbole $ suivi d'un chiffre ou d'une lettre) et ont pour indicateurs numériques les chiffres 0 et 1. La ponctuation de catalogue disparaît alors. La notice peut être retrouvée par une recherche par accès, est disponible dans les bibliothèques du monde entier (dont la Bnf et l'ABES), et peut être récupérée par bande magnétique, support optique ou par Internet.

Forme	Bloc	Description
0XX	Bloc des numéros d'identification	ISBN, ISBN, numéro dans le catalogage associé
1XX	Bloc des informations codées	Dates, langues, pays,...
2XX	Bloc des informations descriptives	« pavé ISBD »
3XX	Bloc des notes	Reproduction facsimilé, contenu,...
4XX	Bloc des liens	Collection, histoire,...
5XX	Bloc des titres associés	Titre uniforme, titre parallèle
6XX	Bloc de l'indexation matière	Nom commun, classification
7XX	Bloc des responsabilités intellectuelles	Auteur principal, personne physique, collectivité
8XX	Bloc des données internationales	Source de catalogage. Centre ISSN
9XX	Bloc des données locales	Données d'exemplaire

L'inclusion dans une base de données

Une fois les métadonnées, le format et l'encodage fixé, il s'agit de l'inclure dans une base de données dans lequel chaque élément dispose de ses propres caractéristiques, comme par exemple l'auteur dont on peut définir le nom, le prénom et la date de naissance, le tout précédé d'un "pk_id xxxx" (dans le format MLD, la caractéristique peut se terminer par "#fk_xxxx") . Il existe un intermédiaire dans lequel on a un titre évoquant une action, en dessous de quoi on a deux "Fk_id_xxx" qui sont pour chacun des caractéristiques du document (il s'agit ici d'un des nombreuses formes de base de données). On peut donner un court schéma comme indiqué dans la figure de l'annexe. Toutefois dans le cas de l'ANRT, on n'inclut pas les données personnelles

de l'auteur de la thèse comme la date de naissance ou son adresse (présent dans le formulaire) par souci de confidentialité.

L'inclusion dans le site

Une fois les métadonnées réalisées, il s'agit de mettre le contenu du fichier (c'est-à-dire le texte et les images numérisées) dans le site. Pour cela, il s'agit d'expédier le fichier avec le bon format (ex: PDF) vers le webmestre (ou *webmaster*), dont l'ANRT n'en dispose pas dans l'établissement même mais dont on en fera une description générale.
Un webmestre est une personne "responsable d'un site web, de sa conception à sa maintenance » (1). Il est donc une personne chargée de la conception d'un site (en créant des interfaces accessibles à tous et validées par le W3C car en effet, il doit veiller à la bonne cohérence et à la bonne architecture du site, en effectuant une certaine mise en scène, dans lequel le respect de la cohérence portera à la fois sur la forme (charte graphique) et sur le fond, qui peuvent cependant s'opposer l'un et l'autre. La personne chargée de cela, qui peut être infographiste, ergonomiste, développeur, rédacteur de contenu ou autre, doit être ainsi étroitement associée à la définition du contenu), de son développement, de sa maintenance (au moyen d'une veille documentaire) et de sa mise à jour, tout en créant des contacts avec les services et les utilisateurs par des enquêtes (2), en tenant compte de leurs avis ainsi que du marché et des lois qui sont appliquées dans ce domaine. Il doit disposer d'une bonne connaissance du fonctionnement de son institution, de son entreprise, de son administration et d'une bonne connaissance informatique générale.
Dans le cadre des fichiers, le webmestre sélectionne le contenu (après l'établissement d'un comité de direction et d'un comité de pilotage pour détecter les éventuelles obstacles, le convertit selon le format souhaité (ex: PDF si le fichier n'était pas converti avant) et selon le standard du Web (comme le W3C), après quoi le document est publié sur le serveur (3) en incluant pour cela un URL propre à un document et/ou à son auteur (ex : http://www.diffusiontheses.fr/3837-thèse-de-larroque-michel.html). Enfin, le webmestre l'inclut dans la page ciblée, qui peut être une suite de nombreuses pages dans la mesure où la page ciblée peut faire partie d'une sous-catégorie d'un thème. On peut en donner un exemple dans le cadre du site de l'ANRT :

« Histoire » --> « 19ème siècle » --> « Thèse de GARMENDIA Vincent » → « L'IDEOLOGIE CARLISTE (1868-1876) AUX ORIGINES DU NATIONALISME BASQUE. 1070 p. »

Ainsi grâce au webmestre, l'internaute ou le service peut commander le document en question, moyennant quelque somme. Le document en question peut être envoyé au destinataire selon une durée variable selon les clients (de 15 jours à plusieurs semaines).

