h e g

Haute école de gestion
Genève

Conception d'un service de numérisation pour ArchiLab

Travail de master réalisé

par :

Anastase HATEGEKIMANA

Sous la direction de :

Basma MAKHLOUF SHABOU, Professeure HES et responsable du Master IS

Carouge, le 15 août 2022

Information documentaire
Haute École de Gestion de Genève (HEG-GE)

Hes·SO GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Remerciements

Au terme de ce travail, je voudrais adresser mes remerciements à toutes les personnes qui ont contribué à sa réalisation :

Basma Makhlouf Shabou, mandante et responsable du laboratoire ArchiLab, professeure et responsable du Master en sciences de l'information à la Haute école de gestion de Genève (HEG), pour la proposition de ce projet de mandat et pour l'encadrement fourni.

Aurèle Nicolet, collaborateur scientifique à la Haute école de gestion de Genève (HEG) pour l'encadrement de ce travail.

Les représentants des entreprises interviewées pour leur participation active à ce projet, en m'accordant des entretiens et en me fournissant d'autres informations dont j'avais besoin après les séances d'entretiens.

Jérôme Guisolan, archiviste auprès des Archives cantonales vaudoises (ACV), pour les informations fournies sur les entreprises prestataires de numérisation d'archives en Suisse.

Liliane Regamey, responsable de la Section utilisation auprès de la Bibliothèque nationale suisse (BN), pour les informations fournies sur les spécifications techniques que la BN exige à ses fournisseurs de numérisation d'archives.

Résumé

ArchiLab, le nouveau laboratoire archivistique de la Haute école de Gestion de Genève (HEG), envisage de mettre en place un service de numérisation ayant pour finalité de dynamiser les cours de la filière Information science, mais aussi d'étendre les prestations de service de la filière dans le domaine archivistique. C'est dans ce contexte que ce travail de master, effectué entre le 24 février et le 15 août 2022 et dans le cadre d'un mandat, a été proposé.

Il dresse un état des lieux des pratiques de numérisation des archives en vigueur dans certaines entreprises basées en Suisse romande et en Suisse alémanique. Nous avons mené des entretiens semi-directifs avec sept entreprises réparties en trois catégories du point de vue de la typologie des documents numérisés. Pour des raisons liées à l'anonymisation des données, ces entreprises sont symbolisées par des lettres alphabétiques A, B, C, D, E, F et G. Les entreprises A, B, C et D numérisent des collections et archives papier et iconographiques, l'entreprise E est mixte tandis que les entreprises F et G numérisent les archives documentaires classiques.

Pour chaque entreprise, nous avons analysé les éléments suivants : typologie des documents numérisés, services offerts, matériel de numérisation utilisé (scanners et logiciels), paramètres techniques de numérisation (résolution, profondeur des couleurs, formats de fichiers images, colorimétrie, type et niveau de compression, formats des métadonnées, formats de conservation…), étapes du workflow de numérisation, aspects juridiques, normatifs et qualitatifs de la numérisation, ainsi que les méthodes de de tarification.

Nous avons analysé tous ces éléments dans le but de concevoir un flux de travail (workflow complet) de numérisation pour le laboratoire ArchiLab. Pour les prix, nous avons privilégié l'approche analytique et comparative, c'est-à-dire savoir comment chaque entreprise facture ses prestations de numérisation et quelles sont les étapes de la chaîne de numérisation qui influencent, de manière significative, le prix total appliqué.

Cette étude montre la diversité du matériel de numérisation (scanners et logiciels) utilisé par les entreprises ; diversité qui s'explique surtout par le type de documents numérisés, mais aussi par les objectifs des projets de numérisations. Elle montre aussi l'importance des traitements d'avant et d'après numérisation dans la réussite de tout projet de numérisation d'archives.

Au-delà du résultat final proposé, cette étude souligne aussi la complexité d'un projet de numérisation puisqu'il est au carrefour de plusieurs disciplines : sciences de l'information documentaire, informatique, droit et management. Mais cette multidisciplinarité, qui doit naturellement caractériser une équipe de projet de numérisation, est aussi un des facteurs de sa réussite.

Mots-clés : archives, numérisation, numérisation patrimoniale, numérisation documentaire, workflow de numérisation, paramètres techniques de numérisation, scanner, logiciel, aspects normatifs de la numérisation, aspects juridiques de la numérisation, aspects qualitatifs et quantitatifs de la numérisation, document et image numérique, tarification.

Table des matières

Liste des tableaux

Liste des figures

1. Introduction

1.1 Contexte

Inauguré en juin 2021, ArchiLab est le premier laboratoire d'archivistique en Suisse romande, mis en place par la filière Information science de la Haute école de gestion (HEG) et avec le soutien de la direction. ArchiLab est structuré autour d'une salle de cours et d'un atelier de numérisation.

La salle de cours comprend 20 postes de travail avec des ordinateurs permettant de tester différents logiciels de gestion documentaire utilisés par les professionnels. L'atelier de numérisation comprend 6 postes de traitement, un scanner type Fujitsu fi-7600 et les logiciels Kofax Express et Acrobat professionnel pour la chaîne de numérisation (capture, contrôle de dossiers, indexation automatique par OCR, codes-barres…).

La finalité du laboratoire ArchiLab est de dynamiser les cours de la filière Information science et d'étendre les prestations de services de la filière dans le domaine archivistique, en offrant un espace sécurisé pour accueillir des fonds d'archives de diverses institutions et les traiter.

Les archives à traiter appartiennent au domaine privé (personnes physiques ou morales) ou public, et elles peuvent être acquises par **don***[1], **legs*** (disposition testamentaire), **dépôt provisoire*** ou par **partenariat*** institutionnel (mutualisation des ressources et des résultats produits). L'origine géographique de ces fonds peut se situer en Suisse ou ailleurs.

Il est prévu que :

> « *Pour chaque mode d'acquisition, une convention est obligatoirement rédigée et signée, stipulant les actions, rôles, droits et responsabilités des parties représentées, ainsi que l'acheminement, le traitement et le statut des archives, ainsi que la date d'acquisition et le délai de traitement. Un modèle standard est proposé selon le mode d'acquisition.* » (Makhlouf Shabou, 2021, p.9)

Par exemple, et d'après Makhlouf Shabou, « un dépôt provisoire peut impliquer une **rétrocession*** des archives. » Déposées dans un cadre strict de traitement à des fins d'enseignement, de recherche ou de mandat et ce pour une période convenue entre les parties signataires, les archives restent la propriété du déposant, qui se réserve le droit de définir « le transfert des responsabilités pour la période de traitement entre les parties », précise la même source.

1.2 Objectifs du mandat

ArchiLab souhaite mettre en place un service de numérisation sous forme de mandats effectués par des étudiants dans le cadre de cours ou par des membres du corps enseignant lors de prestations de service. Dans cette perspective, les objectifs de ce projet de mandat sont les suivants :

[1] Les mots suivis d'un astérisque figurent dans un glossaire de l'**annexe 1.**

a) Réaliser un état des lieux :

- Identifier les bonnes pratiques et les principaux services en matière de numérisation d'archives.
- Comparer les prestations des principaux services de Suisse afin de définir les tarifs.

b) Concevoir un portefeuille de services possibles pour ArchiLab :

- Proposer un flux de travail (workflow) complet pour chacun de ses services.
- Établir une tarification (pricing) en fonction des différents services proposés.

1.3 Problématique

Mettre en place un service de numérisation d'archives nécessite de prendre en compte un certain nombre de contraintes organisationnelles (intellectuelles), techniques et juridiques liées au statut du document numérique. Comprendre l'ampleur et l'étendue de chacune de ces contraintes constitue la clé de réussite d'un projet de numérisation dans sa globalité. Autrement dit, pour réussir une mise en place d'un service de numérisation d'archives, il faut bien planifier le processus de numérisation dans son ensemble, bien choisir les différents éléments techniques (scanners et logiciels), bien respecter les lois, bien appliquer les normes et standards et bien effectuer des contrôles qualitatifs et quantitatifs de tout le processus de numérisation.

1.3.1 Planification du processus de numérisation

La planification du processus de numérisation est l'essence même d'un projet de numérisation efficace. Pour la réussir, Claerr et Westeel (2011, p.137) conseillent de représenter le processus de numérisation sous forme d'un **workflow*[2]** ou enchaînement de tâches unitaires. Et pour pouvoir mettre en place ce workflow, les deux auteurs conseillent au responsable de projet de trouver d'abord des réponses à une série de questions soulevées par chacune des étapes unitaires de ce workflow de numérisation telles que récapitulées dans le **tableau 1**. Trouver de bonnes réponses à ces questions est une tâche difficile et suppose une bonne entente et une étroite collaboration entre le prestataire et le client (discussions en amont du projet).

La Bibliothèque et Archives nationales du Québec (BAnQ), la Bibliothèque nationale de France (BnF) et le Musée canadien de l'histoire (MCH) insistent, eux-aussi, sur l'importance de la planification du processus de numérisation :

> « Plusieurs étapes précédant (catalogage, préparation physique des documents) ou suivant la numérisation (contrôle de qualité, versement sur des serveurs, archivage et diffusion) sont indispensables et doivent être considérées avant d'entamer un programme de numérisation. » (BnQ, BnF, MCH, 2014, p.6)

Pour l'étape de transport et de stockage des documents par exemple, la pratique en vigueur dans les entreprises veut que ce soit le prestataire de numérisation qui s'en charge. Se pose alors la question de savoir sous quelles conditions (moyen de transport et sa sécurité, mode

[2] Voir plus de détail à la section 2.4.

de conditionnement particulier) les documents doivent être acheminés et stockés. Sans oublier la question concernant la méthode de facturation. Pour l'étape de vérification (de la complétude et de l'état physique) des documents reçus par le prestataire, Claerr et Westeel (2011, p.130) recommandent la présence du client afin d'éviter toute litige, car selon eux, « le bordereau de livraison, signé par chacune des parties, matérialise leur accord réciproque. » Mais alors, comment faire si le client ne peut pas être présent ou être représenté pour diverses raisons ?

Pour l'étape de préparation des documents (manutention), étant donné les difficultés que peut poser la manipulation de certains d'entre eux (fragilité, présence de l'encre trop pâle, etc.), le client est tenu de préciser au prestataire le soin dont ses opérateurs-trices devront faire preuve (port de gants par exemple) ainsi que toute manipulation non autorisée lors de la numérisation. La question concernant les ressources humaines doit aussi être posée : il s'agit de savoir si l'équipe de projet possède les compétences requises et le matériel nécessaire pour mener à bien toutes les étapes du processus de numérisation.

La réponse à chacune des questions des **tableaux 1 et 2** « influence les prestations à effectuer, la qualité attendue, le seuil de qualité minimale ou encore la gestion du contrôle qualité » (Brault, 2021, p.18).

Tableau 1 : Liste de questions à (se) poser en vue de mettre en place un workflow de numérisation fonctionnelle

Étape	Questions
1) Transporter les documents physiques	Quel est le délai d'indisponibilité souhaité des documents ?
	Est-il possible de faire sortir les documents de l'institution ?
	Y a-t-il une priorité dans l'ordre de numérisation ?
	Est-il possible de regrouper les documents en lots présentant des caractéristiques homogènes, que ce soit en termes de formats (pour favoriser la prise de vue) ou de contenu (pour favoriser l'OCR et l'indexation) ?
	A-t-on prévu d'organiser les transports en lots, de quelle taille, à quelle fréquence ?
	Qui est chargé d'identifier les documents qui partent vers la numérisation, de les empaqueter, de créer une liste de départ, d'identifier leur emplacement d'origine ?
	Faut-il souscrire une assurance spécifique au projet de numérisation ?
2) Stocker les documents physiques	Combien de documents pourront être stockés à proximité du numériseur ? Pendant combien de temps ?
	Quelles infrastructures sont prévues pour le stockage temporaire ?
	Comment sait-on où se trouve un document donné ?
3) Préparer	Existe-t-il des sources de données décrivant les documents à numériser (catalogues, index, etc.) ?
	Quels sont les liens entre les documents et ces descriptions ?
	Dans quel état sont les documents à numériser ?
	Faut-il prévoir une étape de dépoussiérage, ne serait-ce que pour préserver le numériseur et par respect des opérateurs ?
	Y a-t-il des pages à déplier, à aplatir, à découper ?
	Doit-on prévoir une restauration des documents ?
	Peut-on, au contraire, massicoter les documents ?
	Doit-on numériser une mire de contrôle avant chaque nouveau document ?
4) Numériser	Quel est le format des originaux ? Quel est le nombre de pages par document ?
	Quelle va être l'utilisation des images créées, en termes de diffusion et de préservation ?
	Vaut-il mieux numériser en mode double page, quitte à créer ensuite un fichier de diffusion monopage ?
	L'utilisation de caches sera-t-elle nécessaire ?
	Quels formats d'image/résolution/couleur seront nécessaires pour ces usages ?

(Tableau construit sur la base du texte de Claerr et Westeel, 2011, pp.137-139)

Tableau 2 : Liste de questions à (se) poser en vue de mettre en place un workflow de numérisation fonctionnel (suite)

Étape	Questions
5) Où va-t-on stocker les fichiers produits ?	Y a-t-il un moyen univoque d'identifier chaque document numérique et de le relier à l'original ?
	Comment est-il possible de récupérer les métadonnées techniques de la capture d'image (EXIFF, fichier log) ?
	Quels sont les critères qualité (mesurables objectivement) souhaités pour les images ?
	Comment va-t-on les mesurer ?
6) Restituer	Mêmes questions que pour le transport
	Qui vérifie l'état des documents après numérisation ?
	Contrôle-t-on que tout document rendu a bien été numérisé ?
7) Indexer/Océriser	Quel usage des documents numériques est prévu pour les dix années qui viennent ?
	Quels sont les publics attendus ?
	Un partage des ressources numériques est-il prévu avec d'autres institutions ? Si oui, des formats communs de métadonnées, d'images, d'OCR sont-ils définis ?
	Y a-t-il certains formats préférés par l'outil de diffusion choisi ?
	Dispose-t-on d'outils OCR capables de générer les formats attendus ?
8) Sauvegarder	Quel est le poids, par page, des différents fichiers produits : texte, images, index, etc. ? Sur cette base, un simple tableau reprenant les volumes unitaires et les quantités prévues permettra d'estimer le volume total de données ainsi que son accroissement éventuel (hebdomadaire, mensuel, annuel…).
	Comment nomme-t-on les répertoires et fichiers devant être échangés, publiés, stockés sur le long terme ?
	Une politique de sauvegarde est-elle en place ?
	Doit-on tout conserver en ligne ?
	Quelle est l'infrastructure informatique pour la publication de données sur Internet, sur Intranet, pour l'archivage numérique ?
9) Contrôler	Qui réalise les contrôles-qualité sur les différents éléments : originaux, images, index, texte reconnu, différents formats de fichiers produits ?
	Un contrôle par échantillonnage, type ISO 2859-1, est-il applicable ?
10) Publier	Comment vont être réalisées les publications ?
	Comment va être réalisée, ensuite, l'alimentation du système de publication, par exemple en cas de correction d'une ressource, ou ajout d'un nouveau document numérique ?
	Un système de suivi des consultations est-il prévu ?
	Un mécanisme de retour des utilisateurs, quant à la qualité perçue des données, est-il prévu ?
	Quels sont les droits dont on dispose en termes de diffusion ?
	Doit-on demander une création de microfiches/microfilms, à partir des images numérisées, pour servir de support de préservation à long terme, qui soit indépendant des problématiques de migration de format ?

(Tableau construit sur la base du texte de Claerr et Westeel, 2011, pp.137-139)

1.3.2 Les aspects techniques de la numérisation

Les aspects techniques de la numérisation concernent les scanners, les logiciels et les paramètres de numérisation. Différents scanners et logiciels adaptés à chaque type de document à numériser existent déjà sur le marché (Mkadmi, 2021, pp. 16-19) :

- Scanners à plat pour les documents imprimés non reliés et les feuilles manuscrites ;
- Scanners verticaux pour les documents reliés, documents fragiles, documents de grand format ;
- Scanners avec chargeur pour les pages séparés de même format ;
- Scanners à tambour pour les documents fragiles (ils coûtent plus chers) ;
- Scanners à livres pour les documents reliés ;

- Scanners avec appareil photo numérique pour les objets en 3D et les documents fragiles ;
- Scanners à transparents pour la numérisation des négatifs, diapositives, microformes, microfilm, etc.

A cette liste, Essevaz-Roulet (2016, pp.54-57) y ajoute le scanner de production réservé au monde professionnel et très pratique pour numériser les documents à feuilles volantes de grammage et de format standard, et le scanner de plans.

Le choix d'un scanner doit être minutieux car celui-ci est souvent associé avec le logiciel qui le pilote et ils sont donc vendus ensemble. Ce choix doit être guidé par un certain nombre de critères dont les plus importants sont (Essevaz-Roulet, 2016, pp.52-53) :

- le format des documents à numériser ;
- les caractéristiques des **lots*** à numériser ;
- le degré de fidélité à l'original exigé ;
- la résolution optique maximale et ;
- le budget disponible.

A cette liste, Mkadmi (2021, p.19) ajoute deux autres critères, à savoir **la vitesse de numérisation** (exprimée en nombre de pages par minute) et le **type d'acquisition ou colorimétrie** (noir et blanc, niveaux de gris, couleur).

Pour le Bureau de coopération interuniversitaire (BCI, 2014, p.20) au Canada, le choix d'un scanner doit être guidé par les critères suivants :

- volume des pages à numériser ;
- taille des documents ;
- vitesse de traitement ;
- présence de plusieurs fonctionnalités ;
- alimentation automatique ;
- possibilité de numériser en recto-verso ;
- degré de facilité pour numériser les documents fragiles ou déchirés ;
- présence de résolution-couleurs ;
- et prix.

La Direction générale des systèmes d'information (DGSI) du canton de Genève préconise des modèles de scanners équipés de dispositifs VRS (Virtual ReScan) d'amélioration des images qui, de manière automatique, permettent de

> « *supprimer les pages blanches, redresser les pages, améliorer le contraste et la luminosité, détecter et si besoin conserver la couleur, affiner les écritures et caractères pour une meilleure lecture OCR, nettoyer le bruit, les fonds des pages et les perforations sur les documents.* » (DGSI, 2015, p.12)

Concernant le choix d'**un logiciel**, le BCI (2014, p.20) conseille de prendre en compte les caractéristiques suivantes :

- fiabilité et maturité technologique (répond-il aux besoins techniques de la numérisation ?) ;
- convivialité (facile à utiliser et à comprendre par les utilisateurs) ;
- et la possibilité de l'adapter aux besoins spécifiques (présence de nombreuses options).

La Bibliothèque et Archives nationales du Québec (BAnQ, 2019, p.12) conseille, quant à elle, de considérer les caractéristiques suivantes :

- sa fiabilité ou son degré de perfectionnement technologique ;
- son interface et sa facilité d'utilisation ;
- ses possibilités d'adaptation à des besoins spécifiques ;
- les formats offerts pour l'impression et pour l'exportation : TIFF, PDF/A-1, PDF, JPEG 2000, JPEG, etc. ;
- les fonctions de traitement par lots et de macros pour des opérations répétitives ;
- les options et la fiabilité du module de reconnaissance optique de caractères (OCR ou ROC) ;
- l'enregistrement de l'espace de travail afin de reprendre un travail en cours ;
- les fonctions intégrées de gestion d'images ou de métadonnées ;
- le réglage de la luminosité, du contraste, des blancs, de l'échelle de gris, des couleurs et de la plage dynamique ;
- la possibilité de sauvegarder les réglages.

Outre le scanner et son logiciel de capture, c'est aussi et surtout les paramètres techniques de numérisation et leur réglage qu'il faut bien comprendre et choisir, car ils influent significativement sur la qualité de l'image numérisée, ainsi que sur la qualité de l'OCR. D'après la Direction des Archives de France (DAF, 2008, pp.31-36), les principaux paramètres techniques de numérisation à contrôler sont[3] : la résolution optique, la colorimétrie (noir et blanc, niveaux de gris, couleurs RVB, couleurs CIE Lab[4]), la profondeur des couleurs, le profil colorimétrique, le format de fichier image (pour la conservation et la diffusion), le type et le taux de compression, le cadrage-orientation des zones à numériser (document entier ou zones délimitées), et les traitements d'images post-numérisation.

1.3.3 Les aspects normatifs de la numérisation

Par rapport aux aspects juridiques, les aspects normatifs de la numérisation (normes techniques) ne sont en principe pas contraignants, car les normes sont émises par des organisations de droit privé, c'est-à-dire non dotées d'un quelconque pourvoir législatif (VSS, 2013, p.2). Mais ils influent sur l'image du prestataire de numérisation dans la mesure où le fait d'être certifié pour telle ou telle norme de qualité ou de sécurité par exemple, peut renforcer sa notoriété vis-à-vis des clients potentiels.

Parmi les normes de certification pour la qualité, la sécurité, l'intégrité et la pérennisation de l'information, on peut citer, entre autres, l'ISO 9001 pour le système de management de la qualité (SMQ), l'ISO 27001 pour la sécurité du système d'information, l'ISO 14641-1 pour l'archivage électronique probant et pérenne, l'ISO 19264-1 pour la qualité des documents numérisés et la norme AFNOR NF Z42-026 pour la **copie fidèle***.

Nous verrons plus loin qu'à cause de leur caractère non contraignant, ces aspects normatifs sont gérés différemment par les entreprises prestataires de numérisation.

[3] Voir le détail à la section 2.5
[4] Voir signification à la section 2.5.3

1.3.4 Les aspects juridiques de la numérisation

Les implications juridiques du document numérique sont nombreuses et concernent les droits physiques ou intellectuels des personnes (droit d'auteur, droit de propriété, droit de protection de la vie privée, droit d'accès à l'information) ainsi que la valeur probante dudit document, c'est-à-dire son équivalence fonctionnelle par rapport à son original papier.

Les droits d'auteurs peuvent être moraux (droit de divulgation, droit à la paternité, droit au respect de l'œuvre) ou patrimoniaux (droit de reproduction, droit de représentation). Alors que les droits moraux sont inaliénables, les droits patrimoniaux sont cessibles, de manière contractuelle, par l'auteur-e ou ses ayants droit. Les droits d'auteurs restent en vigueur jusqu'à la mort de l'auteur-e et pendant 50 ans ou 70 ans à partir de la date de décès au profit de ses ayants droit (art.29, LDA). La diffusion d'une œuvre numérisée est régie par les droits voisins, en particulier le droit à l'image (droit de la personnalité) qui oblige tout diffuseur d'images à demander une autorisation de la personne concernée. Si celle-ci est décédée, la loi interdit toute diffusion d'image susceptible de porter atteinte à la dignité de la personne humaine.

Tout ceci pour dire qu'avant de numériser un document, le prestataire doit d'abord vérifier si celui-ci est protégé par le droit d'auteur ou de ses ayants droit. Si c'est le cas, le prestataire doit disposer de l'autorisation du titulaire des droits de reproduction. En Suisse, le droit d'auteur est protégé par la **Loi fédérale sur le droit d'auteur et les droits voisins (Loi sur le droit d'auteur, LDA)** du 9 octobre 1992.

La **protection des données personnelles** est garantie par les dispositions de la Loi fédérale sur la protection des données (LPD) du 19 juin 1992 et les différentes lois cantonales sur la protection des données. A Genève par exemple, c'est la **LIPAD (Loi sur l'information du public, l'accès aux documents et la protection des données personnelles)** du 5 octobre 2001, ainsi que de son règlement d'application (RIPAD) du 21 décembre 2011. Dans les autres cantons romands, où sont basées des entreprises ayant participé à cette étude, ces législations sont :

- Canton du Valais : **Loi du 9 octobre 2008 sur l'information du public, la protection des données et l'archivage** (LIPDA, 170.2) ;
- Canton de Vaud : **Loi du 11 septembre 2007 sur la protection des données personnelles** (LPrD, 172.65) ;
- Canton de Berne : **Loi du 19 février 1986 sur la protection des données** (LCPD, 152.04).

La mise en place d'un service de numérisation d'archives doit aussi respecter les différentes lois cantonales sur l'archivage en ce qui concerne l'accès aux archives publiques (**délais de protection à observer**). Les délais de protection, inscrits dans les inventaires, sont des durées prévues par les bases légales (au niveau fédéral et cantonal) durant lesquelles les documents archivés ne sont pas accessibles au public. A Genève par exemple, la loi sur l'archivage prévoit un délai général de protection de 25 ans après la clôture du dossier. Pour les documents contenant les données personnelles sensibles ou des profils de la personnalité, la loi prévoit un délai supplémentaire de 10 ans après la mort de la personne, soit 35 ans de protection.

De manière générale, les bases légales pour l'archivage électronique en Suisse sont fondées sur le droit commercial, principalement sur **l'Ordonnance concernant la tenue et la conservation des livres de comptes (Olico)** du 24 avril 2002. Celle-ci définit les exigences techniques auxquelles doit satisfaire la conservation de documents électroniques, particulièrement importants afin qu'ils puissent avoir la même **valeur probante*** que leurs originaux papiers. Ces règles de conformité sont : intégrité/authenticité/fiabilité (art.3), disponibilité (art.6), documentation (art.4), devoir de diligence (art.5) et organisation de la gestion documentaire (art.7). La valeur probante d'un document électronique repose donc sur le respect de ces exigences de qualité. C'est pourquoi l'ordonnance Olico insiste particulièrement sur l'importance des processus documentés (de numérisation et de conservation), c'est-à-dire sur la constitution et la maintenance de la documentation de la plate-forme de **dématérialisation*** de chaque organisation. Depuis le 1er janvier 2011, les dispositions de l'Olico ont été renforcées par le nouveau Code de procédure civile (CPC) unifié, particulièrement en ses articles 177-180 (Dunant-Gonzenbach et Droze, 2016, p.5).

Du point de vue administratif, le processus de numérisation

> « *doit garantir aux documents numériques leur fiabilité, intégrité, authenticité, pérennité, exploitabilité et traçabilité pour que ceux-ci puissent conserver leur* ***valeur probante*** *et être admissibles lors d'une procédure.* » (DGSI, 2015, p.21)

1.3.5 Les aspects qualitatifs et quantitatifs de la numérisation

Le contrôle de la qualité du contenu des documents numérisés consiste à s'assurer que la reproduction numérique est conforme aux documents sources. Le contrôle doit aussi porter sur l'ensemble des produits de la chaîne de numérisation tels que le fichier de récolement, d'indexation, etc. Ces contrôles ont lieu chez le prestataire durant le processus de numérisation et à la fin de celui-ci, mais aussi chez le client lors de la restitution des documents par le prestataire. Ces contrôles peuvent prendre plusieurs formes (**tableau 3**).