3) La diffusion des documents : les destinataires et les usages

(1) http://fr.wikipedia.org/wiki/Webmestre
(2) http://metiers-internet.eu/webmestre.html
(3) http://admi.net/industrie/jmycs/divers/webmaitre.html#public

Lorsque l'on parle de destinataires dans le cadre de la diffusion, on parle aussi bien des internautes que de certains services ne disposant pas de moyen de numérisation. Dans le cas du premier, les raisons peuvent être multiples comme par exemple la curiosité intellectuelle ou le fait de faire une référence dans le cadre d'une recherche pour un exposé, une thèse même un mémoire (dans le cas des étudiants) ou même l'enseignement et la recherche (dans le cas des professeurs). Dans le cas du second, il s'agit d'une part de pallier l'absence de moyen de numériser, d'autre part d'agrandir la collection sur un domaine, un thème ou un sujet précis, comme c'est le cas des bibliothèques.
Tandis que dans le premier, on s'oriente volontiers vers les sites proposant les documents, dans le second cas, le document envoyé vers le destinataire fait souvent l'objet d'une commande de la part du service en question. En effet, le commanditaire fait souvent une commande dans lequel il indique le titre du document ciblé, le nom et prénom de l'auteur, le format physique souhaité (papier, microfilms ou numérique), la présentation, le format numérique (si le document est numérique), les caractéristiques du document (par exemple au niveau de la colorimétrie ou de la résolution) et enfin du délai et des conditions d'exécution. Ce qui n'est pas le cas ici de l'ANRT qui, après avoir obtenu le document (sous forme papier) auprès d'une université, le numérise et le microfilme pour ensuite le faire inclure dans le site (à noter que le document papier est détruit après la numérisation). Ainsi nous sommes dans une dimension de marché dans lequel s'effectuent l'offre et la demande. Un marché qui cependant est limité d'une part par le Code du marché, et d'autre part par des lois impliquant la diffusion et l'exploitation d'un document.
En effet, les documents font l'objet ces lois selon le contexte de leurs usages, qui ne se font pas sans contraintes. Dans le cas de la copie privée (que ce soit une partie ou tout le document), l'utilisateur peut le réaliser sans autorisation (1) à condition qu'il applique certaines conditions, à savoir que la copie ne se fait que par la personne le faisant personnellement, aves son matériel et uniquement pour un usage personnel. Toutefois dans la pratique, il faut distinguer dans ce cas deux situations (2):

- Dans le cas d'une copie par la bibliothèque, "*il faut se référer à la cession consentie par l'auteur à la bibliothèque pour la numérisation et la mise à disposition du public de son oeuvre. Les usages [...]doivent être précisés*" (2) [1]
- Si la copie est réalisé par l'usager avec son propre matériel numérique, la bibliothèque n'intervient pas mais les auteurs et les éditeurs peuvent avoir une rémunération dont le montant est perçu lors de la vente de support. (3)

---

(1) Article L.122-5 du CPI

(2) **Sous la direction de Thierry Claerr et Isabelle Westeel.** *Manuel de constitution d'une bibliothèque numérique.* Editions du Cercle de la librairie. Paris: Electre, 2013. 408 pages. ISBN : 978-2-7654-1413-1. (pages 108-109)

(3) Article L.311-1 du CPI: "*Les auteurs et les artistes-interprètes des oeuvres fixées sur phonogrammes ou vidéogrammes, ainsi que les producteurs de ces phonogrammes ou vidéogrammes, ont droit à une rémunération au titre de la reproduction desdites oeuvres, réalisée à partir d'une source licite dans les conditions mentionnées au 2° de l'article L. 122-5 et au 2° de l'article L. 211-3. Cette rémunération est également due aux auteurs et aux éditeurs des oeuvres fixées sur tout autre support, au titre de leur reproduction réalisée à partir d'une source licite, dans les conditions prévues au 2° de l'article L. 122-5, sur un support d'enregistrement numérique. »*

Dans le cas de l'utilisation des documents à des fins non privés, il existe au moins deux cas de figures auxquels l'auteur dispose de l'autorisation de reproduire : les analyses et les courtes les courtes citations si elles ont un caractères critique, polémique, scientifique ou d'information, si le nom de l'auteur et la source du document sont indiqués, et si elles sont brèves; et l'utilisation pour l'enseignement et la recherche, qui est une exception du droit d'auteur. Dans le cas présent, elle dispose de nombreuses contraintes, à savoir:

- Que le document soit des extraits d'oeuvres à des fins exclusives d'illustration dans ce cadre, avec des compensations par rémunération forfaitaire;
- Le fait que la mise en oeuvre soit encadrée par un protocole d'accord sur tout type de support à ces fins avec les auteurs en question;
- Le fait que la mise en ligne des travaux pédagogiques ou de recherche impliquant des oeuvres ou des extraits d'oeuvres protégées doit faire l'objet d'une déclaration sur le site du Centre français d'exploitation du droit de copie (CFC), dont les utilisateurs se doivent de fréquenter.

Ainsi l'usage et la diffusion du document dans un but de recherche sont des choses étant très encadrées par les lois, qui ont pour but de protéger les droits d'auteurs contre tout risque de copie. Ce qui n'empêche pas l'utilisateur d'utiliser le document de son choix, qu'il soit numérique, papier ou sous forme de microfilms, choses que propose l'ANRT. Reste à savoir si ce sont majoritairement des documents numériques qui sont utilisées par les destinataires.

4) Le document numérique : l'avenir du document ?

Comme on l'a expliqué dans la sous-partie précédente, de plus en plus de personnes tendent à être attirés par le document numérique, notamment grâce aux navigateurs, applications et logiciels et aux supports qui en sont rattachés. Ce qui n'empêche pas de se poser des questions le conçernant. En effet, outre les problèmes juridiques propres aux documents numériques, ces derniers sont également concernés par l'obsolescence technologique. Cette obsolescence technologique se traduit par le fait que les outils informatiques tendent à évoluer dans le temps, influençant de ce fait l'accessibilité au document et à l'information numérique. Cette accessibilité se repose à la fois sur un empilement de couches technologiques (ex: l'empilement "support d'enregistrement de l'information – matériel de lecture – logiciel décodant et restituant l'information – périphérique de restitution"), ainsi que de technologies interdépendants entre eux, sans lesquelles aucune forme de technologie ne pourrait fonctionner.
D'autre part le problème peut aussi venir tantôt des formats de données, qui disposent des mêmes problèmes d'obsolescence technologiques en plus des problèmes d'ordre juridique et au niveau du coût et de la lisibilité de l'information, tantôt des supports au niveau de deux aspects:

- la première est au niveau matériel, dans la mesure où il existe une corrélation entre la densité d'information d'un support, la complexité de l'encodage et sa fragilité. En effet, tantôt la dégradation partielle d'un support peut entraîner la perte d'information au niveau de la partie dégradé, tantôt une erreur de bit dans un fichier peut entraîner la perte totale

de l'information. Dans ce dernier cas l'information (qui est représentée avec une suite de 0 et de 1) est perdue définitivement car avec la dégradation d'un support, la qualité du signal binaire se dégrade également et par conséquent, des erreurs d'interprétation peuvent survenir avec l'inversion des 0 et des 1. Le résultat est que le document devient illisible, cela sans que l'on sache à quel point le document à été dégradé. Cette dégradation peut survenir lors de la compression des données (en particulier les compression avec perte) et peut être cependant résolu à condition d'intervenir avant que les perte de données se fassent au moyen de procédure de restauration de données extrêmement coûteuses. D'où le terme "préservation" qui est un anglicisme de "conservation", dans le sens de "conservation sur le long terme d'un contenu"

- La seconde est au niveau technologique. En effet, la technologie est quelque chose qui dépend du marché de l'informatique, de la logique des éditeurs et de la concurrence entre plusieurs marques de l'informatique. Si le format de données peut être utilisé pendant plusieurs dizaines d'années, ce n'est pas le cas des supports car un support comme un CD dispose d'une durée de vie limité (généralement de trois à cinq ans), risque de ne pas trouver le modèle de lecture ou le système d'exploitation adéquat, et peut se révéler dans certains cas fragile et instable. Ce qui fait d'ailleurs que certains formats bureautiques comme WordPerfect ne sont plus maintenu, tout comme les logiciels de lecture.

Dans le cadre de la création de données, son contexte (autrement dit l'auteur, la date de publication, le statut, le sujet et la durée de conservation envisagée) peut être perdu même si le document et ses informations sont lisibles, d'où la nécessité des métadonnées permettant de le décrire, cela dès la création du document numérique. Car en effet, si les outils informatiques permettent de créer facilement un document, elle a pour inconvénient de faire accumuler des données peu qualifiées et non classées jusqu'à les rendre de plus en plus difficiles à les trouver, en particulier lorsque l'on a affaire à des systèmes d'exploitation grand public. Une information non qualifiée qui ne dispose pas de renseignement sur les métadonnées ou de mots-clés d'indexation ni de nom de fichier devient une information perdue.
Ainsi, si le document numérique à l'avantage d'être reproduit à l'infini grâce à son immatérialité, il dispose donc de certains inconvénients. Inconvénients dont le document papier ou les microfims ne disposent pas, même si ces derniers peuvent dépendre du numérique grâce à l'impression commandée par ordinateur et à la réalisation des microfiches par ordinateur (1). De ce fait ces inconvénients tendent à freiner l'utilisation du document numérique au profit du papier et d'autres types de document. Par ailleurs, un rapport d'étude de l'AIIM, réalisé entre novembre 2011 et janvier 2012 sur un panel de 395 personnes et dont le titre est « *The Paper Free Office-dream or reality ?* », tend à démontrer que 77% des factures PDF sont imprimées au format papier, dont 10% d'entre elles réalisées plus d'une fois et 16% numérisées à nouveau ; bien que selon ce même rapport la tendance serait à la baisse, en particulier dans les grandes entreprises, dont le nombre de documents papier et de documents numérique est extrêmement élevé. En outre, pour une bonne partie des répondants, l'adoption du « tout numérique » permettrait une réduction de l'espace de stockage de 15% sur une année et de 35% sur cinq ans (2).