Les contrôles visuels portent généralement sur les critères d'appréciation subjectifs pour lesquels la prise de mesure seule ne peut être suffisante, comme la couleur, le cadrage et la lisibilité, la netteté et le contraste, et l'exactitude de l'indexation du contenu d'une image numérique. Le contrôle par mesure peut se faire manuellement en utilisant des logiciels adéquats. Mais si le volume des documents numérisés est important, ce type de contrôle peut être effectué sur des échantillons.

Tableau 3 : Méthodes de contrôle qualité possibles et leurs caractéristiques

Méthode	Avantages	Inconvénients	Remarques
Contrôle exhaustif	Exhaustivité	Coûteux et long	Faible volumétrie, ou en phase de test ou automatisé.
Contrôle par l'échantillonnage	Peut reposer sur **une norme**[5] (dans ce cas, l'appliquer rigoureusement).	Suivi précis et régulier du respect de la norme : calcul de la taille de l'échantillon selon les **NQA** (niveau de qualité acceptable), taille des lots à contrôler, et les changements de niveau de contrôle (renforcé, normal, réduit) selon le nombre de lots admis ou rejetés.	Projets à volumétrie de moyenne à forte et de supports identiques ; contrôle renforcé en phase de test.
Mise à disposition d'outils de contrôle calibrés	Investissements très limités ; s'assurer de la maintenance et du calibrage par la société.	Maîtrise limitée sur les outils. Augmentation des coûts.	CCTP (Cahier des clauses techniques particulières) rigoureux sur les besoins et les critères techniques ; limiter la demande.
Audits de la chaîne du prestataire (contrôle du respect des engagements qualité et des méthodes de production, en amont des livraisons)	Anticipe sur les reprises à faire et donc réduit les coûts. Englobe l'ensemble des paramètres de production et pas uniquement le produit fini. C'est un dialogue avec le prestataire pour l'amélioration continue.	Lourd à mettre en place et à mener. Pour être crédible et efficace, elle doit s'appuyer sur des méthodes standardisées et des documents contractuels (plan assurance qualité incluant des procédures et des indicateurs qualité, relevés régulièrement).	Pour tous projets. Ne se substitue pas au contrôle du prestataire sur sa production. Implique de la régularité et la rédaction de rapports d'audits validés par les deux parties, la mise en œuvre des recommandations d'un audit est vérifiée à l'audit suivant.

(Mocellin, 2010, p.36)

Claerr et Westeel (2011, pp.277-278) donnent une liste non exhaustive de critères de contrôle cités dans huit appels d'offres publiés en France par des centres d'archives départementaux et des bibliothèques. Dans la liste qui suit, les erreurs d'imprécisions signalées comme majeures dans les cahiers des charges sont soulignées.

a) Indexation et organisation des données

- Nommage des fichiers ;
- Structure des données (répertoires…) ;
- Cohérence de la nomenclature avec le fichier de récolement ;
- Page manquante ;
- Nombre de fichiers livrés ;

[5] Norme NF-ISO 2859-1 : règles d'échantillonnage pour les contrôles par attributs.

- Ouverture, décompression des fichiers et supports.

b) Prise de vue et spécifications de base

- Information tronquée ;
- Complétude intellectuelle ;
- Cadrage des images ;
- Distorsion géométrique non présente sur le document original ;
- Travers supérieurs à + ou – 1°30', image inclinée par rapport à son axe ;
- Sens de lecture, orientation ;
- Niveau d'accentuation ;
- Format des fichiers (signalé par un appel d'offre sur huit);
- Mauvaise résolution.

c) Couleur

- Point blanc, point noir, balance des gris ;
- Gamma ;
- Dérive entre le document d'origine et sa représentation numérisée ;
- Dérive chromatique ;
- Halos sur les contours des images ;
- Ombres portées, et présence d'éléments étrangers au document ;
- Contraste et luminosité inadaptés.

d) Autres éléments subjectifs

- Qualité intrinsèque des images et de leur présentation ;
- Lisibilité ;
- Netteté insuffisante.

Les auteurs précisent que la plupart de ces critères souffrent du manque de **seuils de conformité** (ou seuils d'acceptation de non-conformité), d'**unités** et d'**informations méthodologiques**. Ils soulignent aussi

> « *le caractère très subjectif de certains de ces contrôles, qui demanderait quelques précisions comme celui de la "Qualité intrinsèque des images et de leur présentation" pourtant relevé dans quatre CCTP (cahiers des clauses techniques particulières) sur huit*. » (Claerr et Westeel, 2011, pp.277)

Claerr et Westeel constatent aussi que la plupart des appels d'offres en numérisation ne mentionnent pas « le seuil relatif à la quantité de fichiers non conformes ou présentant des défauts tolérables dans une population d'images donnée et qui déterminerait la validité ou le rejet du **lot***.» (p.280). Sur huit appels d'offres, seulement trois le mentionnent. Concernant la méthode d'échantillonnage, Claerr et Westeel soulignent l'usage largement répandu de la **norme ISO 2859-1** dans les projets de numérisation.

La norme ISO 2859-1 (ancienne NF X 06-022 :1991) a été élaborée par le comité technique ISO/TEC 69, *Application des méthodes statistiques*, sous-comité SC 5, *Échantillonnage en vue d'acceptation*. La première version date de 1989 tandis que la deuxième a été publiée en 1999. Cette norme spécifie un **système d'échantillonnage par attributs** pouvant être effectués dans différents domaines : produits finis, composants ou matières premières, opérations, matériels en cours de fabrication, fournitures en stock, opérations d'entretien, informations ou enregistrements, procédures administratives.

Le but de la norme ISO 2859-1 est de :

> « *contraindre un fournisseur à maintenir la qualité moyenne de la fabrication à un niveau au moins égal au niveau de qualité acceptable spécifié, tout en limitant le risque pour le client d'accepter occasionnellement un lot défectueux.* » (ISO, 1999, p.1)

L'ISO 2859-1 comprend plusieurs **tableaux des données** : **plans d'échantillonnage**, **risque du fournisseur**, **qualité du risque client**, **lettre-code d'effectif d'échantillon**, etc. Chaque plan d'échantillonnage se décline en échantillonnage simple, double ou multiple, qui à son tour, peut être effectué en contrôle normal, renforcé ou réduit.

A partir d'un cas concret, Claerr et Westeel (2011) montrent comment utiliser les abaques de la norme ISO 2859-1 pour établir un plan d'échantillonnage qui permet de déterminer l'effectif de l'échantillon à contrôler.

Rappelons qu'au départ, le client et le prestataire doivent convenir d'un **niveau de qualité acceptable (NQA)**. « *Les caractéristiques des images qui sont contrôlées et la manière dont une image est qualifiée de défectueuse doivent naturellement être définies avec précision.* », souligne la Direction des Archives de France (DAF, 2008, p.48).

Une Bibliothèque B, ici client, veut numériser, par un prestataire externe, un lot de 20 000 documents avec un **seuil d'acceptation de non conforme de 0.15 %.** Le **tableau 4** montre que l'effectif de l'échantillon à contrôler est de **315**, car le nombre 20 000 se trouve dans l'intervalle 10 001-35 000, et coïncide avec la lettre-code **M**. L'**annexe 8** donne une autre présentation de ces deux tableaux en incluant le NQA.

Tableau 4 : Détermination de l'effectif de l'échantillon que la Bibliothèque B doit constituer à partir d'un extrait du tableau 1 de la norme ISO 2859-1

Effectif du lot	Lettre code	Effectif de l'échantillon
501 à 1 200	J	80
1 201 à 3 200	K	125
3 201 à 10 000	L	200
10 001 à 35 000	**M**	**315**
35 001 à 150 000	N	500
150 001 à 500 000	P	800

(Claerr et Westeel, 2011, p.284)

D'après le **tableau 5**, le seuil d'acceptation de non conforme de **0.15 %** se trouve à l'intersection de la **lettre-code M** (et donc de l'effectif de l'échantillon de 315 individus) et des seuils d'acceptation (Ac) et de rejet (Re) respectivement de 1 et de 2.

Tableau 5 : Détermination des seuils d'acceptation Ac et de rejet Re en fonction de la lette-code, de l'effectif de l'échantillon et du seuil d'acceptation de non conforme (extrait du tableau 2A de la norme ISO 2859-1)

Lettre code	N	Ac =0 Re=1	**Ac =1 Re=2**	Ac =2 Re=3	Ac =3 Re=4	Ac =5 Re=6	Ac =7 Re=8	Ac =10 Re=11	Ac =14 Re=15	Ac =21 Re=22
J	80	0,15	0,65	1	1,5	2,5	4	6,5	10	
K	125	0,1	0,4	0,65	1	1,5	2,5	4	6,5	10
L	200	0,065	0,25	0,4	0,65	1	1,5	2,5	4	6,5
M	**315**	0,04	**0,15**	0,25	0,4	0,65	1	1,5	2,5	4
N	500	0,025	0,1	0,15	0,25	0,4	0,65	1	1,5	2,5
P	800	0,015	0,065	0,1	0,15	0,25	0,4	0,65	1	1,5

(Claerr et Westeel, 2011, p.284)

Le **plan d'échantillonnage** que la Bibliothèque B peut dresser à partir de ces deux tableaux est le suivant :

- Type de contrôle à effectuer : individus non conformes par comptage ;
- Type de prélèvement : échantillonnage simple ;
- Effectif de l'échantillon à contrôler : 315 ;
- Seuil d'acceptation (Ac) = 1 ;
- Seuil de rejet (Re) = 2.

Concrètement, la Bibliothèque B doit contrôler 315 fichiers parmi les 20 000 fichiers qui lui seront livrés par le prestataire. Le lot en question sera validé si le nombre de fichiers non-conformes est inférieur ou égal à 1 (seuil d'acceptation). Ce lot sera rejeté si le nombre de fichiers non-conformes est supérieur ou égal à 2 (seuil de rejet).

Claerr et Westeel précisent les conditions dans lesquelles les plans d'échantillonnage selon la norme ISO 2859-1 doivent être réalisées (p.282)[6] :

- Une population (ici le lot) homogène et individus nombreux ;
- Pas d'inférence du taux qualité sur le lot : il faut vérifier si celui-ci a de grandes probabilités de présenter un taux qualité donné à l'avance ;
- En cas de rejet d'un lot, corriger l'échantillon ne suffit pas. Il faut reparcourir et corriger, ou reproduire l'ensemble du lot ;
- Appliquer, comme le prévoit la norme, les règles de passage en contrôle réduit et renforcé.

En résumé, le type de contrôle de la qualité des images numérisées et son degré d'exhaustivité doit, selon le Bureau de Coopération interuniversitaire du Canada (BCI), dépendre de la finalité de la numérisation :

> « Par exemple, *une numérisation de substitution requiert un contrôle exhaustif de la quantité et de la qualité pendant la numérisation et par échantillonnage à la fin de celle-ci. Alors qu'une numérisation de diffusion se contente d'un contrôle par échantillonnage sur un certain nombre de documents numérisés, effectué généralement après numérisation*. » (BCI, 2014, p.20)

[6] Les auteurs précisent qu'à « l'origine, ce type de contrôle sert à vérifier des productions industrielles. »

La reconnaissance optique de caractères (OCR) est un autre enjeu qualitatif dont l'engagement par le prestataire de numérisation doit être pris « après examen des corpus candidats à la numérisation. » (Tessier, 2010, p.51). En effet, le mode d'impression de la page, les variations typographiques internes, les alphabets et graphies syllabaires, les langues, etc., exercent une influence sur la faisabilité et la qualité de l'OCR. En d'autres termes, les difficultés de l'océrisation sont multiples : la qualité de l'image numérisée (qui doit être suffisamment redressée et contrastée), l'état du document original (défauts d'impression, dégradation de l'encre, éventuelles lacunes), logiciels trop orientés autour de la reconnaissance de l'alphabet latin rendant ainsi difficile la transcription d'autres alphabets, etc. Tous ces facteurs « rendent difficile la reconnaissance de caractères pour le logiciel comme pour l'opérateur et entravent le processus. » (Brault, 2021, p.27).

2. Concepts et définitions

2.1. Archives : définition et typologie

Le mot "archives" fait référence à la fois aux documents et à l'institution qui les abrite. En tant que documents,

> « *les archives sont l'ensemble de documents, quels que soient leur date, leur forme et leur support matériel (imprimé, audiovisuel, sonore ou électronique) produits ou reçus et conservés par toute personne physique ou morale ou par tout service ou organisme public ou privé dans l'exercice de son activité.* » (AMUE, 2010, p.72)

Selon le cycle de vie du document, l'archivistique définit trois âges des archives (AEG, 2010, p.1) :

- archives courantes ou documents qui sont d'utilisation habituelle et fréquente pour le traitement des affaires courantes ;
- archives intermédiaires ou documents qui doivent être conservés temporairement pour des besoins administratifs et juridiques ;
- et archives définitives ou documents qui ne sont plus susceptibles d'élimination et conservés pour les besoins de la gestion, de la justification des droits des personnes physiques ou morales et pour la documentation historique de la recherche.

D'après Gueit-Montchal (2020, p.329), « Le mot "archives" est couramment employé dans le sens restrictif de documents ayant fait l'objet d'un archivage, par opposition aux archives courantes. » Essevaz-Roulet (2016, p.14), lui, classe les archives en deux catégories de document : les documents iconographiques ou figurés (dessins, estampes, photographies, gravures, lithographies, affiches, cartes générales, géographiques, administratives, militaires, plans d'architecte…) et les documents textuels (livres, publications en générales, presse, correspondance, documents administratifs, commerciaux, comptables). Tout en précisant qu'il n'y a pas de frontière nette entre les deux catégories puisqu'un même document peut contenir des images et du texte. Ce qui est important selon lui, c'est de savoir si un document contient une majorité d'images ou une majorité de texte, car cette information, précise l'auteur, « est pertinente dans le contexte de la numérisation puisque les spécifications techniques, mais surtout les objectifs, diffèrent. » (p.14). Alors que l'information contenue dans un document iconographique tient au document dans son

ensemble, dans un document textuel, elle correspond aux caractères qui composent le texte. Ainsi, l'intérêt d'un document textuel

> « *n'est pas son image, mais la capacité qu'il offre à lire et à interpréter l'information portée. La numérisation de documents textuels a pour objectif de conserver cette lisibilité, voire de l'améliorer. Elle ne subit pas les mêmes contraintes que pour la numérisation d'une image.* » (Essevaz-Roulet, 2016, p.14)

2.2. Document numérique vs image numérique

Un document numérique ou électronique est un fichier informatique contenant des données codées. Il peut être un fichier texte, un fichier son, une image, une vidéo ou un programme logiciel. L'utilisateur peut accéder à ces données grâce à un ordinateur ou un appareil électronique capable de les interpréter.

Le document électronique est une succession des chiffres « 0 » et « 1 » qui, pour pouvoir coder une information, doivent être groupés par 8. Un « 0 » ou un « 1 » est un bit, un groupe de 8 bits forme un « octet » et chaque octet code un type d'information. Par exemple, le chiffre 1 correspond à l'octet « 00110001 ».

Une image numérique (ou image matricielle) est une collection de petits carrés élémentaires régulièrement ordonnés appelés pixels, disposés dans une grille rectangulaire appelée « bitmap ». Plus le nombre de pixels est grand (donc pixels plus petits), meilleure est la qualité de l'image. Un pixel n'ayant pas de taille fixe, il est convenu de définir la taille d'une image (sa définition) en nombre de pixels par pouce[7] ou par centimètre. La taille totale d'une image numérique correspond donc au nombre de pixels en largeur multiplié par le nombre de pixels en hauteur.

Générée par un appareil photo numérique, un scanneur ou directement par un logiciel informatique,

> « *l'image numérique, peut représenter une photographie, un dessin ou un texte. (…) Elle est aussi caractérisée par la profondeur de couleurs, le poids du fichier correspondant, la compression, les balises de métadonnées ou le format d'enregistrement. Pour démarrer un projet de numérisation d'archives, il est indispensable de bien comprendre chacune des caractéristiques d'une image numérique afin de faire les bons choix pour un résultat optimal, cohérent avec les objectifs et les moyens.* » (Essevaz-Roulet, 2016, p.39)

Les principales caractéristiques d'une image numérique selon Essevaz-Roulet, sont résumées dans le **tableau 6**.

[7] 1 pouce = 2.54 cm

Tableau 6 : Résumé des principales caractéristiques d'une image numérique

Caractéristiques	Commentaire
Origine de l'image	Numérisation, photographie numérique, logiciel informatique.
Taille	Exprimée en nombre de pixels total, nombre de pixels par côté ou en unité de longueur en fonction de la résolution.
Résolution	Densité surfacique de pixels par rapport à la taille de l'image physique en système métrique.
Format d'enregistrement	JPEG, TIFF, PNG, etc.
Mode et intensité de compression	Varie selon le format de l'image.
Profondeur des couleurs	16 millions de couleurs, niveaux de gris ou noir blanc…
Poids du fichier	Varie en fonction des paramètres précédents.
Métadonnées	Balises textuelles caractérisant l'image.

(Essevaz-Roulet, 2016, p.40)

La Bibliothèque et Archives nationales du Québec (BAnQ), la Bibliothèque nationale de France (BnF) et le Musée canadien de l'histoire (MCH) caractérisent un fichier numérique par les cinq paramètres suivants : sa résolution, sa profondeur de codage, son mode, son espace colorimétrique et son format (BAnQ, BnF & MCH, 2014, p.7).

2.3. Numérisation : définition et typologie

La numérisation est un processus mécanique qui consiste à transformer une information d'un support quelconque (texte, image classique, audio, vidéo, etc.) ou d'un signal électrique en une information numérique (nombres binaires) lisible par un ordinateur. Cette transformation peut se faire à l'aide d'un scanner, d'un appareil photo numérique ou d'une caméra à haute résolution, associés à un logiciel de capture. A ne pas confondre avec **dématérialisation*** qui désigne un processus global incluant l'acte de numériser (scanner) et qui permet à une organisation de se passer totalement de papier. Autrement dit, dans le contexte de la dématérialisation, les documents numériques de l'entreprise ont deux origines : natifs (produits ou reçus au format électronique) ou numérisés (c'est-à-dire scannés).

Concernant la typologie de numérisation, Essevaz-Roulet (2016, p.14) distingue trois grandes classes d'archives à numériser : la numérisation patrimoniale, la numérisation de production et la numérisation bureautique.

La numérisation patrimoniale **(figure 1)** consiste à numériser les **fonds d'archives*** ayant une valeur documentaire historique, informationnelle et de témoignage (fonds de documents écrits, plans, livres, fonds photographiques, œuvres artistiques…) appartenant à un particulier, une association, une entreprise ou une collectivité territoriale. La particularité de ce type de numérisation

> « *consiste en la qualité requise du travail plutôt qu'en la quantité et la rapidité de traitement. La copie numérique doit être aussi fidèle à l'original que possible de façon à pouvoir lui être substituée autant que de besoin.* » (Essevaz-Roulet, 2016, p.15)

Les raisons à l'origine d'une numérisation patrimoniale sont diverses (partage des documents, exploitation des données, réédition, mise à disposition du public, mise en valeur d'un fonds, extraction de l'information dormante, substitution, préservation, etc.), mais il s'agit principalement de préserver et de valoriser les fonds en les faisant connaître au plus grand nombre grâce à un accès ouvert (Internet).

La numérisation de production **(figure 2)** est une numérisation automatisée et massive des documents administratifs et comptables ou commerciales. Au contraire de la numérisation patrimoniale, la numérisation de production vise avant tout l'efficacité et la rapidité du travail. La finalité de ce type de numérisation

> « *n'est pas de partager les documents avec le public, mais de conserver une copie des documents pour exploitation ou avant destruction ou délocalisation. Les organismes concernés par la numérisation de production sont les entreprises et les collectivités territoriales qui produisent beaucoup d'archives papier ou les institutions qui doivent gérer des milliers, voire des millions de formulaires ou autres dossiers.* »
>
> (Essevaz-Roulet, 2016, p.17)

Enfin, la numérisation « de bureautique » est une numérisation tout-venant avec laquelle tout le monde est familier, car c'est celle qui peut se faire au secrétariat, à la maison, dans une papeterie, etc.

L'intérêt de cette classification pour ce travail, est qu'elle va nous servir de guide dans le choix d'entreprises à interviewer (voir le chapitre méthodologie plus loin).

2.4. Workflow de numérisation et concepts connexes

Un **workflow*** de numérisation (**figures 1 et 2**) est composé de phases (prénumérisation, numérisation proprement dite et post-numérisation), d'étapes et de tâches unitaires dont la nature et le nombre peuvent varier en fonction de la typologie de documents (numérisation patrimoniale et numérisation de production ou documentaire classique) et des objectifs poursuivis par le client. Il faut noter que l'ordre de ces étapes n'est pas fixe.

Pour réussir un projet de numérisation, Claerr et Westeel proposent de l'organiser sous forme d'un workflow ou flux de travail, car selon ces auteurs, la réussite d'un projet de numérisation

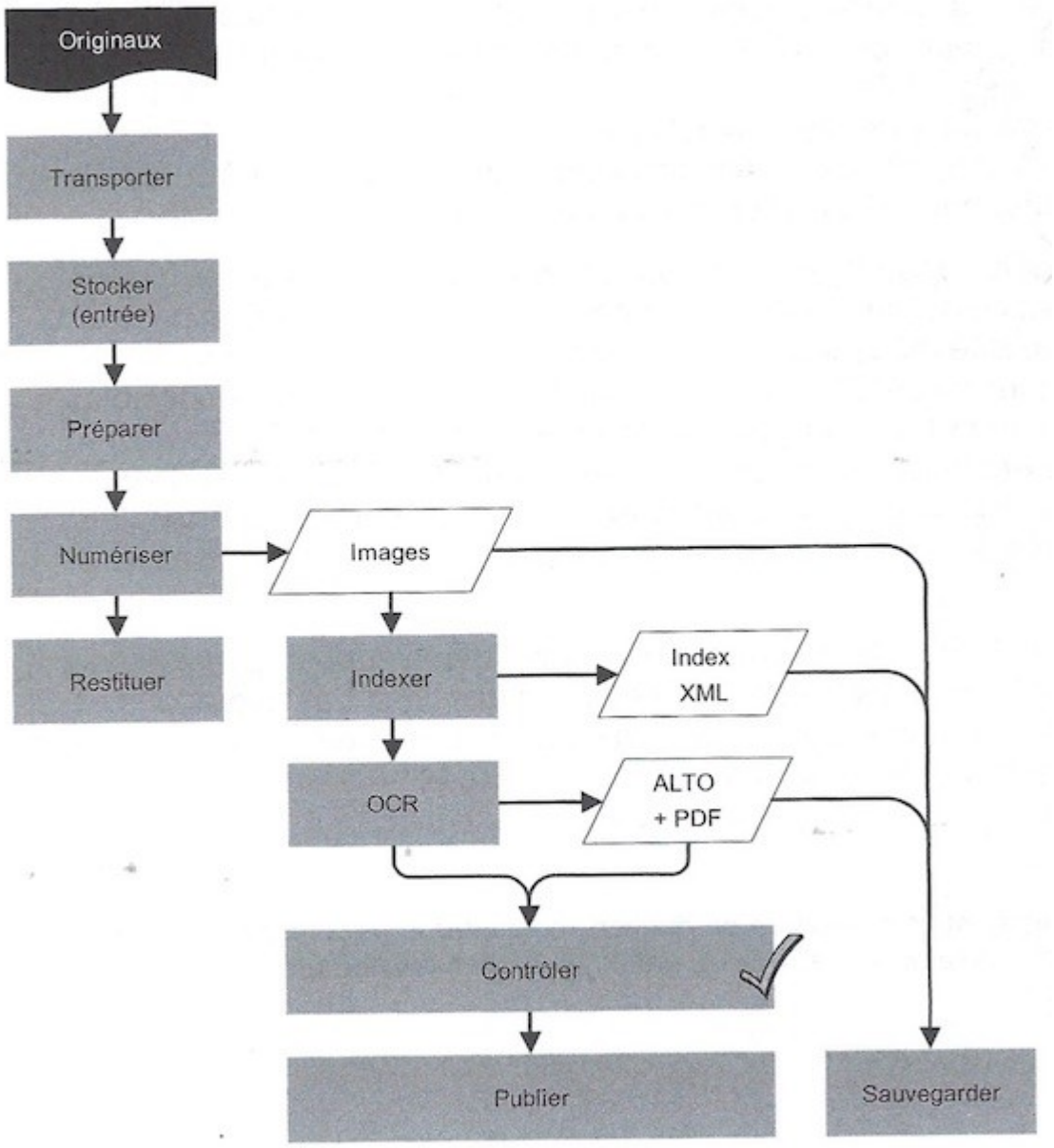
> *« ne dépend pas seulement de la qualité des matériels et logiciels mis en œuvre, mais également du processus élaboré et suivi, qui dépasse les étapes proposées par les logiciels de manutention.* » (Claerr et Westeel, 2011, p.136).

Claerr et Westeel précisent qu'un tel schéma,

> « *permet d'avoir une vue d'ensemble, utile pour ajouter des contrôles en entrée et sortie de chaque étape, d'identifier les métadonnées alimentées par chaque étape, de réduire les risques de chaque étape, notamment en identifiant les erreurs possibles, et d'établir une planification et un budget sur la base des contraintes, des productivités et des ressources (humaines, matérielles, logicielles, logistiques) propres à chaque tâche.* » (Claerr et Westeel, 2011, pp.136-137)

L'intérêt de maîtriser la chaîne de traitement des documents s'explique aussi par le fait que, lorsque le client demande au prestataire plusieurs services (**inventaire***, **numérisation***, **indexation***, **OCR***…), la qualité de chaque service est tributaire de celle des autres services, d'où la nécessité de prendre des précautions à chaque étape du processus de numérisation. Si la qualité d'une étape précédente est bonne, celle de l'étape suivante devrait aussi l'être si, bien entendu, la même rigueur est maintenue.