(1) A noter également que lors du stage effectué, l'une des tâches à réaliser était d'exporter les fichiers vers tous les postes, y compris les postes de création de microfiches.

(2) IW_Paper-free-Capture_2012.pdf (www.aiim.org)

A travers ce rapport, on peut donc établir plusieurs constats, à savoir que le document papier, malgré la tendance à la baisse et la préférence d'une bonne partie des sondés pour le « tout numérique », occupe encore une place importante dans le cadre des services; que le document papier et le document numérique sont interdépendants, que les services ayant la plus forte part de documents numériques sont ceux étant de taille élevée, et que le numérique aurait pour avantage de réduire l'espace physique (on parle alors d'une émergence d'une sorte de bibliothèque numérique). Ce qui fait que l'idée du « zéro papier », ne sera pas encore pour demain, du moins en ce qui concerne les petites entreprises.

Car en le fait de choisir un tel type de document physique pour un service dépend de plusieurs paramètres, à savoir la taille de ce service, la distance entre ce dernier et un utilisateur, le budget, le coût du document et le besoin des utilisateurs. C'est-à-dire que plus la taille du service et la distance entre ce dernier et l'utilisateur est élevée, plus le document numérique prendra une grande place dans le service. Et plus le budget est serré, plus le besoin de faire des économies en terme matériel se fera sentir et plus cela nécessitera de choisir le type de document le moins coûteux (les méthodes de numérisation et le format les moins coûteux si on choisit le document numérique). Par conséquent, la proportion de documents numériques, de documents papier et des autres types de documents ne sera pas uniforme partout puisqu'il variera en fonction des services et des besoins des utilisateurs.

La deuxième raison tient au fait les types de support peuvent être interdépendants les uns des autres. En effet, pour qu'il y ait un document papier non manuscrit, il faudra effectuer des impressions commandées depuis un ordinateur, et en particulier si le document à imprimer est basé sur une page d'un site qui est originellement numérique ou basé sur des documents papier. De même pour les documents numériques, et en particulier pour les documents numériques ne l'étant pas originellement, ils peuvent faire l'objet d'une numérisation par scanner à partir d'un document papier (2) ou d'un microfilm pour diversifier les supports. Numérisation qui peut conduire par la suite à la réalisation de microfiches, dont le principe est le même que celui des microfilms, à savoir observer les images grâce à des appareils spécialement conçu pour ce support.

Ainsi on peut dire que le document numérique, même s'il occupe une place prépondérante dans le cadre des utilisations, ne sera jamais l'unique document qui sera utilisé à l'avenir pour les raisons évoquées précédemment. D'ailleurs, Stéphane Caro, auteur de l'article «*Document papier, document numérique* », note que le document papier rencontre encore une certaine résistance dans le domaine de l'utilisation (1) [4]. Ce qui n'empêche pas cependant le fait que la technologie évolue dans le temps, tout comme les supports censées contenir le document numérique, comme les tablettes et les smartphones, même si ces derniers tendent à entrer en concurrence avec l'ordinateur (notamment en ce qui concerne la navigation, comme déjà expliqué dans la sous-partie « De l'utilité des sites Internet pour la diffusion » ; ou alors le cloud, dont on peut déposer ses propres données pour ensuite le récupérer, quelque soit le support accueillant le document numérique.

---

(1) **CARO, Stéphane.** *Document papier, Document numérique.* Publié le 10 novembre 2003. Référence : H7225

(2) A noter que le personnel de l'ANRT jette les documents papier ayant été numérisés, les thèses étant réimprimables