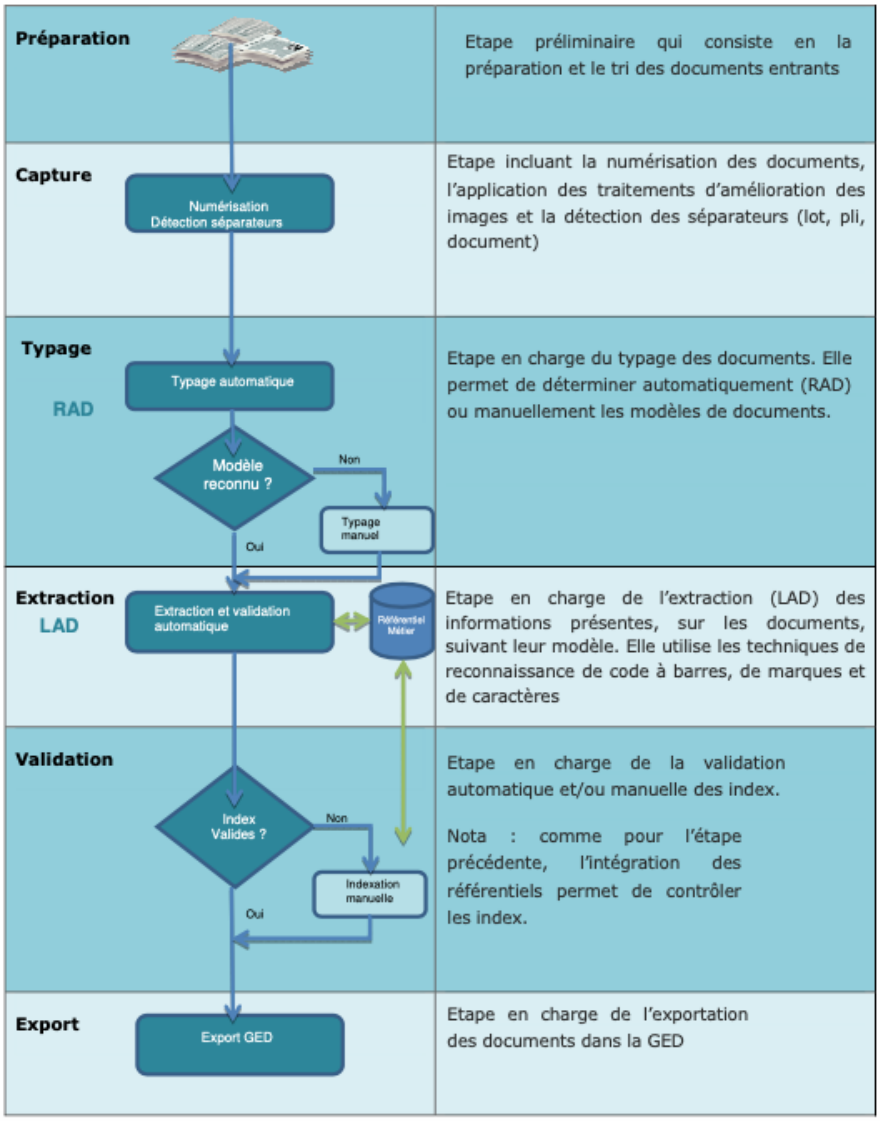
Figure 1 : Exemple générique d'un workflow de numérisation patrimoniale



(Claerr et Westeel, 2011, p.136)

Dans cette figure, chaque étape est représentée par un rectangle et les fichiers produits (images, Index **XML*** et **ALTO*** + PDF) par un trapèze.

Figure 2 : Etapes et tâches d'un workflow de numérisation de production ou numérisation documentaire classique



(DGSI, 2015, p.10)

Le workflow de numérisation de production ou documentaire classique contient deux étapes caractéristiques : la **RAD***(reconnaissance automatique de documents) ou typage ou classification et la **LAD***(lecture automatique de documents) ou extraction des données. Ce sont des technologies utilisées pour traiter les documents structurés (formulaires, chèques, RIB…) ou semi-structurés comme des factures (voir explications à la section 2.4.4).

2.4.1. Inventaire et caractérisation physiques des documents à numériser

L'inventaire est un « instrument de recherche fournissant une énumération descriptive (analyse) plus ou moins détaillée, des pièces composant un ou plusieurs fonds d'archives,

une ou plusieurs séries. » (Auguié et Vialle, 2017, p.128). L'inventaire permet de répondre à la question « Dans quel état sont les documents à numériser ? »

Qu'il soit sommaire ou complet,

> « *l'inventaire est l'occasion de déterminer l'état des archives, mais de façon plus large encore, de les caractériser. L'objectif de la caractérisation est de permettre au prestataire interne ou externe chargé de la numérisation d'adapter son offre ou son process aux documents selon leur état ou leur nature. La caractérisation des archives à scanner doit donc en principe figurer dans le cahier des charges de la prestation envisagée.* » (Essevaz-Roulet, 2016, p.94)

Le client est aussi tenu de signaler toutes les caractéristiques des documents susceptibles d'induire des contraintes techniques particulières pour le prestataire dont entre autres : documents fragiles, documents reliés, documents de très grande dimension (au-delà de A0), documents nécessitant des manipulations particulières de dépliage ou repliage, numérisables en machine à défilement, formats hétérogènes dans la même typologie de documents, négatifs en bandes ou à l'unité, documents très petits impliquant des manipulations lourdes, documents sous cache, etc.(BnF, 2010, p.16).

En prévision de la numérisation, l'inventaire

> « *servira directement à la constitution des bases de données et à l'enrichissement des métadonnées. Dans ce cas, certains champs dans l'inventaire doivent être spécifiquement réservés pour documenter les contraintes de la numérisation : état des documents, résolution à utiliser, etc. Le nom du document numérique peut reprendre la* ***cote***** du document physique*. » (Essevaz-Roulet, 2016, p.94)

2.4.2. Fichier de récolement

Le fichier de récolement « est une liste qui permet au prestataire d'identifier très précisément les documents à numériser et fournit des éléments nécessaires à leur nommage, voire à leur indexation. » (DAF, 2008, p. 22). Il comporte les éléments tels que l'en-tête et des informations précises pour chaque unité documentaire à numériser : cote, intitulé, dates extrêmes, métrage ou nombre d'unités matérielles des documents, état matériel, observations, etc. Se présentant généralement sous forme d'un tableau Excel, le fichier de récolement est un document utile pour suivre la prestation de numérisation et pour servir de base au bordereau de remise des produits réalisés. Il faut donc bien expliquer sa structure, voire la reproduire dans le cahier des charges.

> « *Toutes les précautions devront être prises afin que l'exploitation du fichier par le prestataire soit la plus aisée et complète possible (pas de retour à la ligne dans les cellules du tableur, pas de cellules fusionnées…).* » (Claerr et Westeel, 2011, p.126)

2.4.3. Bordereau d'accompagnement

Il accompagne les documents à numériser à l'aller et retour. Constitué à partir du fichier de récolement (pour rappel : établi par le client), le bordereau d'accompagnement « permet le transfert de responsabilité entre le commanditaire et le prestataire » (BnF, 2010, p.13). D'après ces auteurs, les informations suivantes doivent constituer le minimum à faire figurer sur le bordereau d'accompagnement :

- identification du lot,

- identifiant de chaque document,
- le nombre de volume ou de supports correspondants et ;
- la date d'enlèvement ou de retour des documents.

Dès la réception des documents à numériser, le prestataire est tenu de signaler toute non-conformité du lot avec le bordereau d'accompagnement (le délai de ce signalement doit figurer dans le cahier des charges).

2.4.4. RAD, LAD, OCR

La **RAD** ou reconnaissance automatique des documents (on dit aussi typage), est une technologie basée sur deux approches (DGSI, 2015, p.13) :

- l'approche graphique basée sur des mécanismes de reconnaissance de formes et de repérage de données bien localisées (ancres) grâce à un modèle ;
- l'approche syntaxique basée sur la détection des mots clés, des code-barres, des libellés précis grâce aux logiciels OCR.

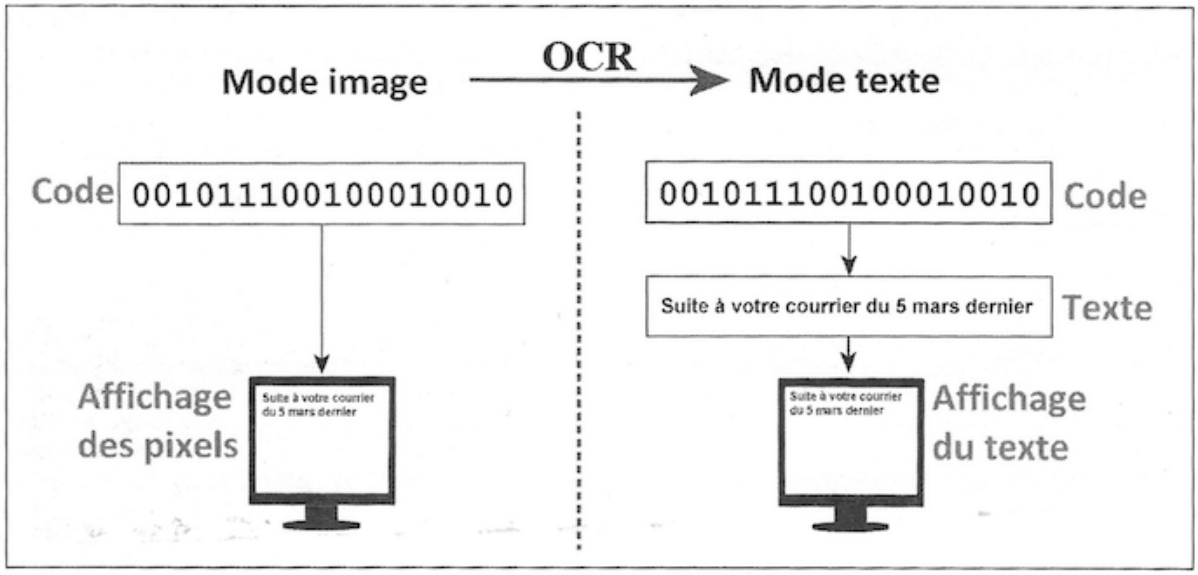
Il est possible de combiner les deux approches afin d'obtenir un niveau de typage élevé.

La **LAD** (lecture automatique de documents) est un ensemble de technologies qui permet de lire, segmenter et extraire, par reconnaissance optique de caractères (OCR, ICR et OMR) des informations sur des documents structurées (formulaires, chèques, etc.) ou semi-structurées (factures : total HT, numéro de facture, références de commande, etc.). Le système permet aussi d'associer les informations ainsi extraites (métadonnées) à une image numérisée afin d'en faciliter sa recherche.

Généralement, la RAD est combinée avec la LAD, car selon le type de document qui a été reconnu, les mêmes informations ne seront pas lues aux mêmes endroits. Inversement, selon les informations lues, le type de document peut être reconnu.

L'**OCR*** (Optical character recognition) ou ROC (reconnaissance optique de caractères) est un processus automatique de reconnaissance optique de caractères, qui, grâce à un logiciel dédié, permet de transformer un fichier image en fichier texte (**figure 3**). Le résultat de cette opération peut être sauvegardé dans un fichier texte, « ce qui permet d'indexer le contenu du document et d'effectuer de la reconnaissance en texte intégral, de la capture des données et, s'il y a lieu, des modifications. » (BAnQ, 2019, p.18).

Figure 3 : Exemple générique du fonctionnement de la technologie OCR



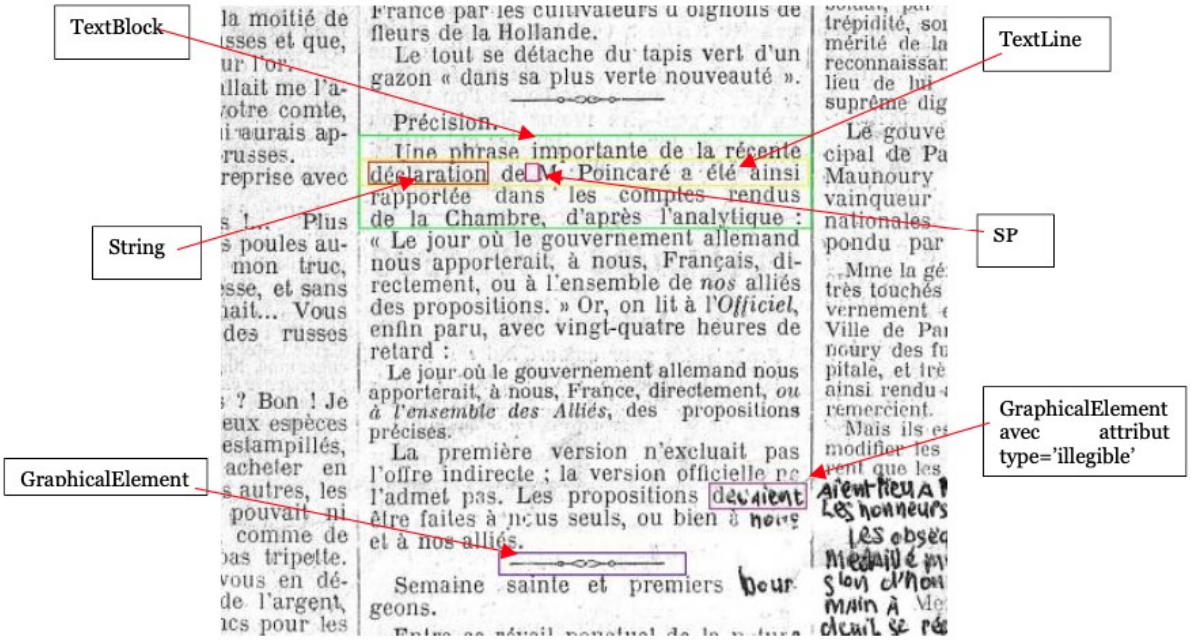
(Essevaz-Roulet, 2011, p.68)

L'océrisation peut être effectuée sur l'ensemble de la page numérisée ou sur certaines zones importantes du texte, des tableaux ou des images. Dans ce dernier cas, on parle de segmentation (**figure 4**). L'opération d'océrisation se fait alors en trois étapes (BnF 2015, p.7) :

> **1.** Préanalyse de l'image (redressement d'images inclinées, binarisation de l'image, correction de contraste) afin de faciliter l'océrisation ;
> **2.** Segmentation physique des contenus de l'image en sous-ensembles des régions ayant des propriétés communes (illustrations, blocs de textes, marges, etc.)[8] ;
> **3.** Reconnaissance des caractères de chaque bloc isolé par comparaison avec des ensembles de formes et à l'aide d'outils linguistiques ; puis conversion de l'ensemble dans un codage normalisé (ISO 8859-1, UTF-8, etc.).

[8] Les coordonnées de ces sous-ensembles ou « boîtes » sont enregistrées « en pixels dans un repère orthonormé dont l'origine est le point supérieur gauche de la page » (**Mocellin, 2010, p.32**).

Figure 4 : OCR-Segmentation d'une zone imprimée et caractérisation des boîtes



(BnF, 2010, Annexes, p.15)

Les taux de reconnaissance de caractères obtenus dépendent de la nature du texte et de l'état physique des documents originaux et de leurs images numérisées. Ces taux « sont calculés sur la base du mot, en moyennant les taux de confiance des mots présents dans un document » (BnF 2015, p.12) :

- OCR haute qualité : 99.9% ;
- OCR taux qualité garantie : 98.5% ;
- OCR brut (pas de correction manuelle) : pour les ouvrages datant de 1750 ou antérieur, ou pour les ouvrages sans date.

Le niveau de qualité requis pour la conversion en mode texte varie selon l'utilisation attendue du texte obtenu (indexation seule, affichage du plein texte). Si le taux recherché est proche de 100%, une intervention manuelle est nécessaire pour ressaisir les mots erronés, ce qui représente « un impact important en termes de coûts de réalisation. » (BnF, 2010, p.5). Selon la BnF, les facteurs suivants peuvent rendre complexe la conversion par OCR : présence de caractères d'alphabets exotiques et non latins, de zones manuscrites sur la majeure partie du document, de lacunes ou de déchirures affectant le texte ainsi que la qualité d'impression et l'hétérogénéité de celle-ci (p.5).

Certains clients peuvent se contenter d'un OCR brut (non corrigé), d'autres demandent un OCR raffiné, tout dépend des besoins de chacun-e ; mais la situation est telle que la majorité de clients demandent un OCR brut.

2.4.5. Nommage des fichiers

Le nommage des fichiers doit respecter les principes contraignants de la norme ISO 9660 (niveau 2) afin de garantir une bonne **interopérabilité***. Selon cette norme, la longueur des noms de répertoires et de fichiers (y compris le séparateur et l'extension), ne doit pas dépasser 31 caractères (DAF, 2008, p.37). DAF recommande « de faire figurer au début du nom de chaque fichier l'identifiant de l'institution responsable de l'opération de numérisation. » Le client doit, à travers le fichier de récolement, communiquer à son prestataire ces règles de nommage des fichiers.

2.4.6. Indexation et métadonnées

L'**indexation*** est une opération qui consiste à décrire le contenu d'une image numérisée afin de la retrouver et de faciliter son utilisation. La création des métadonnées sur le contenu des images numérisées doit se baser sur les normes internationales, comme la norme ISAD(G) pour la description archivistique (voir **annexe 6**). L'**indexation*** peut se faire par type (description formelle, auteur, titre, date, etc.), par concepts ou mots clés sélectionnés d'une façon libre ou contrôlée (thésaurus), ce qui permet d'harmoniser les pratiques. Le choix d'un standard de métadonnées va dépendre des modes d'accès et de recherche envisagés.

Par exemple, le ***Dublin Core**** (voir **annexe 7**), défini par la norme ISO 15836 est un format à caractère généraliste adapté pour la description des ressources bibliographiques et qui « sert souvent de plus petit dénominateur commun dans le cadre de projets interdisciplinaires. » (Gueit-Montchal et Boully, 2020, p.83). Ces formats sont en général utilisés sous forme de schémas encodés, prédéfinis, munis de dictionnaires et structurés en **XML***.

L'indexation se fait en deux étapes : identification, à partir des notices descriptives, des concepts qui seront représentés dans l'index, puis traduction de ces concepts en langage d'indexation.

Littéralement « données sur les données », les **métadonnées*** sont « un ensemble structuré d'informations attachées à un document, servant à en décrire les caractéristiques en vue de faciliter son repérage, sa gestion, son usage ou sa préservation. » (Équipe Vitam, 2020, p.6).

Selon Westeel (2010, p.102), pour caractériser un document numérique, il faut trois types de métadonnées :

- Les **métadonnées descriptives** qui regroupent les informations de contenu et d'identification du document : auteur, titre, mots-clés, identifiant…). Ces métadonnées sont décrites dans un format normalisé, comme **Dublin Core***, **EAD***, MarcXML, etc.
- Les **métadonnées techniques ou de structure** regroupant les informations qui relèvent de la version du document : date, format de fichier, taille, compression…) ;
- Les **métadonnées administratives** qui comprennent : a) les métadonnées de gestion des droits d'accès (droit d'auteur) et d'usage (droit d'impression, de modification…) ; b) les **métadonnées techniques** de préservation (type de fichier, taille du fichier, format, ressource, matériel, facteur de compression, résolution, etc.).

Les métadonnées peuvent être enregistrées, soit à l'intérieur du fichier image, soit stockées à l'extérieur de celui-ci. Certaines métadonnées doivent être saisies (métadonnées sémantiques), d'autres sont produites automatiquement par des scripts ou des conversions de fichiers. Lors de la numérisation, on recommande de créer les métadonnées essentielles suivantes : identifiant(s), titre du document, nom de l'établissement de conservation, date de numérisation, type de scanner et résolution de numérisation utilisés (Westeel, 2010, p.101).

Dans le cahier des charges, le client doit préciser au prestataire les travaux d'indexation à réaliser et les types de métadonnées à relever. Selon les compétences du prestataire en la matière, l'indexation peut être sommaire ou approfondie.

2.5. Paramètres techniques de numérisation

Appelés aussi spécifications techniques, les paramètres de numérisation doivent être bien compris et maîtrisés, car ils influent significativement sur la qualité de l'image numérisée, et par conséquent sur la qualité de l'OCR et de tous les livrables du projet de numérisation.

D'après la Direction des Archives de France (DAF, 2008, pp.31-36), les principaux paramètres techniques à considérer sont :

- le rapport d'agrandissement ;
- la résolution optique ;
- la colorimétrie ;
- la profondeur d'acquisition ;
- le profil colorimétrique ;
- le format de fichier image (pour la conservation et la diffusion) ;
- le type et taux de compression ;
- le cadrage-orientation des zones à numériser ;
- et les traitements d'images post-numérisation.

D'après Claerr et Westeel (2011, p.131), ces spécifications techniques « *doivent tenir compte à la fois des pratiques actuelles, des évolutions technologiques à venir (…), mais aussi des usages que l'on souhaite faire des fichiers.* » Les deux auteurs invitent également le prestataire à se renseigner à temps pour savoir si les applications du client présentent des restrictions vis-à-vis des spécifications techniques choisies.

Les valeurs indicatives des caractéristiques techniques des fichiers images décrites plus haut sont récapitulées dans le **tableau 7.**

Tableau 7 : Caractéristiques techniques des fichiers images issus de documents textuels manuscrits ou dactylographiés, sans images tramées : manuscrits ou dactylographiés feuille à feuille.

Paramètres	Utilisation du document		
	Conservation (Valeur historique prioritaire ou Originaux peu contrastés, écritures encre passée, papiers foncés, carbones…)	**Conservation** (Valeur historique non prioritaire et Originaux normalement contrastés)	**Diffusion**
Rapport d'agrandissement	1/1	1/1	1/1
Résolution de sortie	300 dpi	300 dpi	200 dpi
Colorimétrie ou modèle chromatique	RVB	Bitonal (bitmap seuil)	RVB ou 256 niveaux de gris
Échantillonnage	24 bits (millions de couleurs)	1 bit	24 bits (millions de couleurs)
Formats de fichier image	JFIF, JPEG 2000, PDF/A	TIFF	JFIF, PDF/A
Type et taux de compression	JPEG-Facteur de qualité 8/12 sur l'échelle de Photoshop	Compression LZW ou CCITT G4	JPEG-Facteur de qualité 5/12 sur l'échelle de Photoshop
Cadrage et orientation	Plein cadre fixe, sens de lecture	Plein cadre fixe, sens de lecture	Plein cadre fixe, sens de lecture
Profil colorimétrique	Adobe RGB (1998)	NA	sRGB IEC6 1966-2.1 Gray gamma 2.2
Post-traitements	NA	NA	NA

NA : Non appliqué

(Claerr et Westeel, 2011, pp.235-236)

2.5.1. Rapport d'agrandissement

Il représente le rapport entre la taille originale du document à numériser et celle de son image numérique. En principe, la majorité des documents originaux papier sont numérisés à leur taille réelle (1/1), c'est-à-dire sans agrandissement. En revanche, les documents de petite taille comme les négatifs, les diapositives, les plaques de verre, les microfilms, etc., sont numérisés dans une perspective d'agrandissement pour être visualisés ultérieurement. Dans ce cas, le rapport d'agrandissement devient supérieur à 1/1 (2/1, 3/1, 4/1, 5/1, 6/1, etc.).

2.5.2. Résolution optique

La résolution optique représente le nombre de pixels par unité de mesure, en l'occurrence le pouce (2.54 cm). Elle est exprimée en dpi (dot per inch) ou en ppi (pixel per inch) ou en ppp (pixels par point). Par exemple, une image à "300 dpi" veut dire qu'il faut 300 pixels (ou dots) pour représenter 2.54 cm (1 pouce) en hauteur comme en largeur. Le calcul inverse dit qu'un pixel mesurera, dans ces conditions, 0.0847 mm, soit 2.54cm/300 (Claerr et Westeel, 2011, p.154). Il est possible, après la capture, « d'accroître la résolution par ajout artificiel de pixels de valeurs similaires aux pixels proches », opération qui s'appelle rééchantillonnage ou interpolation, mais « à éviter dans un projet de numérisation qui vise à la préservation à long terme des documents. » (Mocellin, 2010, p.22).

Plus la résolution est haute, mieux est la qualité de l'image, mais en même temps le fichier devient lourd s'il n'est pas compressé. Le **poids de l'image**[9] est déterminé par la résolution et la profondeur d'acquisition selon la formule suivante : **définition x profondeur d'acquisition/8**. Pour avoir le poids en kilo-octet, il faut diviser la valeur obtenue par 1 024. Ainsi, le choix du niveau de résolution est fonction de la taille du document original, de l'utilisation attendue des images numérisées, des capacités du système d'archivage de l'institution et de ses évolutions. Le **tableau 8** donne une liste de résolutions conseillées en fonction des objectifs.

Tableau 8 : Résolutions conseillées en fonction des objectifs du projet de numérisation

Cas de figure	Résolution	Commentaire
Page manuscrite ou imprimée standard pour envoi par courriel	96 dpi	Cette résolution est en général suffisante pour conserver la lisibilité tout en minimisant le poids de l'image.
Page manuscrite ou imprimée pour archivage et OCR	200-300 dpi	A ajuster selon la taille des caractères ou de l'écriture.
Toute image destinée à être mise en ligne sur Internet	72 dpi	A n'utiliser que si l'on n'a pas l'intention de se servir de l'image pour autre chose.
Gravure pour archivage	400-800 dpi	A ajuster selon la finesse du trait et des détails.
Photographie pour archivage	600-1200 dpi	A ajuster selon le piqué de l'image.
Objets (médailles, sceaux…)	300-600 dpi	A ajuster selon l'usage que l'on veut faire des images.
Diapositifs et films négatifs	4000-6000 dpi	Choisir la résolution maximale du scanneur.
Dessin au trait	300 dpi	A ajuster selon l'usage que l'on veut faire de l'image.
Images en tons continus (peintures ou dessins genre pastel)	300-800 dpi	A ajuster selon la taille des détails que l'on souhaite saisir.
Document pour impression	Varie	Selon la formule suivante[10] : RA = RR x coefficient d'agrandissement
Demi-teinte (images imprimées)	Varie	Attention à l'effet de moirage.

(Essevaz-Roulet, 2016, p.42)

En résumé, la Direction des Archives de France (DAF, 2008, p.32) conseille trois cas de figure suivants :

- 300 dpi pour les fichiers de conservation ;

[9] A ne pas confondre avec le poids du fichier définitif qui est plutôt lié au format et à une éventuelle compression.

[10] RA = résolution d'acquisition et RR = résolution de restitution. Exemple : On scanne une photo de format 10 x 12 cm avec une résolution de restitution (RR) de 300 dpi en vue de l'agrandir en poster de format 30 x 40 cm (format A3). Dans ce cas, la résolution d'acquisition (RA) sera de 900 dpi. En effet, le passage du format 10 x 12 cm au format 30 x 40 cm, correspond à un coefficient d'agrandissement d'environ 3. Ainsi, RA (900 dpi) = RR (300 dpi) x 3.

- 150 dpi pour les fichiers de diffusion et ;
- 72 dpi pour les fichiers de visualisation.

2.5.3. Colorimétrie

Elle correspond à la mesure des couleurs selon les modèles chromatiques **noir et blanc**, **niveaux de gris**, **couleurs RVB** (rouge, vert et bleu), et **couleurs CIE L*a*b*[11]** ; qui déterminent la façon dont il faut coder la tonalité de l'image numérisée.

Le choix de l'un ou l'autre de ces modèles dépend de l'apparence du document et du résultat que l'on souhaite obtenir. La numérisation en couleur est en général préférable à la numérisation en niveaux de gris. L'une des problématiques de la gestion de la couleur en numérisation, est de pouvoir maintenir et communiquer cette couleur entre les trois types d'appareils qui caractérisent la chaîne graphique numérique : les appareils d'acquisition (scanner, appareil photo numérique), les appareils d'affichage (écrans d'ordinateurs) et les appareils d'impression (Vinsonneau, 2015, p.90). D'où l'importance de l'opération de calibrage d'un scanner à l'aide d'une **mire colorimétrique** dont les références sont connues.

2.5.4. Profondeur d'acquisition et mode

Appelée aussi profondeur d'échantillonnage, profondeur d'analyse, profondeur de codage ou profondeur de couleurs, la profondeur d'acquisition représente le nombre de bits d'information utilisée pour représenter chaque pixel de l'image (DAF, 2008, p.33).

Un **mode** « est un modèle mathématique destiné à définir les relations des couleurs entre elles par le nombre de couches présent au sein de ces modes. » (Brault, 2021, p.24). Il existe trois principaux modes d'encodage des couleurs définis par le nombre de bits sur lesquels on peut coder un pixel (Mocellin, 2010, p.21) : **le mode bitonal, le mode niveaux de gris et le mode couleurs**.

Le **mode bitonal[12]** correspond à un codage de chaque pixel sur un bit, ce qui lui donne la possibilité de ne prendre que 2 valeurs : 0 et 1 (noir ou blanc). L'image bitonale est obtenue « par binarisation, ou réduction des valeurs de pixels en niveaux de gris ou en couleurs à 0 ou 1, à partir d'une valeur de référence fixée dans les gris. » (Mocellin, 2010, p.21). Sur les imprimés de bonne qualité, ce procédé facilite la lecture tout en permettant de diminuer le poids des fichiers pour le stockage. Sur des papiers ayant trop de taches ou à faible

[11] CIE : Commission internationale de l'éclairage. **L*a*b*** représente un espace colorimétrique de connexion et indépendant de l'appareil, pris comme référence par le Comité international de la couleur ou International Color Consortium (ICC). Cet espace est construit selon deux axes : luminance et information chromatique. Avec L= luminance (indice de luminosité variant de 1 à 100), a = composante chromatique rouge-vert et b = composante chromatique jaune-bleu. L'image est d'abord acquise dans un espace RVB (rouge, vert et bleu), qui est ensuite converti dans l'espace de connexion L*a*b* (Vinsonneau, 2015, pp. 98-100).

[12] Lorsqu'on choisit d'utiliser ce mode de codage, un examen préalable des documents à numériser (état physique et typographie) est nécessaire pour s'assurer, par exemple, que les taches de roussissure ou d'humidité présentes sur le papier ne soient pas interprétées par le scanner comme des points à traduire en noir (Micaletti, 2011, p.2, annexe 8).

contraste, la binarisation peut donner des images illisibles. Ce qui veut dire que, « la profondeur d'acquisition est fonction du support et de son contenu », précise Mocellin.

Le **mode Niveaux de gris** (privilégié pour la numérisation en noir et blanc) correspond à un **codage de chaque pixel sur 8 bits (1 octet)**, ce qui restitue une image à 256 couleurs ou niveaux de gris différents (2^8). Ce mode de codage a l'avantage de mieux restituer les nuances colorimétriques, donc de « préserver correctement l'aspect ancien des documents » (Micaletti, 2011, p.2, annexe 8). En revanche, il n'est pas économique en termes de mémoire. Le **mode Couleurs RVB** (privilégié pour la numérisation en couleur) correspond à un **codage de chaque pixel sur 24 bits (3 canaux x 8 bits)** avec restitution d'une image à 16.7 millions de couleurs vraies (2^{24}) ou sur **48 bits (3 canaux x 16 bits)**. Par rapport au mode bitonal, les fichiers sont ici 24 ou 48 fois plus lourds.

Il y a aussi le **mode CMJN** (Cyan-Magenta-Jaune-Noir), mais qui est privilégié pour l'impression Offset, donc non recommandé pour la numérisation. De même, les modes **couleurs indexées**, **bitmap et bitonal** « ne sont pas envisageables dans le cadre des projets de numérisation. » (Brault, 2021, p.24).

2.5.5. Profil colorimétrique

Le profil ou espace colorimétrique correspond à la gamme de couleurs qu'un scanner, une imprimante ou un écran peut acquérir ou restituer. D'après Brault (2021, p. 25), « Un espace colorimétrique est un modèle mathématique tridimensionnel représentant l'ensemble des couleurs perceptibles, utilisables et reproductibles au sein d'un mode. » Il existe deux types d'espaces colorimétriques : ceux qui dépendent d'un appareil particulier et ceux ne dépendant pas d'un appareil particulier comme l'espace L*a*b*. Dans les deux cas, les deux espaces colorimétriques sont désignés sous le nom de **profil ICC***.

Le profil colorimétrique « est l'un des points délicats dans un projet de numérisation » à cause de l'existence, dans la chaîne numérique, de deux principaux espaces de couleurs, à savoir **le RVB** (rouge, vert, bleu) à l'acquisition et à la restitution sur écran, et le **CMJN** (cyan, magenta, jaune, noir) à l'impression (Mocellin, 2010, p.30). Et dans chaque espace, ces couleurs ne sont pas traduites de la même façon. Le recours à des fichiers normalisés (profils ICC* par exemple) permet, d'après Mocellin, « de convertir les références d'un espace à l'autre tout au long de la chaîne pour garantir une impression des couleurs fidèles à l'original. »

Pour une bonne restitution de couleurs au long du projet de numérisation, Mocellin (2010, pp.30-31) conseille les pratiques suivantes :

- Avoir un système de numérisation performant et adapté au projet avec une bonne gestion de la lumière à la capture ;
- Étalonner le scanner, c'est-à-dire définir un état colorimétrique de référence à partir des mires normalisés, des profils ICC ;
- Ajouter à chaque document numérique l'image des mires numérisées le jour du traitement du document ;
- Calibrer les écarts de contrôle à l'aide d'un spectrophotomètre ou d'un colorimètre ;
- Comparer visuellement les images avec les originaux.

2.5.6. Formats de fichiers images

Le format d'un fichier (comme **TIFF***, **JPEG***, **JPEG2000*** et **PNG*** pour les images ; et PDF, **PDF/A*** et **ALTO*** pour le texte) correspond à la façon de coder les données. Son choix a un effet direct sur les performances de l'image numérique ainsi que des implications pour la gestion à long terme de l'image. Il n'y a pas de format de fichier correct pour toutes les applications, tous les choix de format impliquent des compromis entre la qualité, l'accès et la gestion du cycle de vie. Il est conseillé de préciser le nom du format suivi de sa version. Après avoir choisi le format, il est important de maintenir ce choix tout au long du processus de numérisation. En effet :

> « *un changement de format (…) au milieu d'une opération peut occasionner des problèmes de traçabilité, de nommage, voire des difficultés en cas de retraitements ultérieurs (par exemple pour des opérations de reconnaissance de forme.* »
>
> (DAF, 2008, p.34)

Dans certaines institutions comme la Bibliothèque nationale de France (BnF), la *Library of Congress*, la *Wellcome Library* et la *British Library*, le format JPEG2000 tend à supplanter le format TIFF, car les fichiers sont moins lourds. Mais TIFF comme JPEG2000, « sont toujours utilisés afin de créer un fichier numérique de conservation » à partir duquel sont extraits d'autres formats de fichiers de diffusion comme JPEG et PDF (BAnQ, BnF, MCH, 2014, p.11).

2.5.7. Type et taux de compression

Les techniques de compression visent à réduire le nombre de pixels de l'image. Schématiquement, cela consiste « *à repérer dans l'image les plages contenant des pixels identiques. Plutôt que de les décrire un par un, la compression va les décrire par lots.* » (Essevaz-Roulet, 2016, p.44). Une compression peut être sans perte, comme LZW (Lempel-Ziv-Welch) pour les images en couleur ou CCITT T.4 groupe 3 pour les images en noir et blanc ; ou avec perte, mais avec possibilité d'agir sur l'intensité de la réduction.

Le taux de compression peut être exprimé en pourcentage ou sous forme de rapport : par exemple, le rapport 1 : 2.5 signifie que le fichier a été comprimé 2.5 fois. Un bon taux de compression est celui qui donne le meilleur ratio qualité de l'image/taille du fichier. De manière générale, les experts du traitement d'image recommandent les compressions suivantes (Essevaz-Roulet, 2016, p.47) :

- Documents noir et blanc (plans, cadastres, textes…) : compression Deflate, PackBits ou LZW ;
- Documents en couleurs (photographies…) : JPEG ou JPEG 2000 ;
- Documents historiques : selon compromis.

2.5.8. Cadrage et orientation

Le cadrage correspond à la partie du document qui doit être numérisée : ça peut être la partie contenant seulement les données ou tout le document avec ou sans bordure. La numérisation d'un document destiné à la conservation doit porter sur l'ensemble du document. L'orientation indique le sens de l'image numérisée.

2.5.9. Post-traitements d'images

Ces traitements sont nombreux et ils influent sur la qualité de l'OCR. Généralement coûteux, ils devraient, d'après la Direction des Archives de France, être réservés aux seuls fichiers destinés à l'exploitation (et non aux fichiers destinés à la conservation). Pour les documents textuels, les traitements post-numérisation les plus fréquents sont les suivants (Essevaz-Roulet, 2016, p.67) :

- Ajustement de la luminosité et du contraste ;
- Redressement et rotation de l'image ;
- Conversion de format de fichier ;
- Modification de la taille de l'image ;
- Ajustement de la compression ;
- Indexation ;
- Numérotage des pages ;
- Ajout d'une mention écrite (« Copyright… », origine du fonds, cote…) ;
- Nettoyage/filtrage de l'image ;
- Reconnaissance de caractères (OCR) ;
- Ajout d'un logo, d'une image ou filigrane ;
- Jonction, séparation d'images ;
- Redressement des courbures de page ;
- Correction d'éclairage 1D, 2D ;
- Convention en noir et blanc ou niveaux de gris ;
- Suppression de bordures ;
- Assemblage en PDF ;
- Renommage ;
- Harmonisation du format de page ;
- Centrage des zones de texte ;
- Suppression des pages blanches.

Pour les documents iconographiques, Essevaz-Roulet mentionne les traitements suivants (p.64) :

- Ajustement de la luminosité et du contraste ;
- Ajustement de la balance des couleurs ;
- Conversion de format de fichier ;
- Modification de la taille de l'image ;
- Ajout d'une mention écrite (« Copyright… », origine du fonds, cote…) ;
- Redressement et rotation de l'image ;
- Retouches des défauts (poussières…) ;
- Ajout d'un logo, d'une image ou filigrane ;
- Modification du nombre de couleurs ;
- Superposition, jonction, séparation d'images ;
- Ajout de bordures ;
- Floutage fin (utile pour les images tramées) ;
- Renommage ;
- Augmentation de la netteté.

Généralement coûteux, les traitements post-numérisation devraient, d'après la Direction des Archives de France (2008, p.36), être réservés aux seuls fichiers destinés à l'exploitation (et non à ceux destinés à la conservation). Pour tous ces paramètres techniques, un compris doit être trouvé entre le client et le prestataire de numérisation, et il doit être consigné clairement dans le contrat.

3. Méthodologie

Pour atteindre les objectifs susmentionnés, nous avons adopté une démarche méthodologique en quatre étapes :

- Etape 1 : faire une revue de littérature sur la problématique, la typologie et les concepts clés du workflow de numérisation d'archives papier (sources principales : Internet et l'Infothèque de la Haute école de gestion de Genève). La typologie de numérisation nous a servi de guide dans le choix d'entreprises.
- Etape 2 : choisir les entreprises prestataires de numérisation à interviewer selon la typologie établie à l'étape 1.
- Etape 3 : Préparation et conduite des entretiens avec les entreprises choisies.
- Etape 4 : Analyse et traitement des données d'entretiens.

3.1. Critères de choix des entreprises

Pour choisir les entreprises à interviewer, nous avons été guidé par la typologie de numérisation, et par conséquent la typologie d'entreprises décrite plus haut et proposée par Essevaz-Roulet (2016, pp.14-19). Cette typologie nous a permis de classer les entreprises en trois catégories :

- Entreprises spécialisées en numérisation patrimoniale : entreprises A, B, C et D ;
- Entreprises spécialisées en numérisation de production ou numérisation documentaire classique : entreprises F et G ;
- Entreprises mixtes : entreprise E.

Pour ce faire, nous avons contacté Monsieur Jérôme Guisolan, archiviste auprès des Archives cantonales vaudoises (ACV) qui nous a communiqué plusieurs noms d'entreprises dont quelques-unes font partie de notre échantillon. Nous avons aussi contacté la Bibliothèque nationale suisse (BN)[13] et ce contact nous a permis de découvrir le site www.digicoord.ch, une plateforme d'informations sur les projets de numérisation en Suisse, que la BN gère conjointement avec RERO, réseau des bibliothèques de Suisse occidentale. C'est là que nous avons trouvé la liste complète des adresses de quelques entreprises que Jérôme Guisolan nous a communiquées et d'autres.

Au total, nous avons contacté 13 entreprises :

- 7 ont répondu positivement et elles constituent notre échantillon ;
- 3 entreprises ont répondu négativement, avançant le manque de temps à nous consacrer ;
- une entreprise nous a répondu positivement, mais tout en précisant que la numérisation n'est pas son activité principale ;
- une entreprise nous a donné son accord de participer à ce projet, mais n'a pas donné suite depuis, malgré nos plusieurs rappels ;

[13] Avec Liliane Regamey, responsable de la Section utilisation.

- une entreprise de calure internationale, n'a pas donné suite à notre sollicitation d'entretien malgré de nombreux rappels et nombreuses conversations avec ses réceptionnistes ou secrétaires. A chaque rappel, ces derniers nous affirmaient avoir transmis notre demande au responsable hiérarchique, mais sans effet !

3.2. Préparation et conduite des entretiens

Pour chaque entreprise, nous avons, pour obtenir un entretien, passé par quatre étapes. Premièrement, nous avons contacté l'entreprise par téléphone, d'abord pour obtenir plus de précision sur ses prestations de numérisation d'archives papier (complément aux informations récoltées sur son site Internet). Si cela correspondait à ce que nous attendions, nous avons, dans la foulée, demandé à notre interlocuteur ou interlocutrice s'il était possible d'envoyer un message écrit sollicitant un entretien. Dans l'affirmatif, nous avons, deuxièmement, envoyé ce message de sollicitation d'entretien. Troisièmement, et en cas de réponse positive à notre demande d'entretien, nous avons envoyé à l'entreprise notre guide d'entretien (voir **annexe 2**)[14] afin que notre futur-e répondant-e (représentant de l'entreprise) puisse se préparer convenablement, en particulier pour les questions trop techniques (scanner, logiciels, paramètres techniques de numérisation, aspects juridiques et normatifs de la numérisation). Enfin, en fonction de la date et du lieu d'entretien convenus (dans les locaux de l'entreprise ou sur Teams), nous avons mené des entretiens semi-structurés et ceux-ci ont été enregistrés, sur accord des répondant-e-s, à l'aide d'un dictaphone.

3.3. Méthode d'analyse des résultats

L'analyse des résultats a été effectuée comme suit :

- transcription automatique des fichiers audio (format mp3) avec une application de Microsoft Office 365, via Teams ;
- relecture et correction des verbatims obtenus ;
- analyse qualitative des verbatims (constitution des tableaux de données qualitatives);
- envoi du document de synthèse à chaque représentant de l'entreprise pour appréciation, correction (s'il y a lieu) et validation ;
- analyse qualitative et comparative des résultats finaux.

[14] Elaboré en tenant compte des données de la littérature et des objectifs de ce mandat.

4. Résultats des entretiens

4.1. Informations générales sur les entreprises

Ce sous-chapitre passe en revue les informations générales sur les entreprises ayant participé à cette étude. A la fin de ce sous-chapitre, nous donnons dans les **tableaux 9 et 10** une synthèse des informations importantes sur les entreprises, y compris les services offerts autres que numérisation. Ces entreprises sont symbolisées par des lettres de l'alphabet pour des raisons d'anonymisation des données.

4.1.1. Entreprise A

L'entreprise A est une société privée française prestataire en numérisation et conservation du patrimoine écrit et iconographique dont le siège est établi à Bordeaux. L'entreprise possède aussi des ateliers permanents à Paris, Berlin et Genève. Elle offre ses services de numérisation, d'enrichissement et de valorisation du patrimoine culturel partout dans le monde en installant, depuis 2011, des ateliers de numérisation temporaires chez ses clients. Par exemple, l'entreprise s'est installée à l'ONU-Genève où elle numérise, depuis 2018, les archives de la Société des Nations (SDN), l'ancêtre de l'ONU. La fin du projet (environ 15'000'000 de pages à numériser) est prévue en septembre 2022. La numérisation dans les locaux du client se fait sur demande de celui-ci pour plusieurs raisons, dont entre autres, le souhait de garder les documents sur place pour des questions de sécurité, et la difficulté de les déplacer car trop volumineux. De manière générale, l'entreprise numérise dans les ateliers de ses clients tout ce qui est format standard ; pour de très grands formats, la numérisation se fait dans ses ateliers permanents, sur des scanners bien spécifiques qui ne peuvent pas se déplacer parce beaucoup trop gros.

L'entreprise A propose les principaux services suivants :

- Audit et conseil en stratégie patrimoniale ;
- Audit et étude technique numérisation ;
- Gestion des fonds patrimoniaux ;
- Conservation préventive ;
- (Re)constitution des collections ;
- Offres de numérisation ;
- Numérisation sur site client ;
- Valorisation Web ;
- Création de contenus ;
- Muséographie.

D'après le représentant de l'entreprise A, responsable marketing et business développement, même si l'entreprise a élargi son offre, comme faire des audits des collections papier et iconographiques, faire des inventaires et les fichiers de récolement, « la numérisation reste toujours le cœur de l'activité de l'entreprise. » Par exemple, sur environ 80 employés, 50 sont des opérateurs de numérisation, photographes ou iconographes, répartis un peu partout en France et dans le monde.

L'entreprise A ne s'arrête pas uniquement à la numérisation : il y a aussi, et sur demande des clients, le volet valorisation du patrimoine culturel via la mise en place des bibliothèques numériques. C'est le cas, par exemple, des archives de l'ONU à Genève (projet en cours),

de la bibliothèque de l'Institut des Hautes études internationales et du développement (IHEID) à Genève, de la Bibliothèque Mazarine en France, des cinémathèques, des archives publiques et privées partout dans le monde.

4.1.2. Entreprise B

L'entreprise B est une entreprise suisse prestataire de services d'archivage et de conseil en gestion de l'information qui a son siège à Baden (Argovie) et un bureau en Suisse romande (Yverdon-les-Bains). En matière d'archivage électronique à long terme, l'entreprise B intervient en Suisse, en Allemagne, en France, en Autrice et en Suède. Elle est membre de l'Association des archivistes suisses (AAS), de l'Association vaudoise des archivistes (AVA), de l'Association des archives économiques d'Allemagne et membre collectif du Conseil international des archives (ICA).

Les principales prestations de l'Entreprise B sont les suivantes :

- conseil et gestion de l'information ;
- traitement et gestion d'archives ;
- archivage électronique à long terme.

Le Conseil et gestion de l'information inclut les tâches comme le conseil et gestion documentaire (records management), conseil en gestion de l'information, l'optimisation de la gestion électronique des documents, le conseil en gestion du cycle de vie et en archivage.

Le traitement et gestion d'archives comprend le classement et description, l'évaluation archivistique et tri, l'externalisation, la numérisation, et la conversion de formats de fichiers.

L'archivage électronique à long terme est basé sur la solution logicielle et modulaire docuteam cosmos cloud[15], ainsi que sur Atom (Access to Memory). D'après le site Internet de l'entreprise B, son défi « réside dans le fait de pouvoir utiliser les données à très long terme tout en étant capable de prouver leur authenticité. » La solution logicielle proposée par l'entreprise B utilise le modèle OAIS pour proposer un archivage électronique à long terme en faveur des collectivités publiques, des entreprises et organisations privées qui ne souhaitent pas mettre en place leur propre infrastructure d'archivage numérique. Pour les gros clients qui souhaitent exploiter leurs archives électroniques à long terme sur leur propre infrastructure informatique, l'entreprise B installe docuteam cosmos dans leur centre de données.

La solution docuteam cosmos comprend trois composants : mise en œuvre initiale de l'infrastructure d'archivage, Software as a Service (SaaS) et hébergement, et prestations de gestion d'archives. La mise en œuvre initiale de l'infrastructure d'archivage (déploiement des composants dédiés de la solution docuteam cosmos cloud) comprend les opérations suivantes :

[15] Docuteam cosmos comprend cinq logiciels : 1) **docuteam packer** pour l'empaquetage des données dans un Submission Information Package (SIP) ; 2) **docuteam actions** pour créer, éditer et sauvegarder des SIP ; 3) **docuteam feeder** : outil web pour migrer, valider, modifier et stocker les données ; 4) **docuteam services** qui offre diverses fonctionnalités en lien avec un dépôt Fedora ; et 5) docuteam bridge API (Application Programming Interface), une interface machine qui permet de déposer et récupérer des paquets archivés.

- configuration des services de synchronisation pour le téléchargement/versement des données et des processus sur le serveur Ingest central ;
- installation et configuration du dépôt numérique (Repository) du client dédié (Fedora Cosmos) et des mesures de sauvegarde par réplication ;
- configuration et test de versements issus d'une GED ou d'une application de gestion d'affaires.

Quant à la plateforme Saas ou Software as a Service (« logiciel en tant que service ») et hébergement, elle est mise à la disposition des clients sous forme de services. Les clients n'ont donc pas besoin d'installer ni de gérer eux-mêmes les composants nécessaires à l'archivage. Ils peuvent définir eux-mêmes l'intensité de la prise en charge et décider des travaux qu'ils souhaitent effectuer eux-mêmes ou qu'ils souhaitent déléguer aux archivistes spécialisé-e-s de l'entreprise B.

ATOM est un système d'information archivistique open source, basé sur le web qui peut servir pour le catalogage, la communication et la recherche ; et reposant sur les normes internationales archivistiques (ISAD(G), ISAAR (CPF), ISDF et ISDIAH). L'entreprise B propose de l'héberger dans des centres de calcul suisses partenaires.

Pour ses clients, l'entreprise B installe et entretient AtoM sur leurs infrastructures, paramètre et adapte AtoM en coordination directe avec Artefactual, le constructeur canadien, migre les données et métadonnées de systèmes d'information d'archives existants et adapte l'interface d'AtoM selon les CD/CI (distribution continue/intégration continue) des clients.

En partenariat avec d'autres institutions, l'entreprise B intervient aussi dans la formation des agent-e-s en information documentaire (AID) et des stagiaires.

4.1.3. Entreprise C

L'entreprise C est une entreprise suisse créée en 1998 en tant que bureau d'ingénieurs, active dans l'Engineering et l'Automation. Depuis 2000, l'entreprise C est devenue prestataire de numérisation patrimoniale, principalement des bibliothèques. L'entreprise intervient principalement en Suisse romande et propose à ses clients un autre service, celui des solutions automatiques pour la numérisation de livres, de revues, de magazines et de journaux.

Grâce à cette technologie, l'entreprise C a apporté un savoir-faire révolutionnaire en numérisation patrimoniale grâce à la mise au point de scanners munis de robots qui tournent automatiquement les pages des livres et des autres documents reliés. Actuellement, les scanners de l'entreprise C sont utilisés dans plusieurs pays du monde.

4.1.4. Entreprise D

L'entreprise D est une entreprise suisse basée à Berne, prestataire de numérisation patrimoniale orientée principalement vers les documents iconographiques. Son fondateur et propriétaire travaille seul dans son laboratoire de photographie situé à Berthoud, où il numérise les documents de la plupart de ses clients à l'exception de ceux de la Bibliothèque nationale suisse (BN). Pour la BN, le responsable de l'entreprise D doit se déplacer, car la

BN exige que ses archives ou documents graphiques soient numérisés sur place, dans ses locaux. Le fait, pour l'entreprise D, de ne pas se déplacer pour les autres clients, est dû au poids lourd de ses scanners, difficiles à déplacer. Lorsque le responsable de l'entreprise numérise pour la Bibliothèque nationale, celle-ci met à sa disposition tout le matériel nécessaire dont il a besoin pour mener à bien les termes du contrat.

L'entreprise D propose les services suivants :

- Numérisation, y compris au grand format ;
- Post-traitements des diapositives ;
- Archivage d'images avant leur restitution ;
- Renumérisation.

D'après son site Internet, et à propos de cette dernière prestation de renumérisation (Redigitalierung), l'entreprise D est l'un des rares fournisseurs à pouvoir ramener les séparations en 4 couleurs (séparations de couleurs CMJN sur film) de brochures antérieures, d'anciens motifs publicitaires, de périodiques, etc. dans leur forme d'origine (**figure 5**).