CONCLUSION

Lorsque la numérisation des documents eut lieu dès les années 1990, ce fut le début d'un nouveau type de document qui se mettait alors en place, et dont de nouveaux types de supports sont crées pour les accueillir. Ce sera alors en 2005 avec l'arrivé de Google sur le Web que le document numérique prend un véritable tournant avec le concept de « numérisation de masse », dans lequel on observa une demande croissante en terme de documents numériques, vu le fait que le document numérique est plus accessible que les autres types de document en raison de ubiquité et de sa facilité de transmission. Pour faire face à cela, de l'aide et du soutien de la part de diverses organisations et des politiques, s'avéraient nécessaires pour faire face à ce nouvel enjeu, qui impose de nouvelles formes de numérisation, dont on a décrit les aspects entourant l'image numérique, quelques uns des nombreux formats rattachés, ainsi que l'obtention de l'image numérique par les scanners (c'est d'ailleurs l'une des tâches principales réalisés à l'ANRT). Le tout en tenant compte de nombreux aspects l'entourant comme les droits d'auteur, le budget du service de numérisation, les besoins des utilisateurs et enfin de la diffusion des documents numériques (ce dernier qui est abordé lors du 3ème chapitre).
Ainsi pour répondre à la problématique « Comment expliquer cette importance centrale qu'occupe la numérisation ?", on peut répondre par le fait que la numérisation du document ne permet pas seulement de numériser le document numérique, mais permet aussi de communiquer plus efficacement le document tout en tenant compte des lois sur les droits d'auteur en matière de numérisation et de diffusion, les besoins des clients, du budget et des clauses décrites dans le cahier des charges. De ce fait, cette importance centrale s'explique par le fait que la numérisation regroupe un large panel de thèmes en ce qui concerne l'information, la communication, la documentation et les bibliothèques.
Ce qui n'empêche pas non plus de disposer de nombreux inconvénients comme le fait que le support numérique peut se détériorer, entraînant par conséquent des pertes d'informations qui peuvent être partielles ou totales; un format qui peut être inadapté par rapport à un autre type de support, ou alors une mauvaise numérisation dûe par exemple à un mauvais codage de couleur, une mauvaise compression, ou alors des défaillances techniques de la part du scanner. Par conséquent, et en raison de ces défauts le document numérique ne sera jamais vraiment le seul support valide à l'avenir (et cela même s'il gardera une place prédominante dans la société grâce à son attrait vis-à-vis des utilisateurs) et on peut ajouter d'autres raisons supplémentaires, selon laquelle le document numérique et les autres types de documents seraient complémentaires, et selon laquelle le type de document dépend des nombreuses caractéristiques du service abritant les documents.
Quoi qu'il en soit, il ne faut pas oublier qu'en dépit de ces défauts, le document numérique aura toujours des moyens de corriger ces inconvénients par des logiciels de correction et de préservation d'image, ainsi que de nouveaux supports, de nouvelles technologies et de nouveaux formats qui s'adapteront à ces problèmes ainsi qu'aux nouveaux défis que pose le numérique , comme par exemple la "troisième révolution industrielle".

Bibliographie

Livres :

[1] **Sous la direction de Thierry Claerr et Isabelle Westeel.** *Manuel de constitution d'une bibliothèque numérique*. Editions du Cercle de la librairie. Paris: Electre, 2013. 408 pages. ISBN : 978-2-7654-1413-1.

[2] **Sous la direction de Thierry Claerr et Isabelle Westeel.** *Manuel de la numérisation.* Editions du Cercle de la librairie. Paris: Electre, 2011. 320 pages. ISBN : 978-2-7654-0983-0.

Articles sur Internet:

[3] **BIGOT Régis, DAUDEY Emilie, HOIBIAN Sandra et MÛLLER Jörg ,** *La visite des musées,des expositions et des monument.* Juin 2012 <http://www.credoc.fr/pdf/Rapp/R281.pdf >

[4] **CARO, Stéphane.** *Document papier, Document numérique*. Publié le 10 novembre 2003. Référence : H7225

[5] ***DEBANT, Anne et PERROT, Patrick .****Module 09 - Reproduction par microfilmage et numérisation.* 14 novembre 2011 <http://www.piaf-archives.org/espace-formation/course/view.php?id=11 >

[6] **PÉDAUQUE, Roger.** *Document: forme, signe et médium, les reformulations du numérique.* 8 juillet 2003. (consultation le 27 juillet 2012) <http://archivesic.ccsd.cnrs.fr/docs/00/06/21/99/PDF/sic_00000511.pdf>

Glossaire

- Le bit

Il s'agit de la quantité élémentaire d'information dans laquelle on choisit souvent deux états possibles, par exemple ici « 0 » ou « 1 » dans le cas de l'informatique (à noter que le système à deux états existait déjà depuis plusieurs décennies auparavant avec les automates, les orgues de barbarie et la télégraphie morse). On considère toutefois le plus souvent une suite de groupes de bits dans la mesure où on a souvent à coder des informations dont les unités représentent un choix de plus de deux états (ex : un chiffre), dont le nombre est le plus souvent pair (2 états pour un bit, 4 pour 2 bits, 8 avec 3 bits, etc.). Une série de 8 bits est souvent nécessaire pour un signal pour faciliter les traitements de l'information.