Figure 5 : Un exemple de renumérisation des séparations quatre couleurs avec reconstitution de la forme d'origine des images, basé sur une ancienne couverture "Silver Arrow" de la maison d'édition BASTEI

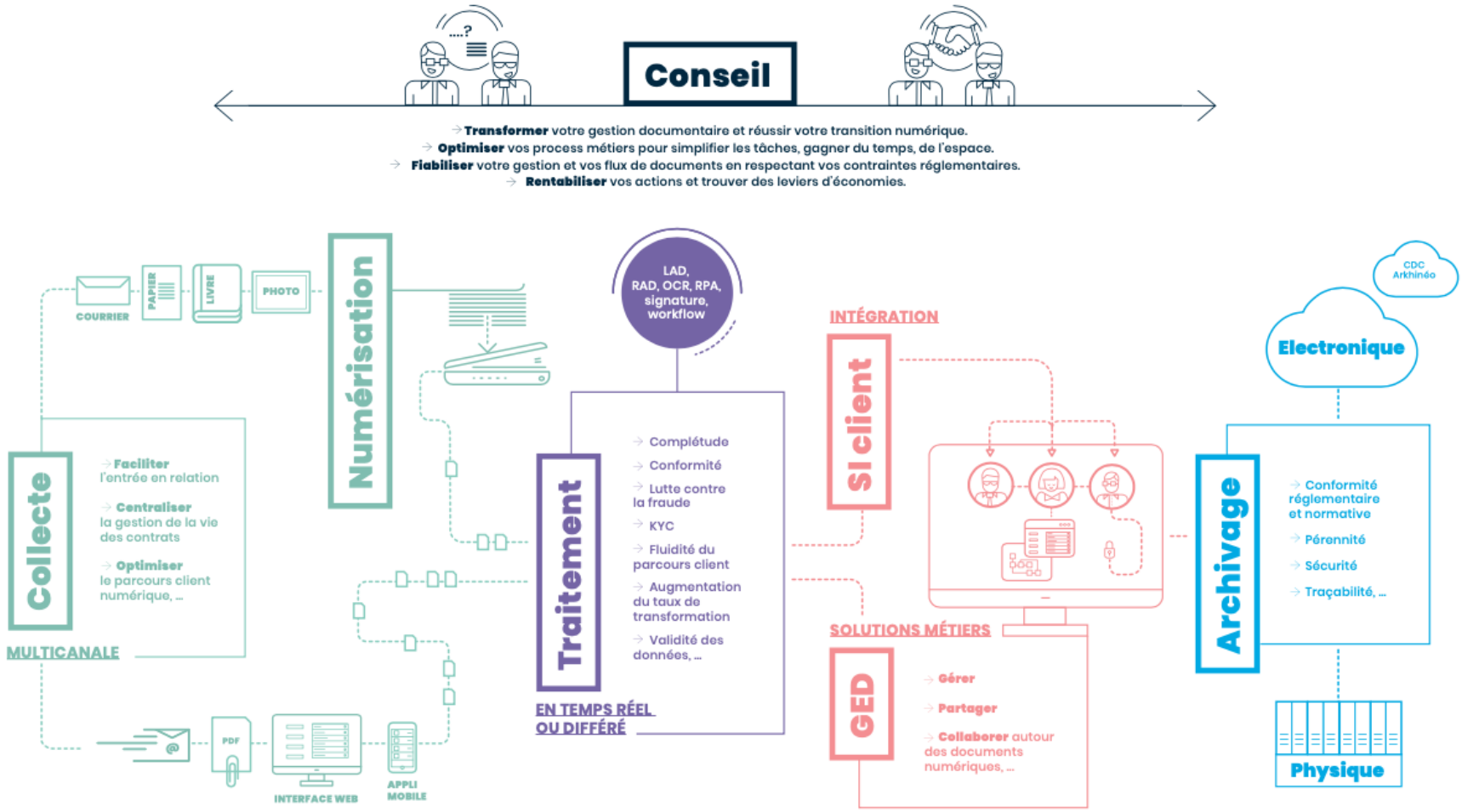
(D'après le site web de l'entreprise, 2022)

4.1.5. Entreprise E

L'entreprise E est une société française par actions simplifiée (SAS) dont le siège social est établi à Rillieux-La Page (Lyon, France), mais qui est aussi active en Suisse, en Espagne et en île-Maurice.

La principale activité de l'entreprise est la gestion du flux documentaire : collecte des archives physiques chez le client, stockage sécurisé dans les locaux de l'entreprise, dématérialisation, archivage physique et électronique, destruction. Les principaux services offerts par l'entreprise E sont illustrés à la **figure 6**.

Figure 6 : Principaux services offerts par l'entreprise E



(Document fourni par l'entreprise, 2020)

4.1.6. Entreprise F

L'entreprise F est une entreprise privée spécialisée dans l'édition et intégration des logiciels pour la gestion électronique des documents, et dans la prestation de services informatiques (hébergement informatique des données-clients dans les datacenters en Suisse, gestion de l'infrastructure et réseaux, formations spécialisées).

L'entreprise dont le siège se trouve à Sierre (Valais, Suisse), est aussi active en France et au Canda. Elle offre des solutions logicielles répondant aux enjeux métiers dans les domaines de la dématérialisation des documents, de la gestion des ressources humaines et de l'infrastructure et cloud.

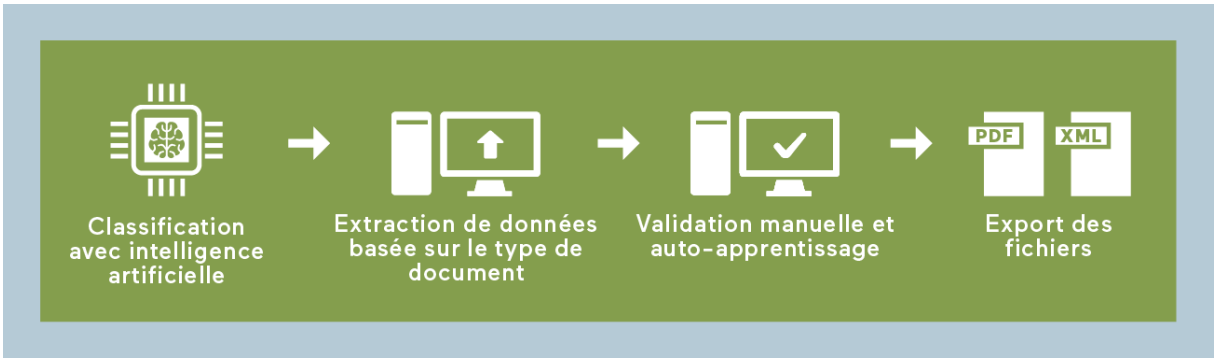
4.1.7. Entreprise G

L'entreprise G est une société suisse prestataire de services informatiques dont les solutions proposées permettent à ses clients de numériser et d'automatiser les processus de gestion des documents y compris l'archivage électronique. L'entreprise compte à peu près 90 collaborateurs répartis entre les sites de Zurich (siège de l'entreprise) et celui de Lausanne.

Les principaux services offerts par l'entreprise G sont les suivants :

- **Input Management :** saisie numérique de données provenant de diverses sources en vue de leur traitement ultérieur. Ce service comprend aussi la numérisation et la capture. D'après le responsable régional (bureau de Lausanne), la **capture*** ou l'extraction automatique d'information sur les documents numérisés en vue d'automatiser les processus, constitue la valeur ajoutée de l'entreprise (voir **figure 7**).

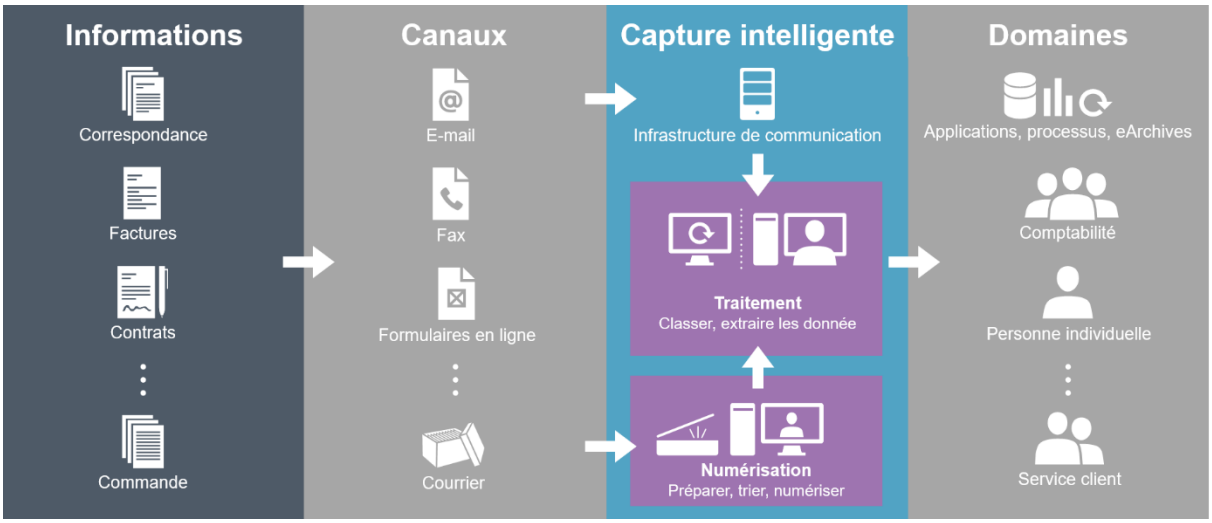
Figure 7 : Capture : classification et extraction des données par des solutions logicielles de l'entreprise G



(Document fourni par l'entreprise, 2022)

D'après le responsable régional, « La capture intelligente permet d'étendre le traitement automatique aux documents non structurés. » (**figure 8**).

Figure 8 : Capture intelligente : traitement hautement standardisé des documents par l'entreprise G, quel que soit le format d'entrée



(Dufey, 2018)

- **Enterprise Content Management** ou gestion de contenu d'entreprise (ECM, Enterprise Content Management) : en tant qu'extension d'une GED, ce service comprend les stratégies et les technologies destinées à capturer, stocker, retrouver, distribuer, conserver ou détruire les documents et les contenus.
- **Traitement de factures / P2P (Purchase to Pay) :** traitement automatique des factures fournisseurs depuis la commande jusqu'à l'achat.
- **Workflow & Case Management :** service pour l'automatisation des processus et des cas liés aux documents.
- **Digital Mailroom :** organisation et optimisation de tout le courrier qu'une entreprise peut recevoir sur différents canaux. Le but est d'externaliser complétement la réception du courrier, la numérisation et la capture, solution idéale pour les entreprises disposant de gros volumes et plusieurs sites.
- **Dossier électronique du personnel :** mise en œuvre de solutions de gestion électronique de documents pour gérer les dossiers du personnel de façon digitale.
- **Numérisation :** digitalisation de dossiers papier existants et de numérisation continue des documents.
- **Archivage probant des documents :** grâce aux technologies GED/ECM efficaces, les solutions d'archivage probant proposées permettent aux clients d'accéder à leurs documents numériques essentiels en ligne de manière rapide et fiable.

Tableau 9 : Informations générales sur les entreprises

	Numérisation patrimoniale				Numérisation mixte	Numérisation documentaire classique	
	Entreprise A	**Entreprise B**	**Entreprise C**	**Entreprise D**	**Entreprise E**	**Entreprise F**	**Entreprise G**
Année de création de l'entreprise	1999	2003	1998	2014-2015	1989	1983	2006
Siège social	Bordeaux	Baden (Argovie)	Ecublens (Lausanne, Vaud)	Burgdorf, laboratoire photo à Berthoud (Berne)	Rillieux-La-Pape (Lyon-France)	Sierre (Valais), succursale à Renens (Vaud)	Zurich
Succursales ou ateliers permanents	**Genève** (Vernier), Paris et Berlin	Yverdon-les-Bains (Vaud)	Non	Non	Morges (Suisse), Chartres et Chalon-sur-Saône (France), Madrid (Espagne)	France (Lyon) et Canada (Montréal)	Lausanne
Zone couverte par les activités de l'entreprise	Mondiale	Toute la Suisse	Suisse romande	Suisse alémanique	France, Espagne, Suisse	France, Suisse romande, Canada	Toute la Suisse
Types de documents numérisés	Tous les fonds patrimoniaux, y compris des objets.	Tous les fonds patrimoniaux, y compris des objets, sauf vidéos et audio	Tous les fonds patrimoniaux, principalement des livres reliés	Documents iconographiques des bibliothèques et archives	Tous les fonds patrimoniaux sauf vidéos et audio. Tous les documents classiques	Documents classiques, principalement les factures et les contrats de vente	Tous les documents classiques, sauf des livres
Lieu de numérisation	Chez le client (ateliers de numérisation in situ) ou locaux de l'entreprise	Locaux de l'entreprise	Locaux de l'entreprise	Locaux de l'entreprise pour les privés*	Locaux de l'entreprise (90%) et chez le client (10%)	Locaux de l'entreprise	Locaux de l'entreprise et chez le client

*Se déplace seulement pour la Bibliothèque nationale suisse qui met à sa disposition un atelier de numérisation bien équipé.

Tableau 10 : Informations générales sur les entreprises (suite)

	Entreprise A	**Entreprise B**	**Entreprise C**	**Entreprise D**	**Entreprise E**	**Entreprise F**	**Entreprise G**
Principaux clients	Établissements culturels patrimoniaux (publiques ou privés) : archives, bibliothèques, cinémathèques, etc.	Entreprises publiques, principalement des communes (60%), un peu d'institutions privées	Établissements culturels patrimoniaux, principalement des bibliothèques publiques (90%)	Établissements culturels patrimoniaux, principalement des bibliothèques, archives et imprimeries (70%). Entreprises et personnes privées pour le reste.	Tous les établissements publiques ou privés : établissements culturels patrimoniaux, banques, assurances, entreprises privées du commerce, etc.	Entreprises publiques, principalement des communes (environ 65%) et privées.	Toutes les catégories de sociétés (finance, immobilier, industrie, etc.) de 100 à 500 collaborateurs
Services autres que numérisation ?	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Entreprise inscrite sur la plateforme digicoord.ch**	Oui	Non	Oui	Oui	Oui	Non	Non

** Plateforme d'informations sur les projets de numérisation en Suisse (mise à disposition par la Bibliothèque nationale suisse et RERO, réseau des bibliothèques de Suisse occidentale).

4.1. Typologie des documents numérisés

Pour la typologie des documents basée sur le support physique (**tableaux 11 et 12**), nous avons adopté la classification proposée par la Bibliothèque et Archives nationales du Québec (BANQ), la Bibliothèque nationale de France (BnF) et le Musée canadien de l'histoire (MCH) dans leur ouvrage intitulé ″Recueil de règles de numérisation″ et publié en 2014. Cette classification regroupe les documents présents dans des collections patrimoniales en trois familles de support :

- les documents sur support opaque ;
- les documents sur supports transparent ;
- et les objets.

Tableau 11 : Typologie des documents numérisés par les entreprises

	Numérisation patrimoniale				Numérisation mixte	Numérisation documentaire (ou de production)	
	Entreprise A	**Entreprise B**	**Entreprise C**	**Entreprise D**	**Entreprise E**	**Entreprise F**	**Entreprise G**
I. Documents sur support opaque							
A. Documents imprimés							
Livres reliés ou non, dictionnaires	X	X	X	-	X	-	-
Incunables	X	X	X	-	X	-	-
Périodiques	X	X	X	-	X	-	-
Catalogues, documentations techniques	X	X	X	-	X	-	-
Bulletins, journaux, Magazines, Revues	X	X	X	-	X	-	-
Presse ancienne	X	X	X	-	X	-	-
Brochures	X	X	X	-	X	-	-
Fonds d'archives, liasses	X	X	X	-	X	-	-
Prospectus	X	X	X	X	X	-	-
B. Documents manuscrits							
Manuscrits médiévaux	X	X	X	-	X	-	-
Registres	X	X	X	-	X	-	-
Registres fonciers	X	X	X	-	X	-	-
Registres : délibérations, état civil, matricules militaires, etc.	X	X	X	-	X	-	-
Correspondance (lettres…)	-	X	X	-	X	-	-
Factures des fournisseurs	-	-	-	-	X	X	X
Contrats de vente	-	-	-	-	X	X	X
Dossiers du personnel	-	-	-	-	X	X	X
Notes de frais	-	-	-	-	X	X	X
C. Documents cartographiques							
Cartes (géologiques, topographiques, etc.)	X	X	X	X	X	-	-
Plans (du cadastre, autres)	X	X	X	X	X	-	-
Dessins (techniques, d'architecture)	X	X	X	X	X	-	-

Tableau 12 : Typologie des documents numérisés par les entreprises (suite)

	Entreprise A	Entreprise B	Entreprise C	Entreprise D	Entreprise E	Entreprise F	Entreprise G
D. Documents iconographiques							
Affiches	X	X	X	X	X	-	-
Cartes postales	X	X	X	X	X	-	-
Estampes	X		X	X	X	-	-
Gravures, peintures, tableaux, dessins	X	X	X	X	X	-	-
Timbres-poste	X	X	X	X	X	-	-
Tirages photographiques (couleur, noir et blanc)	X	X	X	X	X	-	-
Cartes à fenêtre	X	X	X	X	X	-	-
II. Documents transparents (iconographiques)							
Ektas, négatifs, négatifs sur plaques de verre, diapositives (noir et blanc, couleur)	X	X	-	X	X	-	-
Microformes (microfiches, microfilms)	X	X	-	X	X	-	-
Film photographique	X	-	-	X	X	-	-
Fonds audiovisuels	X	-	-	X	X	-	-
III. Objets							
Objets en 2D, 360 °, 3D	X	X	-	-	X	-	-

4.2. Matériel de numérisation

Les entreprises interviewées disposent de matériel de numérisation très varié : scanners à plat, scanner vertical, scanners à chargeur, scanners à tambour, scanners de livres, scanners à plans, scanners à microfilm et appareils photo numériques (**tableau 13**), mais c'est le scanner à plat qui est bien représenté. Dans ce tableau, les chiffres entre parenthèses indiquent le nombre de scanners que chaque entreprise possède. Parmi ces scanners, qui couvrent aussi un large éventail de formats, la majorité numérisent en mode recto-verso.

Il est à remarquer que seuls les scanners type I2S, 4Digitalbooks et Fujitsu embarquent avec eux leurs logiciels de capture de l'information. Il s'agit du matériel à degré de perfectionnement technologique élevé, donc très professionnel. Ces types de scanners ont aussi l'avantage d'avoir des logiciels capables de compter automatiquement le nombre de pages.

Ce qui est important de rappeler ici, c'est que le choix du scanner et de son logiciel doit être guidé par les objectifs du projet de numérisation. Il faut aussi souligner que le scanner à tambour, où le document doit s'enrouler autour d'un cylindre tournant rapidement, offre une résolution et une qualité d'image sans égal, mais qu'il est très peu utilisé de nos jours, car il n'est pas adapté à tous les types de documents (qui ne doivent pas dépasser 1 mm d'épaisseur). De plus, il est lent et son coût est très élevé.

L'entreprise A possède deux types de scanners : I2S et son logiciel Limb Capture pour la numérisation de tout type de documents, et Phase One ainsi que son logiciel propriétaire pour la numérisation des archives iconographiques de petit format. Limb Capture est un logiciel d'interface conçu pour fonctionner sur tous les scanners I2S. Le logiciel propriétaire qui pilote le scanner Phase One est un logiciel de suivi de la production qui permet d'avoir une gestion uniforme sur l'ensemble de la phase de numérisation.

L'entreprise B utilise deux types de scanners : un scanner à plat et un appareil photo numérique qui ne comptent pas automatiquement le nombre de page. Cette entreprise nous a communiqué les critères de choix d'un bon logiciel de numérisation, à savoir la facilité d'utilisation, un large éventail d'options de personnalisation, le contrôle des paramètres individuels, le degré de diffusion, le support et la précision du taux de reconnaissance de texte (tests de comparaison).

L'entreprise C possède une dizaine de scanners dont la majorité sont une fabrication de l'entreprise, une fabrication de l'entreprise. Ces scanners peuvent être à plat ou en V et numérisent à une vitesse maximale de 1500 pages par heure. Ils sont munis d'un dispositif qui permet de tourner automatiquement les pages des livres. Ils sont aussi automatiques ou semi-automatiques pour le traitement de l'image. L'entreprise possède aussi les scanners type I2S, de l'entreprise française du même nom et partenaire de l'entreprise C, ainsi que le type Fujitsu, tous à plat. Certains scanners sont posés sur des tables aspirantes, c'est-à-dire munis de petits trous qui permettent de déplier les plans (les mettre à plat). Les logiciels qui pilotent ces scanners sont souvent associés au scanner, et quand on achète le scanner, on achète en même temps le logiciel. Une des qualités de ces logiciels, comme Kofax, c'est la possibilité de regrouper plusieurs documents en un seul dossier en utilisant des algorithmes.

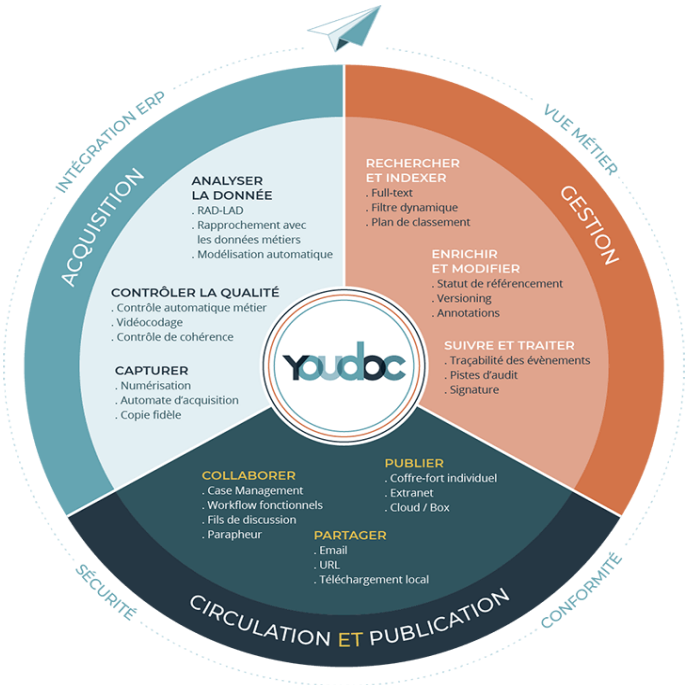
L'entreprise D utilise deux types de scanners : à plat pour les photos et à tambour pour les négatifs et diapositives. Nexscan F4200 est un scanner couleur 48 bits doté d'un CCD couleur trilinéaire avec des éléments de 8'000 pixels et une résolution physique maximale de 5080 dpi (F4100) ou 7040 dpi (F4200). Le Scanner à tambour est un scanner très spécial qui numérise à très haute résolution. Avec ce type de scanner, l'entreprise D est en mesure de préparer des diapositives, des négatifs et également des images couleur de petits formats (24 mm x 36 mm) pour l'impression grand format (430 cm x 297 cm). Les scanners de l'entreprise D sont également capables de détramer les images d'anciennes brochures afin qu'elles puissent être réutilisées pour des imprimés ou des photographies.

L'entreprise E possède quatre types de scanners dont trois à plat (Kodak Capture Pro, Contex pour les plans, I2S pour les livres) et un à chargeur (ScanPro pour les microfiches et microfilms) dont les logiciels ne sont pas associés. La diversité de ces scanners s'explique par le fait que l'entreprise numérise à la fois les documents patrimoniaux et les documents classiques.

L'entreprise F utilise le scanner Kodak Alaris équipé du logiciel Youdoc (**figure 9**) que l'entreprise vend à ses clients pour numériser différentes sortes de documents, principalement les factures fournisseurs et les contrats de vente. Le logiciel Youdoc inclut trois modules (suite d'outils) suivants :

- un module d'acquisition-capture intelligente de documents multiformats ;
- un module de gestion, de classement, de recherche et de stockage des documents ;
- et un module dédié à la circulation et au partage des documents.

Figure 9 : Présentation du logiciel multi-module Youdoc de l'entreprise F



(D'après le site web de l'entreprise, 2022)

Depuis 2018, d'autres fonctionnalités du logiciel se sont ajoutées : indexation des documents non structurés, gestion des droits sur les documents, calendrier de conservation et un module de traçabilité pour répondre aux enjeux du RGPD (Règlement général sur la protection des données).

L'entreprise G utilise les scanners Kodak et Opex, tous deux pilotés par le logiciel DocProstar de TCG Informatik AG. Ce logiciel utilise l'intelligence artificielle pour capturer, classifier et extraire automatiquement, non seulement les données structurées, mais aussi les données non-structurées.

Tableau 13 : Scanners et logiciels de numérisation utilisés par les entreprises

Entreprise	Scanner	Logiciel de pilotage du scanner	Logiciel embarqué dans le scanner ?	Comptage automatique de pages ?	Formats acceptés par le scanner	Mode de numérisation	Numérisation recto-verso ?
Entreprise A	I2S pour des livres (marque CopyBook, DigiBook) (**N= 40-50**)	Limb capture	Oui	Oui	Tout type de documents car toute une gamme de scanners	Â plat ou bien ouverture restreint ou bien des objets, …	Non, manuel
	Phase One (**N=nc**)	Propriétaire	nc	Non	Iconographie petit format	Â plat	Non, manuel
Entreprise B	Appareil photo NIKON D850 (**N=1**)	Capture One	Non	Non	Jusqu'au A2	Caméra planétaire	Oui
	EPSON (**N=1**)	Silverfast	Non	Non	Jusqu'au A4	À plat	Oui
Entreprise C	4Digitalbooks à livres (**N=10**)	4Digitalbooks	Oui	Oui	Jusqu'au A2	À plat et en V	Oui
	I2S à livres (**N= 2**)	I2S	Oui	Oui	Jusqu'au A2, A0	A plat	Oui
	FUJITSU à feuilles (**N= 2**)	Kofax Capture Pro	Non	Oui	Jusqu'au A3	A plat	Oui
Entreprise D	Heidelberg Nexscan F4100 et F4200 (**N=4**)	Newcolor 7000	Non	Non, mais protocole détaillé pour chaque scan	A3+	À plat	Oui
	Heidelberg Topaz (**N= 2)**	Newcolor 7000	Non	Non, mais protocole détaillé pour chaque scan	A3+	À plat	Oui
	Heidelberg Primescan D7100, D8200, D8400 (**N= 3**)	Newcolor 7000	Non	Non, mais protocole détaillé pour chaque scan	A3+	À tambour	Oui
Entreprise E	KODAK Capture Pro i5600 (**N=4**)	Kodak Capture Pro	Non	Oui	Jusqu'au A3	À plat	Oui
	Contex pour les plans (**N=1**)	HD Ultra	Non	Non	A0	À plat	Non
	I2S (marque CopyBook, DigiBook) pour des livres (N**=1**)	Copy book	Non	Oui	Jusqu'au A3	À plat	Oui
	ScanPro Microfilms (**n=1**)	Autoscan	Non	Non	Microfilms, microfiches	Chargeur	Non
Entreprise F	KODAK Alaris (**N=3**)	Youdoc Acquisition	Non	Oui	Tous les formats	Vertical	Oui
Entreprise G	Kodak (**N>10**)	DocProstar	Non	Oui	Jusqu'au A3	A bac	Oui
	Opex (**N= 3**)	DocProstar	Non	Non	Dépend du modèle	Tapis roulant	Oui

nc : non communiqué

4.3. Paramètres techniques de numérisation

Toutes les entreprises sondées numérisent leurs documents avec une résolution optique de sortie de 300 dpi (**tableau 14**). La profondeur des couleurs ou profondeur de codage ou d'échantillonnage (24 bits, 16 bits, 8 bits et 1 bit) dépend, quant à elle, du client. Pour presque toutes les entreprises, TIFF est le format de sortie de l'image après scannage et sur lequel est effectué l'océrisation. C'est aussi à partir de ce format qu'on obtient d'autres formats. PDF ou PDF/A est partout le format de sortie de l'océrisation. L'entreprise C utilise parallèlement le format TIFF et JPEG2000 ainsi que le format **XML-ALTO***. Le format de stockage et de conservation est TIFF pour les entreprises faisant de la numérisation patrimoniale, alors que le format PDF est privilégié par les entreprises qui numérisent les documents classiques. Le format de consultation en ligne est partout JPEG ou PDF. Le type et niveau de compression des fichiers dépendent des besoins du client, mais quand c'est du TIFF, soit c'est non compressé ou compressé avec l'algorithme Lzw. Dublin Core est le schéma de métadonnées largement utilisé par les entreprises interviewées. Le disque dur externe ou le protocole FTP sont les deux moyens très utilisés pour restituer les données numériques aux clients.

Chez l'entreprise A, le niveau de résolution (200-4000 dpi) dépend des demandes de l'établissement ; celles-ci dépendant du document/objet et de la capacité de stockage de l'entreprise. Il en est de même pour la profondeur des couleurs (RVB 24 bits, ton de gris 8 bits, ton de gris 16 bits, 8 bits couleurs indexées et 1 bit noir et blanc), le type et le niveau de compression des fichiers images. Le format de sortie de l'image est RAW (brut), qui regroupe plusieurs formats d'images numériques sortant d'appareils photo numériques ou de scanners. Ces formats possèdent des données brutes provenant des capteurs. TIFF est le format sur lequel est effectué l'OCR, et PDF en est le format de sortie. Dublin Core EAD est le schéma des métadonnées, TIFF ou JPEG2000 sont les formats de conservation. Pour la consultation en ligne, l'entreprise A privilégie trois formats : JPEG, PDF ou ALTO. Pour restituer les données numériques à ses clients, l'entreprise utilise plusieurs canaux : disques durs, NAS, FTP (File Transfer Protocol), AWS (Amazon Web Services)…

L'entreprise B suit les recommandations de Mémoriav (2017) qui recommande de respecter les quatre paramètres suivants : résolution, taille, échantillonnage ou profondeur de couleur et mode chromatique. Les clients de l'entreprise B ont la possibilité d'utiliser sa solution docuteam cosmos pour conserver les données numériques restituées après numérisation.

L'entreprise C utilise, pour la colorimétrie, des mires de Calibration de couleurs de la Société X-RITE, mais aussi des mires FADGI et Metamorfoze. Comme le souligne son directeur, l'entreprise a « mis énormément d'énergie » pour développer des compétences en colorimétrie[16], et c'est peut-être une des raisons qui fait que l'entreprise est considérée par certains comme une référence en numérisation en Suisse romande. L'entreprise C utilise la compression Visually Lossless JPEG qui permet une compression très faible sur les pixels (sur la luminescence), mais qui n'affecte pas la qualité de l'image et qui donne un fichier « 5 fois plus petit que le TIFF en général », déclare le directeur de l'entreprise C. Selon lui,

[16] Entretien du 27 mai 2022

« l'œil humain est capable de distinguer tous les cinq niveaux d'un ton de gris. (…),mais entre 2 niveaux, la transition est tellement faible qu'il ne peut pas le voir. » Visually Lossless JPEG

> « *fait une compression sur ça, sur cette luminescence et permet de prendre un niveau sur 5, mais tout en faisant une image extrêmement fidèle », précise-t-il, tout en ajoutant que « ça permet de faire d'énormes gains de stockage.* »

Lors de l'océrisation, quand le client demande du PDF, l'entreprise C fait toujours un PDF en double couche, c'est-à-dire une image numérisée (première couche visible à l'écran) accompagnée d'un PDF avec texte caché (deuxième couche).

L'entreprise D applique les paramètres techniques de numérisation de la Bibliothèque nationale suisse (BN). L'entreprise numérise les documents de ses clients avec une résolution de sortie de 300 dpi. Un minimum de 400 dpi est requis pour les tirages d'art sur papier couché. Des résolutions plus élevées sont également possibles. En principe, la profondeur de couleur est de 16 bits par canal, soit 48 bits en mode RVB (3 canaux x 16 bits) et pas de compression du fichier images TIFF. Les outils de l'entreprise permettent de supprimer la trame, par exemple sur un prospectus : ajustement ou modélisation des données et leur envoi à l'imprimerie pour l'impression d'un nouveau prospectus. En attendant leur restitution au client (dans une période ne dépassant pas 6 mois), les images numérisées sont conservées sur serveur en backstop avec une copie dans un coffre-fort en banque.

L'entreprise E numérise ses documents, généralement avec une résolution de 300 dpi, mais précise que ce niveau dépend du client, de son objectif et de l'espace de stockage dont il dispose. L'entreprise numérise en mode couleur (24 bits), en mode niveaux de gris (8 bits) et en mode bitonal (1 bit). Le format de sortie est TIFF ou PDF selon les cas, mais dépend du client. Le type de compression du format TIFF est le Groupe G4 (méthode de compression des images en noir et blanc sans perte de qualité). Le format des métadonnées est XML. La restitution des données au client se fait au moyen de File Transfer Protocol, de disques durs ou elles sont directement transférées au portail du client.

L'entreprise F numérise ses documents avec une résolution de 300 dpi en mode couleur ou en gris. La compression utilisée, MRC (Mixed Raster Content) est une méthode de compression d'images contenant à la fois du texte et des éléments à tons continus, qui utilise la segmentation pour améliorer le niveau de compression et la qualité de l'image.

L'entreprise G numérise ses documents avec une résolution de 300 dpi en mode couleur (16 bits). TIFF est le format de sortie de l'image et sur lequel est effectué l'océrisation. PDF ou PDF/A est le format de stockage et de consultation, le niveau de compression dépend des projets des clients et l'entreprise restitue les données aux clients via le SFTP (Secure File Transfer Protocol).

Tableau 14 : Paramètres techniques de numérisation et caractéristiques techniques des fichiers images

Paramètre	Entreprise A	Entreprise B	Entreprise C	Entreprise D	Entreprise E	Entreprise F	Entreprise G
Résolution optique de sortie	200-4000 dpi	300-1200 dpi suivant la taille du format original (le A2 ça sera 300, 1200 ça sera A6)	300 dpi (98% des cas)	300 dpi pour l'imprimé	300 dpi	300 dpi	300 dpi
Profondeur de codage et mode colorimétrique	Tout dépend de la demande client	24 bits eciRGBv2, bzw, 8-bit Grayscale Gamma 2.2, bzw bitonal	ça dépend du client	16 bits en couleur	24 bits en couleur, 8 bits niveau de gris, 1 bit noir et blanc	Numérisation en couleur ou en gris.	16 bits en couleur
Format de sortie de l'image numérisée	**RAW***	TIFF ou JPEG 2000 (selon les besoins des clients)	JPEG (70% de production annuelle) TIFF (30% de production annuelle)	TIFF avec 16 bits en couleur	PDF/A (99%) pour Kodak (format A4/A3). Tiff ou PDF pour les scanners à plans (A2, A1, A0), ça dépend des clients	TIFF, JPEG	TIFF (temporaire)
Format sur lequel est effectué l'OCR	TIFF	TIFF	JPEG (70%), TIFF (30%)	TIFF	TIFF	TIFF noir blanc	TIFF
Format de sortie de l'OCR	PDF	PDF/A	PDF avec texte caché, PDF ou XML (XML ALTO)	PDF	PDF	PDF/A	PDF ou PDF/A
Type et niveau de compression	Dépend de la demande du client	Non compressé ou TIFF Lzw, jpeg (selon les besoins des clients)	JPEG à 95% de qualité	Pas de compression	TIFF groupe G4	MRC 75%	Ça dépend des projets
Format ou Schéma de métadonnées	Dublin Core, EAD	XML	Dublin Core	Dublin Core	Dublin Core	XML/JSON	XML/JSON
Format de stockage/conservation	TIFF ou JPEG 2000	TIFF	TIFF	TIFF	PDF	PDF/1 compressé (compression inversée)	PDF
Format de consultation en ligne	JPEG, PDF avec texte caché ou ALTO.	JPEG, PDF	JPEG, PDF	JPEG (8 bits)	JPEG, PDF	JPEG, PDF	PDF
Support de stockage d'images numérisées avant leur restitution au client	Disques durs, NAS, FTP, AWS, …	docuteam Cosmos sur NAS	Disque dur, Serveur	Serveurs, coffre-fort en banque	Serveurs	Serveurs dans une GED	Serveurs
Support pour la restitution des données numériques au client	Disques durs, Serveurs NAS, FTP, AWS…	Disques durs externes ou clés USB	Disques durs prêtés ou vendus aux clients	Disques durs ou clé USB	FTP(s)	-	SFTP

4.4. Etapes du workflow de numérisation

4.5.1. Entreprise A

Le processus de numérisation de l'entreprise A débute par une discussion en amont avec le client pour déterminer la nature et les caractéristiques des prestations demandées (**tableau 15**). Et puis, l'entreprise reçoit les collections à numériser accompagnées d'un inventaire. Normalement, les établissements culturels patrimoniaux s'occupent de l'inventaire de leurs fonds avant leur transfert au prestataire de numérisation, précise le représentant de l'entreprise et responsable marketing et business développement[17]. L'entreprise A ne reçoit que le **fichier de récolement** et toutes les indications concernant les règles de numérisation (le nombre de pages, numérisation recto-verso, dpi, métadonnées, etc.). L'entreprise ne fait l'inventaire que sur demande du client et sur des critères définis en commun accord avec celui-ci

Le scannage proprement dit est précédé par la phase de tests dont le résultat est envoyé au client pour validation. Le **récolement*** est effectué dès la phase de numérisation et éventuellement complété durant les étapes de contrôle qualité.

L'OCR est fait avec le logiciel ABBYY en quatre étapes : 1) OCR brut, 2) Segmentation du document ou de la page, 3) Indexation manuelle, et 4) Relance de l'OCR. Mais ce n'est pas forcé que l'OCR et indexation manuelle soient concomitants, souligne le représentant de l'entreprise A, car cela dépend du taux OCR souhaité par le client. Le taux de l'OCR brut en automatique peut ou ne pas correspondre à ce qui est attendu par le client. Dans ce cas, l'entreprise recourt à l'indexation manuelle pour atteindre ce taux. Par exemple, lorsque l'entreprise A numérisé les documents d'archives de presse de l'Union Internationale des Télécommunications (UIT) à Genève, l'OCR brut en automatique a reconnu jusqu'à 92% alors que l'UIT voulait un taux de 96%. Pour arriver à ce taux, l'entreprise a procédé à une indexation manuelle avec segmentation pour identifier où étaient par exemple les éléments textes sur des publicités, des couvertures, etc. Il s'agit d'une prestation supplémentaire qui « est peu demandée parce que ça coûte relativement cher », souligne le représentant de l'entreprise, tout en précisant que « la plupart de clients vont se contenter de l'OCR en mode brut. »

Les difficultés rencontrées en océrisation par l'entreprise sont liées à la qualité physique du document original : présence de trous, pages déchirées, encre qui s'est effacé, la transparence du papier qui laisse apparaître la typo du verso, etc.

[17] Entretien du 3 mai 2022.

Tableau 15 : Etapes du workflow de numérisation de l'entreprise A

N°	Etapes
	Prénumérisation
1	Discussion en amont du projet avec le client pour établir les caractéristiques des livrables
2	Inventaire et reconditionnement (si le client le demande)
3	Préparation manuelle des documents : retirer les agrafes, les trombones, les post-it, les formats inappropriés, etc. (si le client le demande)
	Numérisation
4	Phase de tests et envoi du résultat au client pour appréciation et validation
5	Scannage proprement dit
6	1er contrôle qualité : contrôle visuel fait par un opérateur ou une opératrice de numérisation
7	Récolement
8	2ème contrôle qualité : auto-détections automatiques des images avec zéros kilo-octets (0 ko)
	Post-numérisation
9	Océrisation : OCR brut ou raffiné selon les besoins du client
10	Indexation manuelle
11	3ème contrôle qualité fait par le responsable qualité
12	4ème contrôle qualité fait par le chef de projet
13	Récolement
14	Restitution des livrables

4.5.2. Entreprise B

La première étape du processus de numérisation de l'entreprise B (**tableau 16**), consiste à mener des discussions en amont pour être sûr de ce que le client veut comme produit final (résolution, format de sortie, nommage des fichiers, etc.). Vient ensuite l'inventaire qui est « une condition sine qua non vu que les clients ne viennent pas chez nous pour la numérisation, mais pour l'archivage », déclare la représentante de l'entreprise et conseillère en gestion de l'information[18]. « On fait normalement de la description (des inventaires d'archives) sans numérisation en parallèle, en revanche on ne fera pas de la numérisation seule. », souligne-t-elle. L'entreprise B utilise, pour la description archivistique, la norme internationale ISAD(G) avec encodage en XML (**annexe 6**), mais il arrive aussi qu'un client lui propose un autre outil de description. La préparation manuelle et intellectuelle des documents est suivie de leur reconditionnement dans des fourres non acides. Plus précisément, les opérations d'inventaire, de préparation matérielle et de reconditionnement se font simultanément.

L'entreprise B ne procède aux tests de pré-scannage que lorsque le volume des documents est grand ; et le résultat est envoyé au client pour validation. L'OCR plein texte est fait après numérisation avec le logiciel ABBYY ; et le taux de qualité souhaité par les clients est en général supérieur à 98 %. Après océrisation, l'opérateur ou l'opératrice procède à une relecture (exhaustive ou par échantillonnage) du fichier texte généré. Mais l'océrisation n'est

[18] Entretien du 16 mai 2022.

pas faite de façon systématique pour les clients (l'OCR est plus utilisé à l'interne, par exemple pour les anciens inventaires d'archives qui sont dactylographiés[19]).

L'entreprise B ne fait pas d'indexation à proprement parler, car elle n'est pas demandée par les clients, déclare sa représentante, qui précise que l'indexation est une partie de la description qui va jusqu'au niveau du dossier. Par exemple pour l'image, on relève tout juste son titre.

Tableau 16 : Etapes du workflow de numérisation de l'entreprise B

N°	Etapes
	Prénumérisation
1	Discussion en amont du projet avec le client pour établir les caractéristiques des livrables
2	Prise en charge des archives chez le client : mise en conteneurs
3	Transport des archives vers les locaux de l'entreprise
4	Inventaire des archives dans l'application docuteam curator (description archivistique selon la norme ISAD(G)) ou l'outil proposé par le client
5	Préparation matérielle et intellectuelle des documents : retirer les agrafes, les trombones, les post-it, les formats inappropriés, pose de code-barres, de séparateurs, etc.
6	Reconditionnement (mise en fourres non acides)
	Numérisation
7	Phase de tests (si grand volume) et envoi du résultat au client pour appréciation et validation
8	Scannage proprement dit
9	Nommage des fichiers
10	Contrôle automatique (pointage) de la quantité et de la qualité des images numérisées
	Post-numérisation
11	Océrisation si demandée
12	Récolement
13	Transfert vers le serveur NAS de l'entreprise
14	Restitution d'images numérisées et des documents physiques au client (dans des fourres non acides)

Selon sa représentante, l'entreprise B accorde de :

> « *l'importance à comment conserver ensuite les résultats de la numérisation (les fichiers numériques) pour éviter de devoir refaire ce travail dans quelques années, et donc l'importance de réfléchir à une solution pour l'archivage électronique en parallèle des travaux de numérisation proprement dit.* »[20]

4.5.3. Entreprise C

Le processus de numérisation de l'entreprise C (**tableau 17**) commence par analyser la demande du client. Si celle-ci n'est pas claire, l'entreprise qui va voir le client pour clarifier la situation. Cette discussion permet à l'entreprise C d'établir une offre beaucoup plus riche que

[19] Plutôt que de les retranscrire manuellement, ils sont scannés et puis le texte PDF qui en sort est océrisé. L'entreprise utilise aussi des stagiaires pour faire cette retranscription d'inventaire.
[20] Courrier électronique du 30 mai 2022.

la demande initiale et qui rassure le client. En effet, lors de cette discussion, l'entreprise C se rend souvent compte que le client veut beaucoup plus par rapport à sa première demande.

L'entreprise C ne fait pas d'inventaire, car selon son directeur, elle n'a pas de compétence dans ce domaine. Par conséquent, l'entreprise exige à ses clients de lui livrer, en même temps que les documents à numériser, le fichier d'inventaire sous format Excel. Après réception des documents, l'étape suivante consiste en un contrôle de ce fichier. En effet, deux cas de figure peuvent se présenter : un document peut être listé sur l'inventaire alors qu'en réalité il absent (c'est-à-dire non livré), ou un document peut être présent physiquement (c'est-à-dire livré) mais absent de la liste d'inventaire.

Pour la préparation des documents (manutention), deux tâches sont importantes : l'insertion des séparateurs ou intercalaires avec les mentions START et END ainsi que la pose des code-barres. L'entreprise C procède aussi à une phase de tests suivie d'un envoi des résultats au client pour validation. Après numérisation, les images sont structurées manuellement en lots : par exemple pour constituer des chapitres en cas de livres, de mensualités en cas de journal de presse ou de parutions s'il s'agit d'un hebdomadaire.

Le nommage des fichiers se fait selon les indications figurant dans le fichier d'inventaire. S'il s'agit par exemple, d'un périodique, c'est la date (année-mois-jour) ; pour les livres, l'identifiant peut être son ISBN ou un code (pas en tout cas le titre du livre, car celui-ci peut être dans une autre langue comme le chinois, dont les caractères ne sont pas reconnus).

La gestion des métadonnées associés aux images numérisées revient au client qui, à travers le fichier d'inventaire, doit donner un identifiant à chaque document. Lorsqu'elle numérise les documents, l'entreprise C ne fait que reproduire ça en utilisant un code-barres, c'est-à-dire un lien entre le livrable et le système d'information du client.

Tableau 17 : Etapes du workflow de numérisation de l'entreprise C

N°	Etapes
	Prénumérisation
1	Discussion en amont du projet avec le client pour établir les caractéristiques des livrables
2	Réception des documents physiques par l'entreprise
3	Contrôle ou vérification de l'exactitude de la liste d'inventaire accompagnant les documents physiques (fichier Excel).
4	Préparation matérielle et intellectuelle des documents : retirer les agrafes, les trombones et les post-it (on ne les remet pas) ; pose des intercalaires entre lots des documents (feuille START et feuille END) pour déterminer un agrégat et des code-barres.
	Numérisation
5	Phase de tests de pré-scannage et envoi du résultat au client pour appréciation et validation
6	Scannage proprement dit
7	Traitement automatique d'images : détourage, redressement, amélioration du contraste, séparation des pages (pour les livres), etc.
	Post-numérisation
8	Structuration (agrégation) manuelle des pages pour constituer un lot, un document
9	Nommage des fichiers selon ce qui est indiqué dans l'inventaire
10	OCR par structure (un PDF par chapitre ou par parution)
11	Transfert d'images numérisées sur un disque dur[21] (archivage sur disque)
12	Restitution du disque et des documents physiques au client

[21] Avec possibilité que le client reste avec ce disque dur en l'achetant.

4.5.4. Entreprise D

Le workflow de numérisation de l'entreprise D (**tableau 18**) commence bien par discussion en amont avec le client pour déterminer ensemble les paramètres et les règles de numérisation. Lors de la préparation et de la phase de tests (étalonnage selon les spécifications de la Bibliothèque nationale suisse), l'entreprise utilise un chart de couleurs pour calibrer ses scanners.

Pour chaque projet, les documents à numériser sont toujours accompagnés d'un fichier de description et d'un index pour faciliter la saisie des métadonnées par l'entreprise. Mais quand il s'agit d'un grand projet, il revient au client de faire lui-même l'indexation, car l'entreprise n'a pas de compétence en la matière. Le schéma des métadonnées utilisé est le Dublin Core dont le minimum de métadonnées à saisir comprend le titre de la photo, la date de numérisation et le nom du logiciel utilisé.

Chaque fois que le document à numériser contient du texte, l'entreprise pratique l'OCR plein texte, en utilisant Adobe Acrobat comme logiciel. Le taux de reconnaissance de caractères visé est 100 % puisqu'il y a une relecture du fichier texte généré. Les difficultés rencontrées avec l'OCR sont liées aux vieux documents et à la présence de caractères spéciaux sur certaines lettres.

Les post-traitements peuvent être légers (élimination grossière des rayures et des taches, recadrage et redressement de l'image) ou moyens (corrections des couleurs, élimination grossière des rayures et des taches, recadrage et redressement de l'image y compris l'affûtage).

Tableau 18 : Étapes du workflow de numérisation de l'entreprise D

N°	Etapes
	Prénumérisation
1	Discussion en amont du projet avec le client pour établir les caractéristiques des livrables
2	Préparation manuelle et intellectuelle des documents, pose de la charte de couleur pour calibrage
3	Premier test basé sur le cahier des charges : étalonnage du matériel en enregistrant et en mesurant les cibles de test (pour vérifier la grandeur de l'image)
4	Deuxième test (pré scan) basé sur le cahier des charges (pour vérifier la couleur, la luminosité, le contraste…) et envoi du résultat au client pour validation
	Numérisation
5	Scannage proprement dit (numérisation fine)
6	Exportation d'images numérisées dans Adobe Photoshop Lightroom
	Post-numérisation
7	Contrôle qualité et traitements d'images numérisées (netteté, retouches, positionnement, redressement…)
8	Océrisation éventuelle (en cas de présence de texte sur l'image numérisée)
9	Indexation selon les consignes données par le client
10	Transfert d'images numérisées sur le serveur et backstop, puis dépôt en banque (coffre-fort)
11	Restitution des images numérisées au client sur disques durs externes ou clé USB
12	Restitution des documents physiques au client

4.5.5. Entreprise E

Toute la chaîne de numérisation de l'entreprise E (**tableau 19**) est vue en amont avec le client pour savoir comment il est organisé afin de pouvoir lui fournir quelque chose qui est exploitable. D'après le directeur de l'entreprise côté Suisse, une des difficultés qui sort souvent de ces discussions, c'est que le client ne sait pas exactement comment il va rechercher ses documents numériques :

> « *Généralement, le client dit toujours, moi je veux numériser. Et je dis oui mais comment vous recherchez ? Eh bien, je ne sais pas. Alors, on l'aide vraiment à faire intervenir soit ses collaborateurs finaux qui disent bon, moi je recherche comme ça dans mon armoire, c'est rangé comme ça, voilà. Il faut que ça colle un petit peu à l'organisation d'entreprise.* »[22]

En conséquence, l'entreprise se voit dans une situation de devoir diriger les échanges pour éclaircir la situation.

La préparation matérielle et intellectuelle des documents est faite en fonction de ce qui est attendu à la sortie de la numérisation, c'est-à-dire en fonction de ce que veut le client. La numérisation se fait par lots des documents correspondant à une certaine typologie. Le code-barres, est d'abord posé sur chaque carton lors de la prise en charge des archives chez le client, puis sur un séparateur des lots des documents pendant la phase de préparation. Il permet de séparer les documents, de garder et suivre le lien entre l'archive physique et son image numérisée. C'est une sorte d'inventaire comme le dit le directeur :

> « *Ça nous permet de lier, à vrai dire pendant la chaîne de numérisation, l'archive physique à un carton et à une image. Donc on sait que cette image-là elle se trouve sur ce serveur-là, mais elle se trouve également dans ce carton-là. Donc on a fait l'inventaire*. »[23]

Tableau 19 : Étapes du workflow de numérisation de l'entreprise E

N°	Etapes
	Prénumérisation
1	Discussion en amont du projet avec le client pour établir les caractéristiques des livrables
2	Prise en charge des archives chez le client : mise en cartons et pose des code-barres sur des cartons
3	Transport vers les locaux sécurisés de l'entreprise
4	Préparation matérielle et intellectuelle des documents : retirer les agrafes, trombones, post-it, les fourres, les reliures, les formats inappropriés, mettre un code-barres sur le séparateur de chaque lot de documents…
	Numérisation
5	Phase de tests et envoi du résultat au client pour appréciation et validation de la qualité de scannage, d'indexation et de format.
6	Scannage proprement dit et océrisation plein texte
7	Premier contrôle de la qualité d'images
	Post-numérisation
8	LAD (lecture automatique de documents)
9	Indexation manuelle (**par vidéocodage*** : mettre des métadonnées de recherche sur chaque image et par vidéotypage : donner une catégorie de document (facture, courrier, contrat…).
10	Deuxième et dernier contrôle de la qualité d'images
11	Transfert d'images numérisées vers une GED (ENOXA)
12	Restitution des documents numériques et documents physiques au client
13	Ou archivage numérique et/ou physique dans les locaux de l'entreprise.

[22] Entretien du 28 avril 2022.
[23] Entretien du 28 avril 2022.

Tout document scanné est océrisé (logiciel Kodak Capture Pro) en même temps que la numérisation. Le taux de qualité souhaité est de 90% et une relecture du fichier texte généré par l'OCR est effectuée.

L'indexation est à la fois automatique (LAD) et manuelle (vidéocodage). En effet, si le numéro de contrat est toujours au même endroit dans la feuille (dans la première page), ce qui facilite la reconnaissance automatique des documents pendant la phase de LAD et d'OCR, ce n'est pas le cas pour la date, qui elle, n'est pas forcément toujours au bon endroit. Ce changement demande de faire un complément avec de la saisie manuelle. Les critères d'indexation sont discutés en amont avec le client en fonction de la façon dont il prévoit effectuer des recherches de ses documents. Généralement, l'indexation pratiquée consiste à mettre un nombre limité de métadonnées. Par exemple, si le document concerné est un contrat bancaire, les métadonnées seront le numéro du contrat et la date à laquelle le contrat a été établi. La qualité d'extraction des informations figurant sur le document dépend de sa nature (structuré, semi-structuré, manuscrit) et de son état physique. Les métadonnées techniques sont par exemple la date de numérisation, le nom de celui ou celle qui a numérisé les documents, la résolution utilisée, le nombre d'images qui a été supprimé, le nombre de recto-verso, le nombre d'images dans le document, etc.

4.5.6. Entreprise F

Pour son workflow de numérisation (**tableau 20**), l'entreprise F n'a pas de préparation spéciale à faire sauf s'il manque un bulletin de versement sur la facture, car celui-ci est utilisé comme séparateur de lots. Pour les contrats, c'est la même chose, vu que c'est l'entreprise qui les génère. L'entreprise établit le contrat, elle l'imprime avec un code-barres qui contient le numéro du contrat. Le client le reçoit par la poste, il le signe, il le renvoie à l'entreprise qui le scanne. Une fois scanné, le contrat est automatiquement reconnu par le logiciel via son numéro et classé automatiquement.

La RAD (reconnaissance automatique des documents) consiste en une combinaison du modèle graphique (recherche des formes, des ancres…) et du modèle syntaxique (recherche par mots-clés, code-barres, libellés précis grâce à l'OCR plein texte).