- La compression

Une compression de données est une opération qui consiste à transformer un signal numérique, autrement dit une suite de bits A en une suite de bits B plus courte, tout en contenant la même information et le fait qu'il faut le stocker sur un support dont la capacité de stockage est limitée, comme une clé USB. Ainsi, on peut souhaiter par exemple réduire un fichier dont la taille excède la capacité d'une clé USB. On peut vouloir économiser de la place disponible sur son disque dur. Le système d'exploitation Windows propose par exemple de compresser les fichiers qui n'ont pas été utilisés depuis longtemps. Il existe deux types de compression : la compression sans perte (ou réversible), qui permet de reconstituer le fichier d'origine avec une exactitude absolue et la compression avec perte (ou irréversible), qui est utilisée pour des images, des fichiers sons ou vidéos et dont on peut compresser un document de manière à ne perdre quelques détails, souvent indécelables par un humain et ainsi réduire parfois de façon considérable la taille du fichier

- L'interpolation

Opération mathématique qui consiste à déduire une valeur X des valeurs prises par des valeurs voisines. Pour cela, on prend en compte un nombre plus ou moins élevé de voisins et de fonction d'approximation plus ou moins complexe selon l'algorithme réalisé pour le calcul.
Par exemple, une valeur interpolée à 6 peut être encadrée par deux valeurs quantifiées de 4 à 8.

- Océrisation

Procédé technologique visant à accélérer la conversion d'une base de données physique en une base de données numériques par une reconnaissance automatique de caractères.

- Le signal

Un signal désigne un ensemble de valeurs prises par une information en fonction d'une variable, comme par exemple le temps qui peut avoir pour fonction une information sonore (ex : avec la

voix comme signal). On distingue par ailleurs le signal continu, qui dispose d'un signal linéaire et constant dans le temps, du signal discontinu qui dispose d'un taux de variation minimum.

Annexes

LISTE DES TABLEAUX :

Format ISO (ici 9660)	Largeur	Hauteur	Résolution d'échantillonnage (dpi)	Définition (Mpix)
A5	14,9 cm (5,87'')	21 cm (8,27'')	200	1,9
			300	4,4
			400	7,8
			600	17,5
A4	21 cm (8,27'')	29,7 cm (11,69'')	200	3,9
			300	8,7
			400	15,5
			600	34,8
A3	29,7 cm (11,69'')	42 cm (16,54'')	200	7,7
			300	17,4
			400	30,9
			600	69,6
A2	42 cm (16,54'')	59,4 cm (23,39'')	200	15,5
			300	34,8
			400	61,9
			600	139,2

Tableau 1 : Définitions d'images

Couleur	ultraviolets	violet	bleu	vert	jaune	orange	rouge	infrarouges
Longueur d'onde	< 400 nm	420	470	520	570	620	670	> 700 nm
		Couleurs visibles par l'œil humain						