Muni d'une intelligence artificielle (module externe qui n'a pas été développé par l'entreprise), le logiciel Youdoc est aussi capable de faire des captures intelligentes des documents non structurés. C'est-à-dire des captures qui ne nécessitent pas de faire un modèle[24] de document (ex. facture Swisscom, facture d'une commune…) pour faciliter la recherche et l'extraction de l'information.

> « *Lors de la numérisation, déclare le représentant de l'entreprise F et product owner, on lui dit simplement : je numérise une facture, donne-moi les informations de la facture, mais je ne veux pas faire un modèle*. »[25].

[24] D'après le représentant de cette entreprise, faire un modèle consiste à donner des instructions du genre : « ça c'est une facture de Swisscom, le numéro d'une facture Swisscom se trouve à tel endroit, le montant total se trouve à tel endroit, la date à tel endroit… » (entretien du 9 mai 2022).

[25] Entretien du 9 mai 2022.

Mais c'est surtout pour des documents un petit peu semi-structurés, précise-t-il. Par exemple, la capture intelligente n'a aucun intérêt pour les formulaires, car ce sont des documents structurés ; on sait que c'est à tel endroit où se trouve tel ou tel type d'information. Souvent, la capture par modèle et la capture intelligente sont combinées, car les modèles sont plus fiables et coutent moins chers. « Nous, on n'est vraiment souples parce qu'on peut faire les étapes qu'on veut dans l'ordre qu'on veut », conclut-il.

Avec la LAD, l'entreprise extrait les données suivantes : le numéro de contrat et le nom du fournisseur pour le contrat ; numéro de facture, date et montant pour la facture. Il y a toujours une étape manuelle qui permet de vérifier que la lecture est faite correctement et puis de valider ces données.

L'OCR ne fait pas partie des étapes principales du workflow de numérisation de l'entreprise F pour la simple raison que ses documents sont structurés ou semi-structurés. En cas de besoin (par exemple pour identifier un modèle ou constituer un fichier PDF texte pour la publication), l'OCR intervient au moment de la classification (RAD) et cible certaines zones. Il est aussi toujours fait sur les images noirs et blancs (sur couleur, l'OCR est beaucoup plus lent et moins fiable) avec un taux de reconnaissance de 90-95% (pas de relecture du fichier texte généré).

Tableau 20 : Étapes du workflow de numérisation de l'entreprise F

N°	Etapes
	Prénumérisation
1	Préparation physique des documents : insérer le bulletin de versement s'il est absent (sert de séparateur de lots de factures), poser un code-barres sur les contrats.
	Numérisation
2	Scannage proprement dit
3	Premier contrôle visuel de la qualité par un opérateur ou une opératrice : suppression de pages blanches, de publicité, contrôler l'ordre et l'orientation des pages.
	Post-numérisation
4	RAD (reconnaissance automatique de documents) : classification ou détermination automatique du modèle de document (facture, contrat)
5	Validation manuelle par un opérateur ou une opératrice qui vérifie si les modèles ont été bien reconnus et si le découpage de différents documents d'un lot a été bien effectué.
6	OCR en cas de besoin
7	LAD (lecture automatique de documents) : extraction et validation automatique des informations présentes sur les factures suivant leur modèle (en utilisant la reconnaissance de code-barres, de marques et de caractères).
8	Validation manuelle des données extraites
9	Publication des données extraites
10	Transfert des données vers une GED

4.5.7. Entreprise G

Le workflow de numérisation de l'entreprise G (**tableau 21**) commence aussi par discussion en amont avec le client pour établir les caractéristiques des prestations demandées. Après la préparation matérielle et intellectuelle des documents (étape qui prend beaucoup de temps

d'après le représentant régional de l'entreprise), l'entreprise G procède à la phase de tests pour vérifier la qualité des profils des scans. Pour la suite, les tâches centrales de la plateforme-capture de l'entreprise sont la classification (reconnaissance automatique des documents : facture, contrat…) et l'extraction ou lecture automatique de données se trouvant sur des documents reconnus. Effectuées grâce à l'intelligence artificielle (algorithmes d'apprentissage préalable avec des training sets et d'autoapprentissage constant), chacune de ces deux étapes est suivie d'une éventuelle étape de validation manuelle ; qui intervient lorsque tous les documents n'ont pas pu être automatiquement classés ou lorsque toutes les données n'ont pas pu être automatiquement extraites.

Tableau 21 : Étapes du workflow de numérisation de l'entreprise G

N°	Étapes
	Prénumérisation
1	Discussion en amont du projet avec le client pour établir les caractéristiques des livrables
2	Préparation manuelle et intellectuelle des documents : enlèvement des agrafes, trombones, post-it, pose des codes-barres sur chaque première page du document, etc.
	Numérisation
3	Phase de tests pour vérifier la qualité des profils des scans (qualité, compression, etc.) et envoi au client pour validation.
4	Scannage proprement dit
5	Premier contrôle qualitatif et quantitatif : netteté de l'image, complétude du dossier, suppression des pages blanches, redressement des images.
	Post-numérisation
6	Classification automatique des documents à l'aide de l'intelligence artificielle
7	Validation manuelle (éventuelle, en complément de la classification automatique)
8	Extraction automatique des informations métier pertinentes (LAD : lecture automatique de documents)
9	Validation manuelle des informations (éventuelle, en complément de l'extraction automatique)
10	Export vers un serveur de l'entreprise
11	Océrisation
12	Contrôle par pointage de la qualité d'OCR
13	Restitution des livrables au client (sur le serveur client via SFTP)

L'OCR plein texte intervient après numérisation et après que tous les traitements post-numérisation aient été effectués. Le taux de qualité souhaité dépend des besoins du client et il n'y a pas de relecture du fichier texte généré par l'OCR. En effet, même s'il est difficile de garantir à l'avance le taux de reconnaissance attendu, celui-ci est en général bon, vu que les logiciels OCR d'aujourd'hui intègrent des dictionnaires, souligne le représentant régional de l'entreprise.

4.6. Aspects qualitatifs et quantitatifs de la numérisation

Toutes les entreprises sondées procèdent à des contrôles qualitatifs et quantitatifs d'images numérisées, mais le nombre de contrôles est différent d'une entreprise à l'autre. Dans l'ensemble, chaque entreprise effectue au moins deux contrôles.

Chez l'entreprise A, le contrôle de la qualité d'images se fait en quatre niveaux, et il a pour but de vérifier, entre autres, la netteté des images numérisées, l'ordre, l'orientation et la complétude des pages. Le premier contrôle exhaustif (visuel et manuel) intervient pendant la numérisation, et il est fait par l'opérateur ou l'opératrice de numérisation. Le deuxième niveau de contrôle qui intervient toujours pendant la numérisation, est une autodétection automatique des pages avec zéros kilo-octets (0ko). Le troisième contrôle exhaustif (visuel et manuel) est effectué juste après numérisation par le responsable qualité. Le quatrième et dernier contrôle exhaustif et manuel intervient juste avant la restitution des livrables au client, et il est fait par le chef de projet.

L'entreprise B effectue deux contrôles qualité : le premier contrôle (vérification de la netteté d'images) intervient pendant le scannage et se fait par échantillonnage automatique (pointage). Le deuxième et dernier contrôle, le récolement, est un contrôle manuel et exhaustif qui intervient à la fin de la numérisation et qui vise à vérifier la complétude et la qualité des livrables finaux (vérifier si tout ce qui est décrit dans l'inventaire est bel et bien là physiquement : boîtes, fourres, etc.).

L'entreprise C effectue des contrôles qualitatifs et quantitatifs, exhaustifs et manuels à la fin de chaque étape du processus de numérisation, sauf pour l'océrisation où aucun contrôle n'est effectué (pas de relecture du fichier texte généré par l'OCR automatique, car c'est trop cher). Par exemple, après le scannage, l'entreprise fait un contrôle exhaustif et automatique par analyse de mire selon les standards FADGI, Metamorfoze et ISO. Le traitement de l'image par un contrôle exhaustif manuel inclut le détourage, le redressement, l'amélioration du contraste et la séparation des pages. Le même type de contrôle est effectué après structuration (segmentation) et vérifie la présence de toutes les pages d'une parution de journaux périodiques.

Chez l'entreprise D, le contrôle qualité intervient lors de la phase de test (deux tests de pré-scannage) et après le transfert d'images numérisées dans Photoshop Lightroom. Selon le responsable de l'entreprise, cette étape d'exportation des données et le contrôle-traitement (exhaustif et visuel) qui suit prennent beaucoup de temps, vu que l'entreprise peut faire entre 400 et 500 scans par jour (par une seule personne).

L'entreprise E effectue deux contrôles qualité : le premier, exhaustif et visuel, intervient pendant le scannage et le deuxième, qui est un échantillonnage manuel, se fait à la fin du scannage, c'est-à-dire après indexation.

Pour l'entreprise F, l'opérateur ou opératrice effectue trois contrôles visuels de la qualité des images numérisées. Le premier contrôle intervient pendant le scannage et vise à supprimer les pages blanches et la publicité, et à vérifier l'ordre et l'orientation des pages. Le deuxième et troisième contrôle, sont effectués après scannage et concernent le vidéocodage-classification (2ème contrôle) et le vidéocodage-reconnaissance (3ème contrôle).

Le workflow de numérisation de l'entreprise G comprend quatre niveaux de contrôles quantitatifs et qualitatifs. Le premier, le deuxième et le troisième niveau de contrôle sont à la fois exhaustifs, manuels et visuels alors que le quatrième et dernier contrôle se fait par échantillonnage. Le premier contrôle intervient tout juste après scannage d'un lot de documents et il a pour but de contrôler la netteté de l'image, la complétude du dossier, de

supprimer les pages blanches et de redresser les images. Le deuxième contrôle intervient après la classification des documents et il a pour but de vérifier si tous les documents ont été reconnus. Le troisième contrôle intervient après l'extraction des données et il a pour but de vérifier si toutes les données ont été extraites. Le quatrième et dernier contrôle intervient après océrisation et il a pour but de vérifier la qualité de l'OCR.

4.7. Les aspects juridiques et normatifs de la numérisation

4.7.1. Entreprise A

Selon son représentant, l'entreprise, du fait qu'elle n'est pas propriétaire des images qu'elle numérise, est sujette à peu de textes législatifs. Lorsqu'elle travaille avec des sociétés privées, l'entreprise A est tout simplement tenue de respecter les dispositions du droit contractuel en vigueur dans le pays où se déroule le projet de numérisation. Pour les établissements culturels patrimoniaux, en majorité publics ou parapublics, l'entreprise doit respecter ses engagements sur la base d'un Cahier des charges.

Concernant les normes, l'entreprise A applique les standards en vigueur dans la profession (ISO, AFNOR et autres normes internationales et nationales). Par exemple, quand l'entreprise numérise aux Etats-Unis d'Amérique, elle est tenue d'appliquer les normes FADGI (Directives techniques pour la numérisation des documents du patrimoine culturel). Le processus de certification pour la norme ISO 9001 (système de management de la qualité) est déjà en cours. L'Entreprise respecte aussi les dispositions des lois françaises et suisses en matière de droit d'auteur et de protection des données personnelles.

4.7.2. Entreprise B

Les aspects juridiques que l'entreprise B respecte concernent la conformité de l'entreprise aux lois fédérale et cantonales sur l'archivage (respect des délais de protection), la protection des données personnelles et le droit d'auteur et droits voisins.

D'après sa représentante, pour l'accès aux données personnelles,

> *« c'est aux clients de l'entreprise de respecter ces dispositions quand ils font de la diffusion et non l'entreprise quand elle fait de la numérisation. En revanche, l'entreprise suit ces prescriptions quand elle devient l'archiviste et quand elle veut mettre ces objets numériques à disposition dans les inventaires d'archives de ses clients*. »[26]

Numérisant les archives patrimoniales, l'entreprise B applique les directives de la norme ISO 14721 (modèle OAIS) concernant l'archivage pérenne. Elle applique aussi les directives techniques pour la numérisation des documents du patrimoine culturel, telles que FADGI, Metamorfoze et Memoriav. Le **tableau 22** donne les noms de quelques normes et de quelques textes de lois que l'entreprise B est tenu de respecter.

[26] Entretien du 16 mai 2022

Tableau 22 : Normes, lois et règlements encadrant les activités de numérisation et d'archivage de l'entreprise B

Normes ISO	**Domaine concerné**
ISO 21550	Photography-Electronic scanners for photographic images-Dynamic range measurements. Photographie-Scanners électroniques pour images photographiques-Mesures de plage dynamiques
ISO 12233	Photography-Electronic still-picture cameras-Resolution measurements. Photographie-Appareils photo électroniques-Mesures de résolution
ISO 14721 (OAIS)	Système d'archivage électronique (SAE) pérenne (docuteam cosmos)
ISO 19005 (divers)	PDF/A (supports des données)
Best Practices and Guidelines	
Metamorfoze (version 1.0-January 2012)	Metamorfoze Preservation Imaging Guidelines. Image quality
Directives FADGI (2016)	Technical Guidelines for Digitizing Cultural Heritage Materials/ Directives techniques pour la numérisation des documents du patrimoine culturel
Memoriav (2017)	Recommandations photo : Conservation des photographies
Lois et règlements	
Loi fédérale sur le droit d'auteur et droits voisins	Diffusion des données numériques : droit d'auteur
Lois cantonales sur l'archivage	Accès aux archives : délais de protection
Lois fédérale et cantonales sur la protection des données personnelles ou sur la transparence	Accès aux données personnelles

4.7.3. Entreprise C

Pour les aspects juridiques, l'entreprise C demande à ses clients de lui signer un document le déchargeant de toute responsabilité vis-à-vis des exigences juridiques liées au droit d'auteur et à tous les droits liés au document. Dans tous ses projets de numérisation, l'entreprise prend soin d'informer clairement tous ses clients que c'est à eux que revient la responsabilité si les documents à scanner sont encore soumis au droit d'auteur. Le client prend aussi la responsabilité quant à l'usage non autorisé des documents scannés par l'entreprise.

Concernant les normes, l'entreprise C n'est pas certifiée pour telle ou telle norme ISO. D'après son directeur, l'entreprise préfère travailler sur une base de réputation, qui s'appuie à son tour sur un travail bien fait, la confidentialité et la sécurité totale des documents numérisés. Par exemple, les images numérisées ne quittent pas les serveurs de l'entreprise, et elles sont supprimées aussitôt que le client signale qu'il les a bien reçues.

4.7.4. Entreprise D

Lorsqu'elle travaille pour une Archive ou une bibliothèque, l'entreprise C est tenue de respecter les dispositions légales et réglementaires telles qu'elles sont stipulées dans le contrat de numérisation. Ces dispositions sont celles qui figurent dans le document ″Ligne de conduite pour la numérisation″ publié par la Bibliothèque nationale suisse (BN). Par exemple, en matière de droit, il est mentionné que :

> « *La BN numérise ses collections dans le respect du droit d'auteur et du droit de la personnalité. Elle numérise et met en ligne sans contrôle préalable les documents dont la publication remonte à plus de 110 ans et plus. Cette règle s'applique au dernier numéro paru pour les publications périodiques. Elle contrôle systématiquement la date de décès des auteurs des ouvrages parus plus récemment et complète les notices d'autorité de ses catalogues avec cette information.* »
>
> (Bibliothèque nationale suisse, 2017, p.7)

.

Il en est de même pour les normes de qualité, comme la norme ISO 19264-1 relative à la qualité du matériel numérisé et l'ISO 2895-1 concernant les procédures d'échantillonnage pour le contrôle par attributs[27]. Également les normes ISO 12647-2 à 12647-8 concernant les différents procédés d'impressions professionnels pour les imprimeries ainsi que les directives techniques FADGI sont appliquées par l'entreprise D.

Lorsque l'entreprise D numérise pour les entreprises et personnes privées, elle applique les mêmes règles et les mêmes normes, car comme le dit son propriétaire, la Bibliothèque nationale suisse (BN) lui sert surtout de « benchmark pour la qualité. » Enfin, la Bibliothèque nationale et d'autres institutions patrimoniales exigent à leurs fournisseurs de services de numérisation de contracter une assurance pour garantir les éventuels dégâts matériels.

4.7.5. Entreprise E

L'entreprise E respecte les dispositions des lois fédérales et cantonales sur la protection des données, le droit d'auteur et droits voisins et l'archivage (délais de protection des données). Il en est de même pour les législations européennes, principalement le Règlement général sur la protection des données (RGPD) qui oblige l'entreprise à anonymiser les noms de ses clients lors de la numérisation (éviter de saisir les métadonnées liées à l'individu). Mais aussi les législations françaises, par exemple sur la valeur probante des copies numériques. Pour ce dernier point, l'entreprise E applique les normes françaises en Suisse (mais l'inverse n'est pas vrai), car selon son directeur, les règles suisses sont moins exigeantes en la matière.

Concernant les normes, l'entreprise E est certifiée conforme à la norme ISO 9001 :2015 (système de management de la qualité), à la norme ISO 14641 (système d'archivage électronique probant), à la norme AFNOR NF Z42-020 (coffre-fort électronique) et à la norme AFNOR NF-Z42-026 régissant la numérisation fidèle (copie fidèle). Pour rappel, la

[27] Extrait des spécifications techniques (en allemand) reçu de Liliane Regamey en mai 2022 puis traduit en français par Google. Titre du document : Exemple anonyme de ce que la Bibliothèque nationale suisse demande à ses fournisseurs de numérisation pour le type de document « journal » (7p.).

numérisation fidèle ajoute à la chaîne de numérisation classique les quatre critères suivants (Locarchives, 2022) :

- Qualification et paramétrage de la chaîne de numérisation probante ;
- Justification de la fiabilité de la numérisation probante ;
- Mise en place d'un système complet de traçabilité sur l'ensemble des opérations de numérisation probante ;
- Mise en œuvre des moyens cryptographiques afin d'assurer l'intégrité des copies numériques (empreinte, horodatage et cachet électronique).

4.7.6. Entreprise F

Côté législation, l'entreprise F respecte les dispositions du Code de procédure civil (CPC) suisse concernant la force probante des documents électroniques, celles de la Loi fédérale sur la signature électronique et celles du Règlement général sur la protection des données (RGPD) de l'Union européenne (traitement des données personnelles). Pour la signature électronique, l'entreprise F utilise les trois types de signature électronique reconnus en Suisse : signature électronique simple, signature électronique avancée et signature électronique sûre ou qualifiée. Le document est signé à l'aide d'un certificat numérique installé sur le serveur, puis envoyé vers un serveur de signature. Si le document doit être signé manuellement, cette signature sera gérée dans un flux de validation géré par la GED de l'entreprise.

Concernant les normes, l'application Youdoc de l'entreprise est certifiée pour la copie fiable (NF-Z42-026) et pour le système d'archivage électronique probant (ISO 14641) :

> « *… et on a un suivi de tout ce qui est modifié ; dès qu'il y a une modification, c'est stocké dans un log avec du blockchain, on ne peut pas supprimer l'enregistrement dans les logs il y a toute cette partie-là* »[28].

L'entreprise est aussi certifiée pour son système de management de la qualité (ISO 9001) ainsi que pour la sécurité de son système d'information (ISO 27000).

4.7.7. Entreprise G

Les solutions d'archivage proposées par l'entreprise sont conformes à la législation et aux réglementations suisses ou internationales régissant l'archivage probant. Pour le cas suisse, il s'agit des dispositions relatives aux textes suivants :

- L'Ordonnance OLICO pour la réglementation suisse sur l'archivage et la gestion des documents des entreprises (force probante des documents électronique) ;
- Le Code des obligations (CO), art. 957ss concernant les dispositions relatives à la comptabilité commerciale et à la présentation des comptes ;
- Lois fédérales et cantonales sur l'archivage (délais de protection des données) ;
- Lois fédérales et cantonales sur la protection des données personnelles ou sur la transparence.

[28] Représentant de l'entreprise F, entretien du 9 mai 2022.

Les normes pour lesquelles l'entreprise G est certifiée sont plus liées à la partie ECM/GED qu'à la numérisation. C'est ainsi que l'entreprise est certifiée pour la norme ISO 270001 concernant la sécurité de l'information.

4.8. Méthodes de tarification des prestations de numérisation

Dans cette section, nous exposons seulement les méthodes de facturation des prestations de numérisation et de post-numérisation telles que pratiquées par les entreprises sondées. Quelques entreprises ont accepté de nous communiquer des prix indicatifs, mais pour des raisons liées à la concurrence et pour respecter les consignes qui nous ont été données par ces entreprises, ces prix ne serons pas mis au public (ne sont pas publiés dans ce document, mais dans une annexe à part).

4.8.1. Entreprise A

Pour la numérisation, la facturation se fait par page et les formats A4 et A3 ont le même prix toutes couleurs confondues. Pour l'OCR brut et l'indexation manuelle, la facturation est faite par champ. Les prestations archivistiques, quant à elles, sont facturées à la journée ou au mois de travail. D'après le représentant de l'entreprise, le fait de réaliser des prestations post-numérisation (OCR, contrôle qualité exhaustif, segmentation manuelle, indexation, réalisation de la table des matières…) entraînent des coûts supplémentaires.

4.8.2. Entreprise B

Le bureau régional (Suisse romande) de l'entreprise B et le siège (Argovie) ne facturent pas les prestations de la même façon. Le bureau régional facture à l'heure toutes ses prestations, y compris même celles liées à la numérisation car celle-ci est toujours incluse dans un grand projet de traitement archivistique. C'est seulement au siège de l'entreprise B que la facturation peut être à la fois à l'heure et à la page ; et les formats A4 et A3 ont le même prix, toutes couleurs confondues.

4.8.3. Entreprise C

L'entreprise C facture ses prestations de numérisation à la page et au document ou agrégat, « parce que faire un agrégat ça coûte, (…) il doit passer dans un workflow spécial », souligne son directeur. Par exemple, un client qui veut numériser un livre de 500 pages en le divisant en deux parties, sera facturé comme suit : (500 pages x prix à la page) + 2 fois le prix à l'agrégat. Le prix à la page dépend de la difficulté liée à sa préparation (temps mis), variable selon qu'il s'agit de pages libres, pages reliées, pages pliées. Pour les grands formats (A0, A1), le prix est forfaitaire, car il faut déplier, mettre le plan à plat avec une table aspirante, le replier… Pour les couleurs, le gris est de l'ordre de 15 à 20% moins cher que la couleur.

4.8.4. Entreprise D

L'entreprise D facture ses clients par photo, par page (plan, dessin, carte…) ou au forfait (**tableaux 23, 24, 25 et 26**). Pour chaque commande, l'entreprise applique aussi un prix

forfaitaire de CHF 30 pour le traitement des données (sans support de données). Le montant minimum de commande est de CHF 50.-. Ces prix sont disponibles en allemand sur le site Internet de l'entreprise.

Tableau 23 : Prix de numérisation des diapositives et négatifs appliqués par l'entreprise D

Nature du document et format	Taille imprimable	Prix CHF	Services supplémentaires	Détails du fichier
35 mm secs	A4 (300 dpi)	1.25	Nettoyage grossier à l'air comprimé	Données 16 bits, TIFF et JPEG fournies
	A3 (300 dpi)	1.80		
		0.20	Nettoyage avec un nettoyant pour film	
Restauration de lame de verre	A4 (300 dpi)	6.00	Analyse de sécurité, retirer la vitre, nettoyage, analyse principale, nouveau cadre	Données 16 bits, TIFF et JPEG fournies
	A3 (300 dpi)	6.50		
Moyen format, secs	A4 (300 dpi)	2.50	Nettoyage grossier à l'air comprimé	Données 16 bits, TIFF et JPEG fournies
	A3 (300 dpi)	3.00		
		0.20	Nettoyage avec un nettoyant pour film	
Restauration de lame de verre	A4 (300 dpi)	8.00	Analyse de sécurité, retirer la vitre, nettoyage, analyse principale, nouveau cadre	Données 16 bits, TIFF et JPEG fournies
	A3 (300 dpi)	8.50		

(Document fourni par l'entreprise, 2022)

Tableau 24 : Prix de numérisation des plaques de verre et photographies appliqués par l'entreprise D

Nature du document et format	Taille imprimable	Prix CHF	Services supplémentaires	Détails du fichier
Plaques de verre				
Grand format 4x5 pouces jusqu'à 114x140mm à plat, sèches	A4 (300 dpi)	3.50	Nettoyage grossier à l'air comprimé	Données 16 bits, TIFF et JPEG fournies
Grand format 8x10 pouces à partir de 114x140mm à plat, sèches	A3 (300 dpi)	6.00	Nettoyage grossier à l'air comprimé	Données 16 bits, TIFF et JPEG fournies
Photographies				
bis format 10x15 cm	A4 (300 dpi)	2.00	Nettoyage grossier à l'air comprimé	Données 16 bits, TIFF et JPEG fournies
bis format 10x15 cm	A3 (300 dpi)	3.30		
ab format 10x15 cm bis A4	A4 (300 dpi)	2.50		
ab format 10x15 cm bis A4	A3 (300 dpi)	3.50		

(Document fourni par l'entreprise, 2022)

Tableau 25 : Prix de numérisation des documents iconographiques appliqués par l'entreprise D

Nature et dimensions du document	Taille imprimable	Prix (CHF)	Des services supplémentaires	Détails du fichier
Diapositives et négatifs de 24 x 36 mm à 4 x 5 pouces	30 x 40 cm	39.00	Nettoyage des rayures fines supprimées, couleur optimisée.	Résolution d'impression : 300 dpi, 16 bits, données au format TIFF non compressé.
	50 x 70 cm 60 x 60 cm	59.00		
	70 x 100 cm 80 x 80 cm	79.00		
	100 x 130 cm 100 x 100 cm	129.00		
Diapositives et négatifs de 4 x 5 pouces	50 x 70 cm	79.00	Nettoyage des rayures fines supprimées, couleur optimisée.	Résolution d'impression : 300 dpi, 16 bits, données au format TIFF non compressé.
	70 x 100 cm	99.00		
	100 x 130 cm	159.00		

(Document fourni par l'entreprise, 2022)

Tableau 26 : Prix appliqués par l'entreprise D pour les post-traitements d'images

Post-traitement	Description	Prix
Post-traitement léger	Élimination grossière des rayures et des taches, recadrage et redressement de l'image.	CHF 2.- par photo
Post-traitement moyen	Corrections des couleurs, élimination grossière des rayures et des taches, recadrage et redressement de l'image. Y compris l'affûtage	CHF 5.- par photo
Post-traitement, restauration chronophage	Corrections de couleur chronophages, élimination des rayures et des taches. Recadrez et alignez l'image, éventuellement à l'aide d'un masquage. Y compris l'affûtage.	CHF 90 par heure (facturation à la minute).
Lames de verre : restauration avec numérisation de sauvegarde et nouveau cadre de diapositive	Les vieilles lames de verre sont souvent très sales à l'intérieur. Avant de retirer soigneusement la vitre, nous créons une analyse de sauvegarde. Un second scan a alors lieu à partir de la lame nettoyée. La diapositive est alors recadrée. Post-traitement selon les tarifs ci-dessus. Bien sûr, vous pouvez également ouvrir vous-même les lames de verre et nous envoyer simplement les lames sans cadre. Cela réduit le prix de CHF 3.-	Tarifs voir ci-dessus
Supports de données et accessoires	DVD par morceau	CHF 5.-
	Clé USB 8Go	CHF 12.-
	Clé USB 16Go	CHF 15.-
	Clé USB 32Go	CHF 20.-
	Clé USB 64Go	CHF 25.-
	Cadre coulissant par pièce	CHF 1.-
	Pochette pergamine (qualité musée) pour petits films négatifs par feuille (7 x 6 négatifs)	CHF 1.-
	Pochette pergamine (qualité musée) pour négatifs moyen format par feuille (4 x 3 négatifs)	CHF 1.-
	Pochette pergamine (qualité musée) pour grand format par pochette pour une photo	CHF 1.50

(Document fourni par l'entreprise, 2022)

4.8.5. Entreprise E

L'entreprise E facture la mise en conteneurs des boites d'archives chez le client et les frais de transport des documents vers ses locaux.