Tableau 2 : Spectres électromagnétiques visibles par l'œil humain

FIGURE

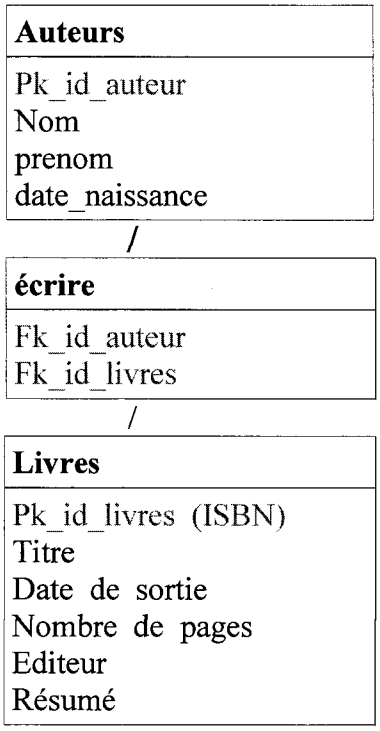

Figure : exemple de base de données

ILLUSTRATIONS

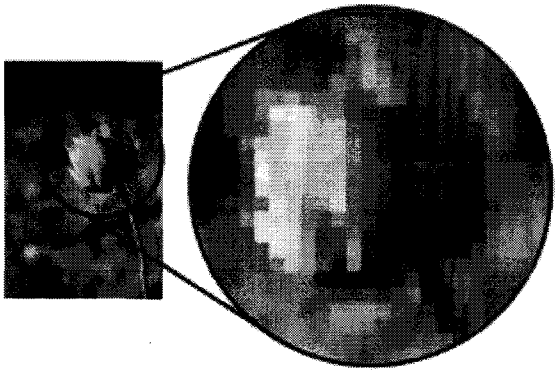

Illustration 1: exemple d'image matricielle

Illustration 2: exemple d'image vectorielle

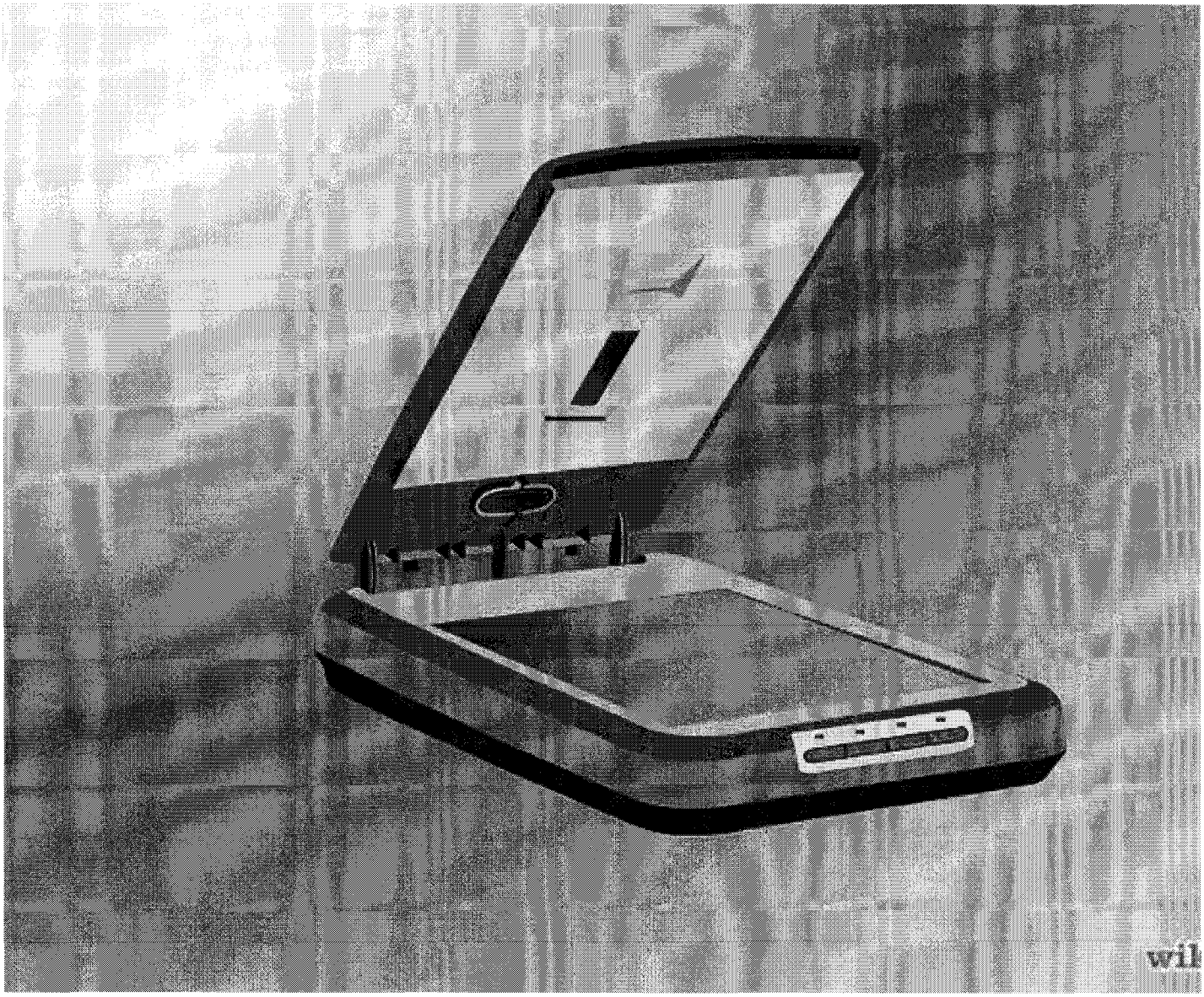

Illustration 3 : exemple de scanner à livre ouvert

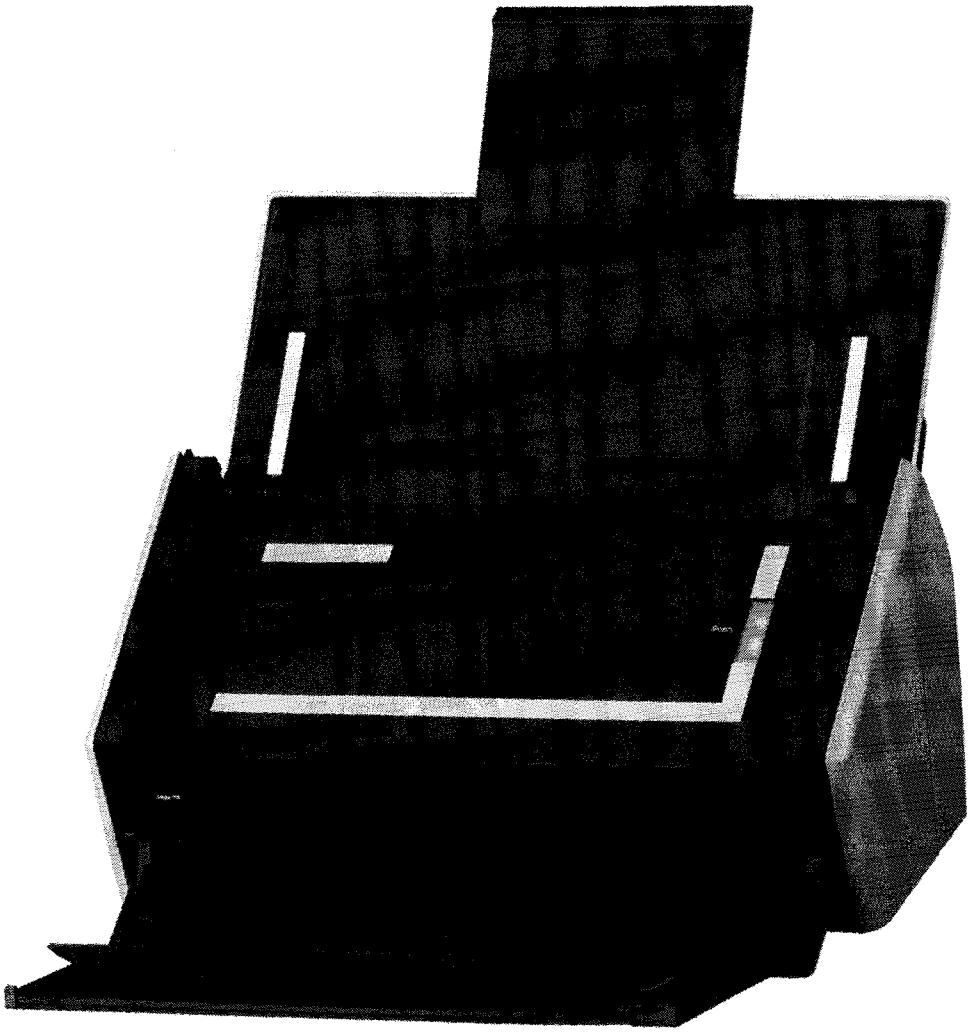

Illustration 4 : exemple de scanner à défilement

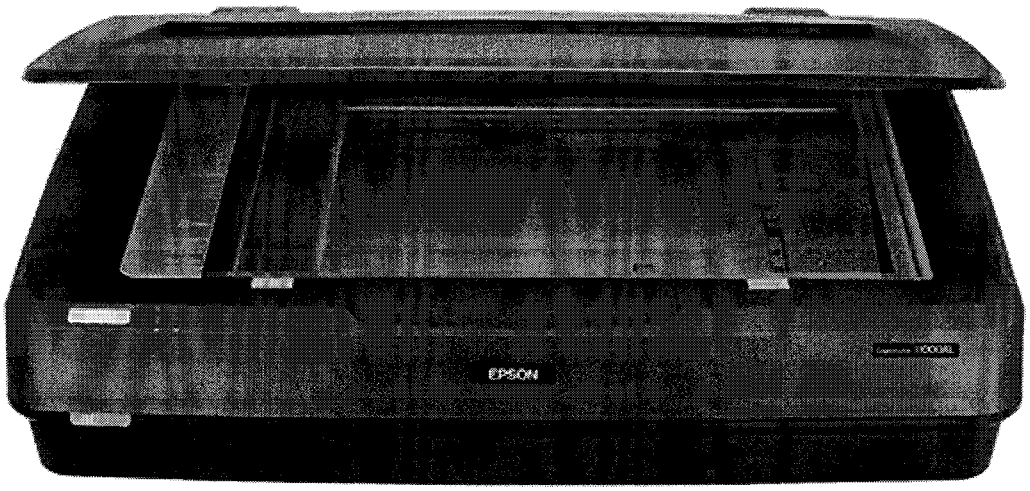

Illustration 5: exemple de scanner à plat

Résumé

Dans un contexte de développement accru de l'Internet et des demandes de la part des utilisateurs, la numérisation et la diffusion de l'information se doivent de répondre aux défis engendrés par ces développements. Pour cela, il convient au documentaliste et à l'archiviste de scanner l'information contenue dans le document papier à travers un scanner, utiliser des formats adaptés à la numérisation et à l'exportation des données, vérifier si les couleurs ne sont pas déséquilibrées au niveau de la disposition, ou si possible les supprimer si celles-ci devaient gêner la numérisation, faire attention à d'éventuelles erreurs de codage qui pourraient nuire à la lecture de l'information ainsi qu'à sa protection, et enfin à exporter ces informations à travers différents sites sous différents formats, cela bien entendu avec l'autorisation de l'auteur du texte contenants ces informations dans le cadre de la loi.
Par conséquent, la numérisation et la diffusion de l'information sont concernées par de multiples enjeux comme par exemple les enjeux stratégiques, juridiques et informatiques. Enjeux qui sont eux-mêmes liés à de multiples problèmes comme celle de la protection informatiques des données à travers le codage et le choix des formats, celle de la protection des droits d'auteur au moyen des lois les encadrant et enfin celle des liens par partenariat avec différents sites de diffusion de documents numérisés (comme par exemple la diffusion de thèses avec l'ABES (Agence bibliographique de l'enseignement supérieur)), ainsi que diverses bibliothèques (comme les bibliothèques universitaires) et centres de documentation .

Mots-clés : Bibliothèques, codage, couleur, diffusion, documentation, droit d'auteur, format, information, Internet, numérisation