Les prestations de numérisation (préparation des documents et scannage) sont facturées à la page et celles d'indexation au document. Les formats A4 et A3 ont le même prix (toutes les couleurs confondues) qui dépend surtout de la complexité du document original à numériser : son état physique (document froissé ou pas), document avec divers formats, avec beaucoup d'agrafes, etc. Car plus c'est compliqué, plus l'étape de préparation des documents prend beaucoup de temps.

4.8.6. Entreprise F

L'entreprise F n'est pas un prestataire de numérisation tarifée. Elle numérise ses propres documents.

4.8.7. Entreprise G

Les prestations de numérisation de l'entreprise G sont facturées à la page et au dossier. Le prix dépend de l'état du document, c'est-à-dire la durée de sa préparation, le reste du projet de numérisation est facturé à l'heure ou à la journée.

4.9. Difficultés rencontrées par les entreprises prestataires de numérisation interviewées

De manière générale, les entreprises prestataires de numérisation ayant participé à cette étude ne rencontrent pas beaucoup de difficultés, car sur le plan technique, « les projets de numérisation, c'est quelque chose qui est relativement bien maîtrisé », colle le souligne le représentant régional de l'entreprise G. Selon lui, c'est surtout la concurrence qui fait qu'il y ait beaucoup d'acteurs qui travaillent différemment, ce qui complique les choix du client : « Parfois, celui-ci ne comprend pas pourquoi tel ou tel prestataire de numérisation est 2 ou 3 fois plus cher que l'autre alors qu'ils font la même chose », déclare-t-il. Une autre difficulté relevée par le représentant de l'entreprise G concerne la gestion de l'après-numérisation par le client : « les clients disent, oui on veut numériser, mais en fait ils n'ont pas forcément l'idée de ce qu'ils vont en faire derrière. »[29]

Le directeur et représentant de l'entreprise E évoque lui aussi la même difficulté observée du côté client : selon lui, le client ne voit pas clairement comment il veut retrouver ses données numériques une fois archivées et ce qu'il va faire avec les originaux (80% des clients de l'entreprise conservent les originaux).

Et d'ajouter :

> « *La principale difficulté est d'un côté de conseiller le client mais également de l'accompagner dans l'organisation de son entreprise, car cela a un impact dans la*

[29] Entretien du 1er juillet 2022.

manière de travailler et les modes opératoires du flux d'information de l'entreprise en général. »[30]

Toujours du côté client, le directeur et représentant de l'entreprise C note ceci : « souvent, il demande ça et après l'avoir interrogé, on découvre qu'il veut beaucoup plus. »[31] La grande variabilité des documents d'archives est une autre difficulté évoquée par le directeur et représentant de l'entreprise C dans la mesure où leur manutention prend beaucoup de temps, ce qui se répercute sur le coût total de numérisation.

Le représentant de l'entreprise A constate, quant à lui, que « les clients ne connaissent pas forcément la volumétrie exacte de leurs fonds, ne se doutent pas toujours du coût de la numérisation et leurs documents ne sont pas préparés » (pour la numérisation)[32].

Du côté de l'entreprise B, ce sont plutôt des difficultés d'ordre technique, notamment celles liées au volume et aux formats des documents : « nous renonçons de numériser de gros volumes de documents car notre équipement n'est pas concurrentiel, et nous ne sommes pas équipés pour de très grands formats », déclare sa représentante[33] Une autre difficulté qu'elle évoque est celle de pouvoir intégrer le travail de numérisation et de description dans les projets de l'entreprise : conseil et gestion de l'information, traitement et gestion d'archives, archivage électronique.

[30] Courrier électronique du 8 août 2022.
[31] Entretien du 27 mai 2022.
[32] Courrier électronique du 16 août 2022
[33] Courrier électronique du 3 août 2022.

5. Proposition d'un portefeuille de services de numérisation possibles pour ArchiLab

5.1. Synthèse des prestations de numérisation proposées par les entreprises étudiées

Pour pouvoir proposer un workflow complet de numérisation d'archives papier pour ArchiLab, nous devons d'abord faire une synthèse de tous les workflows proposés par les entreprises interviewées tels qu'exposés dans les pages précédentes. Le **tableau 27** donne une synthèse des étapes d'un workflow de numérisation patrimoniale alors que le **tableau 28** illustre celui de la numérisation documentaire classique ou de production.

Tableau 27 : Synthèse des étapes des workflows de numérisation patrimoniale

N°	Étapes	Entreprise A	Entreprise B	Entreprise C	Entreprise D
	Prénumérisation				
1	Discussion en amont du projet avec le client pour établir les caractéristiques des livrables	X	X	X	X
2	Transport des archives vers les locaux du prestataire	X*	X	X	X
3	Réception des documents physiques	X	X	X	X
4a	Vérification de l'exactitude du fichier de récolement accompagnant les documents physiques	X	X	X	X
4b	Inventaire et reconditionnement	X**	X	-	-
5	Préparation matérielle et intellectuelle des documents	X	X	X	X
	Numérisation				
6	Phase de tests de pré-scannage et envoi au client pour appréciation et validation	X	X	X	X
7	Scannage proprement dit	X	X	X	X
8	Traitement automatique des images numérisées : détourage, redressement des pages, amélioration du contraste, séparer les pages (pour les livres)	X	X	X	X
9	Nommage des fichiers	X	X	X	X
10	Contrôle qualité	X	X	X	X
	Post-numérisation				
11	Typage (structuration, agrégation) manuelle d'images numérisées	X	X	X	X
12	OCR	X	X***	X	X****
13	Indexation automatique	-	-	X	-
14	Indexation manuelle	X	X	X	X
15	Contrôle qualité	X	X	X	X
16	Récolement	X	X	-	-
17	Transfert vers les serveurs dans une GED de l'entreprise	X	X	X	X
18	Restitution des données numériques et documents physiques au client	X	X	X	X

* Pas de façon systématique (l'entreprise privilégie la numérisation sur le site du client). **Sur demande (normalement, l'entreprise reçoit les collections avec un inventaire). ***Pas de façon systématique.
****Se fait de manière occasionnelle (quand il y a du texte sur le document).

Tableau 28 : Synthèse des étapes des workflows de numérisation mixte et de production (numérisation documentaire classique)

N°	Étapes	Entreprise E	Entreprise F	Entreprise G
	Prénumérisation			
1	Discussion en amont du projet avec le client pour établir les caractéristiques des livrables	X	-	X
2	Prendre des archives chez le client : mise en cartons et pose des code-barres sur des cartons	X	-	X
3	Transport des archives vers les locaux du prestataire	X	-	X
4	Inventaire sommaire chez le client et reconditionnement	X	-	-
5	Préparation matérielle et intellectuelle des documents	X	X	X
	Numérisation			
6	Phase de tests de pré-scannage et envoi au client pour appréciation et validation	X	-	X
7	Scannage proprement dit	-	X	X
8	Scannage proprement dit + OCR	X	-	
9	Contrôle de la qualité des images numérisées	X	X	X
10	Nommage des fichiers	X	X	X
	Post-numérisation			
11	RAD (Reconnaissance automatique des documents) ou typage	-	X	X
12	LAD (Lecture automatique des documents)	X	X	X
13	OCR	-	X	X
14	Indexation manuelle : par vidéocodage (mettre des métadonnées de recherche sur chaque image) et par vidéotypage (donner une catégorie de document, comme facture, courrier, contrat…)	X	X	X
15	Contrôle qualité	X	X	X
16	Transfert vers un serveur dans une GED	X	X	X
17	Restitution des données numériques au client	X	-	X
18	Restitution des documents physiques au client	X	-	X
19	Archivage numérique par l'entreprise	X	X	X
20	Archivage physique par l'entreprise	X	X	-

A partir de cette synthèse, nous avons élaboré trois workflows complets tenant compte de la typologie des documents numérisés. Le premier est un workflow de numérisation patrimoniale (fonds d'archives papier) qui est notre workflow principal que nous proposons pour ArchiLab (**figure 10**). Le deuxième workflow (**annexe 3**) concerne les documents iconographiques car, les archives étant par nature des documents hétérogènes, il faut s'attendre à ce que l'on trouve ce type de document dans les fonds d'archives qui seront proposés à ArchiLab. Le troisième workflow est un workflow mixte (**annexe 4**), c'est-à-dire incluant la numérisation documentaire classique (ou de production) et la numérisation patrimoniale (cas typique de l'entreprise E). Ce workflow mixte inclut aussi la numérisation de production (ou documentaire classique) représentée ici par les entreprises F et G.

Pour chacun de ces workflows et au vu des difficultés déclarées par les entreprises, nous insistons sur l'importance de la première étape qui se situe tout juste en amont d'un projet de numérisation, et qui consiste à discuter à fond avec le client pour mieux cerner ses besoins (ses attentes) en matière de numérisation.

Toujours au vu des pratiques de numérisation en vigueur au sein des entreprises interviewées, nous insistons également sur la place que doit prendre **l'inventaire des fonds d'archives** dans les projets de mandat qui seront menés au sein du laboratoire ArchiLab. Nous reviendrons sur ce point au chapitre des recommandations. En effet, nous avons

constaté que l'inventaire comme un des services d'un projet de numérisation est peu proposé par les entreprises ayant participé à cette étude. Quatre cas de figure se présentent : il y a d'abord l'entreprise B pour qui l'inventaire est l'activité principale, puis l'entreprise A qui fait de l'inventaire occasionnellement (sur demande du client), l'entreprise E qui fait de l'inventaire que nous avons qualifié de sommaire (il se limite au carton). Et enfin les entreprise C, D et G qui ne pratiquent pas du tout d'inventaire, soit par manque de compétence ou pour des raisons stratégiques.

5.2. Proposition d'un workflow complet de numérisation d'archives papier pour ArchiLab

La **figure 10** illustre les différents services d'un workflow de numérisation d'archives papier proposé pour ArchiLab. Les rectangles en vert représentent les produits des étapes correspondantes. Selon la nature du projet et les besoins du client, les étapes 12, 13, 14 et 15 peuvent ne pas être demandées ; dans ce cas, les images numérisées sont directement sauvegardées. Mais cela ne veut pas dire que le laboratoire ArchiLab ne doit pas les proposer. Tant qu'il y a toujours de l'imprimé sur un document, l'OCR est toujours demandé. D'ailleurs, outre l'inventaire, ArchiLab devrait proposer des services d'indexation, toujours dans cette perspective de dynamisation des cours. Pour la majorité d'entreprises ayant participé à cette étude, l'indexation complète est faite par le client (établissement ou institution propriétaire des documents à numériser). La raison en est qu'il s'agit d'une prestation très professionnelle pour laquelle les entreprises manquent de compétences.

Étape 1 : Discussion en amont avec le client pour définir les caractéristiques des livrables

C'est une étape incontournable car elle permet de bien cerner les tenants et les aboutissants d'un projet de numérisation porté par le client. C'est au cours de cette étape que toutes les prestations et les règles de numérisation et de contrôle qualité sont clairement définies entre le client (commanditaire) et le prestataire de numérisation. Car selon les propos rapportés par la majorité des représentants d'entreprises, c'est au cours de cette étape que le prestataire et le client doivent se rendre compte s'ils ont la même vision des contours du projet de numérisation (le coût, la préparation des documents, la gestion des données numériques après numérisation, la recherche ou accès aux données numériques, le sort des originaux…).

Étape 2 : Prise en charge des documents : transport et stockage

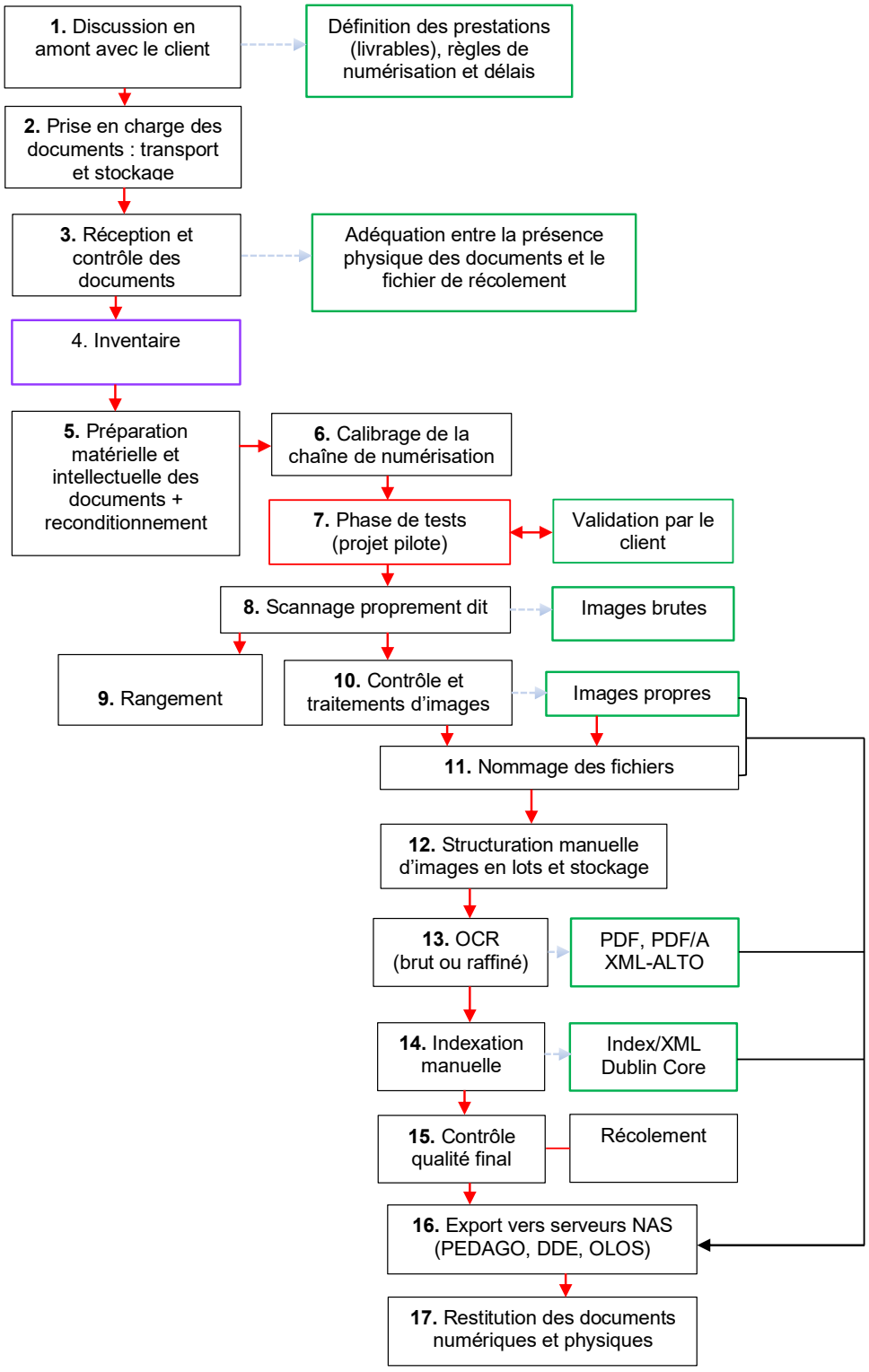
Pour cette étape, on peut se poser plusieurs questions : A qui revient le transport des documents pour l'aller et le retour ? A quel coût ? Comment facturer cette tâche ? Comment sécuriser les documents pendant le transport ? Quel est le délai d'indisponibilité des documents ? …). Des réponses à ces questions et tant d'autres doivent être claires. Selon les pratiques en vigueur au sein des entreprises sondées, le transport d'archives à numériser est généralement assuré par le prestataire.

Étape 3 : Réception et contrôle des documents

Dès réception des documents, le prestataire doit procéder à un **récolement des documents** figurant à l'inventaire (si celui-ci a été déjà fait par le client). Le prestataire doit aussi vérifier

la **conformité du lot avec le bordereau** qui accompagne la livraison. Ce contrôle est pertinent, car certaines entreprises ont évoqué la présence d'irrégularités lorsqu'elles reçoivent les documents à numériser : documents figurant dans le fichier de récolement mais absents physiquement et vice versant.

Figure 10 : Workflow de numérisation d'archives papier proposé pour ArchiLab



Étape 4 : Inventaire

C'est une prestation que nous proposons et qui cadre parfaitement avec la finalité du laboratoire ArchiLab consistant à dynamiser les cours archivistiques. Plus précisément, vu que la Haute école de gestion forme des professionnels des archives, les projets de numérisation menés au sein d'ArchiLab dans le cadre de la coopération avec d'autres institutions, devraient être une occasion pour les étudiant-e-s de mettre en pratique les cours théoriques portant sur le traitement et la gestion d'archives.

Étape 5. Préparation matérielle et intellectuelle des documents, et reconditionnement

Toutes les entreprises interviewées déclarent que la préparation matérielle et intellectuelle des documents est l'étape la plus longue de la chaîne de numérisation et qui, par conséquent, influence beaucoup le coût d'un projet de numérisation des documents. Ce que confirme aussi la Direction générale des systèmes d'information (DGSI) du canton de Genève qui précise que cette étape peut prendre « au moins 50 % du temps total de traitement. » (DGSI, 2015, pp. 10-11). Selon la même source, l'importance de cette étape tient en ce qu'elle « conditionne, dans une large mesure, la qualité de la capture et la productivité de la chaîne. » (p.10). Les étapes 4 et 5 du workflow proposé doivent être effectuées de façon simultanée. Le reconditionnement consiste à mettre les documents déjà préparés dans des fourres non acides et dans des boites afin de les rendre disponibles pour le scannage.

Tous les documents à numériser doivent être évalués pour leur état (physique et chimique) et des mesures appropriées doivent être prises si des problèmes sont découverts. Ces éventuelles mesures doivent être consignées dans le contrat).

Etape 6. Calibrage de la chaîne de numérisation

La calibration se fait à partir d'une mire colorimétrique dont les couleurs sont connues. Comme vu précédemment, le directeur et représentant de l'entreprise C conseille d'utiliser la charte ColorChecker Xrite. Ce point de vue est aussi partagé par la Bibliothèque et Archives nationales du Québec, la Bibliothèque nationale de France et le Musée canadien de l'histoire (BANQ, BnF & MCH, 2014, pp. 12-13) qui recommandent aussi, au même titre, la charte ColorChecker Passport. D'après ces institutions, ces deux chartes sont à privilégier « principalement parce que les valeurs numériques connues de ces chartes sont valables et constantes. » Le calibrage de la chaîne de numérisation consiste à numériser une échelle de gris ou une charte de couleurs dans les mêmes conditions que les documents à numériser, de façon à obtenir une image qui « servira d'image de référence pour la calibration de l'image numérisée. » (BANQ, BnF & MCH, 2014, pp. 12-13).

> « *L'ensemble de la chaîne de production doit faire l'objet de calibrages à l'aide de sondes (scanners et écrans) utilisant des mires normalisées (Colorchecker et mire de résolution et quadrillage noir et blanc) : point blanc, point noir, gamma (luminosité), température de couleurs. L'utilisation de profils ICC (fichiers décrivant la manière dont un matériel informatique restitue les couleurs) permet d'assurer un suivi rigoureux du calibrage tout au long de la prestation.* »
>
> (Grassin, 2008, p.2)

Étape 7 : Phase de tests

Cette phase consiste à tester le processus de numérisation dans son ensemble : les outils, la résolution et les formats, les méthodes de contrôle, l'ajout des métadonnées, et les moyens de sécurité engagés par le prestataire. L'échantillon sur lequel s'effectue les tests doit être représentatif. La phase s'achève par un accord entre prestataire et le client et « la rédaction d'un cahier de procédures (ou plan assurance qualité) décrivant l'ensemble de points de la prestation et incluant les chartes techniques. » (BnF et BMCO, 2010, p.21).

Après l'installation de l'équipement et la familiarisation avec le matériel et les logiciels, les directives techniques FADGI recommandent d'effectuer une évaluation initiale des capacités de performance à l'aide du système DICE afin d'établir une référence pour chaque appareil de numérisation spécifique. Les tests DICE doivent être effectués au début de chaque journée de travail ou au début de chaque lot, selon la première éventualité. Des tests au début et à la fin de chaque lot pour confirmer que la numérisation était cohérente pour l'ensemble du lot sont souhaitables (Rieger, 2016, p.71).

Étape 8 : Scannage proprement dit et nommage des fichiers

Cette étape commence après validation par le client de l'ensemble de tests effectués à l'étape 7. L'étape aboutit à des images numériques brutes et au nommage des fichiers. Un schéma de nommage des fichiers doit être établi avant la capture des données et celui-ci doit être consigné dans le contrat.

Etape 9. Rangement

Il consiste à placer les documents dans des armoires préparer à cette fin.

Étape 10. Contrôle qualité et traitements d'images numérisées

Les contrôles doivent être à la fois qualitatifs et quantitatifs. Ils peuvent se faire sur l'ensemble des documents numérisés (contrôle exhaustif) ou sur un échantillon représentatif. Ces contrôles peuvent aussi être effectués automatiquement (par le logiciel Kofax Express) et manuellement (par un opérateur ou une opératrice). Kofax Express est en effet équipé du dispositif VRS d'amélioration d'images qui permet d'automatiser les contrôles qualité suivants : suppression des pages blanches, redressement des pages, amélioration du contraste et de la luminosité, détection et conservation de la couleur, affinement des écritures et des caractères pour une meilleure océrisation, nettoyage du bruit, des fonds de page et des perforations sur les documents. La fonction Correction VRS de Kofax permet aussi de mettre de côté une image qui sera inspectée et corrigée ultérieurement par un opérateur ou une opératrice sans renumériser le document (Kofax, 2019, p.3).

Il est important de rappeler ici le point de vue du Bureau de coopération interuniversitaire (BCI) au Canada qui conseille d'effectuer **un contrôle exhaustif** des documents numérisés en cas de **numérisation de substitution** et un **contrôle par échantillonnage** en cas de **numérisation de diffusion** (BCI, 2014, p.20).

L'étape de contrôle et traitements produit des images numériques propres à partir desquelles des prestations de conversion en mode texte (océrisation) peuvent être effectuées. Si un contrôle par échantillonnage est prévu dans le cahier des charges, celui-ci doit préciser s'il

doit se faire selon les exigences prévues par la norme ISO 2859-1. A l'image de la **figure 11** ci-après, nous proposons, pour ce workflow de numérisation, d'effectuer un contrôle à la fin de chaque étape du processus de numérisation.

Étape 11. Structuration

La structuration ou agrégation manuelle consiste à mettre ensemble les images numérisées appartenant à un même lot de documents.

Étape 12. OCR

Ici, le client devra préciser au prestataire de quel type d'OCR s'agit-il (OCR brut ou OCR raffiné) et le taux de reconnaissance souhaité. Dans tous les cas, le fonctionnement de l'algorithme est le même (Perrot, 2016, p.64). :

- Redressement de l'image de la page pour aligner les lignes sur la grille d'expertise ;
- Décomposition de la page en élément contenant du texte est des images ;
- Bitonalisation du texte pour la reconnaissance ;
- Établissement de la correspondance entre les pixels de chacun des caractères avec une matrice décrivant le caractère type ;
- Comparaison des mots formés avec un dictionnaire, ce qui permet d'intégrer certains "suspects" pour reformer un mot.

Quant à la gestion du fichier texte généré, au lieu d'exporter le texte vers une application indépendante, par exemple ABBYY fine reader, Perrot souligne que la solution qui intéresse l'archiviste, consiste à mettre le texte « en couche sous-jacente à l'image dans le fichier PDF représentant l'archive. » Cette solution est aussi celle prônée par l'entreprise C qui par ailleurs juge non nécessaire la segmentation de la page.

Étape 13. Indexation automatique et manuelle

L'indexation automatique peut se faire par OCR ou par codes-barres puisque le logiciel de capture Kofax Express qui pilote le scanner Fujitsu est doté de cette fonctionnalité. Pour le bon fonctionnement de cette étape, le prestataire et le client doivent définir ensemble les métadonnées à extraire et à saisir pour chaque lot ou chaque document. Cette entente concerne aussi :

> « *les règles métier qui permettront de vérifier la complétude et l'exactitude de ces données et les référentiels métiers (bases de données) qui pourront être utilisés pour la validation contextuelle des champs d'indexation*. » (DGSI, 2015, p.21)

Les métadonnées descriptives doivent généralement être acquises avant le début de la numérisation. Des exemples ou de schémas d'illustration de la part du client seraient nécessaires pour une meilleure compréhension. « Quelle que soit la qualité de ces dispositifs d'indexation automatique, les risques d'erreurs ne peuvent être totalement éliminés », souligne la DGSI, d'où l'intervention inévitable des opérateurs ou opératrices pour une validation manuelle.

Étape 14. Contrôle final

Il englobe les contrôles techniques et les contrôles visuels (vus à l'étape 10). Les contrôles techniques portent sur les éléments suivants (BnF, 2010, p.23) :

- la qualité des livrables et des supports de livraison ;
- le respect strict des normes et schémas fournis dans le cahier des charges ;
- le respect des paramètres techniques de numérisation (résolution, mode l'image, format…) ;
- l'exhaustivité et lisibilité des fichiers ;
- la conformité et cohérence des métadonnées par rapport aux schémas et fichiers fournis par le client ;
- la conformité des règles de segmentation et de structuration pour l'océrisation.

D'après Soyez (2009, p.25), « il est généralement recommandé de prévoir un contrôle systématique d'environ 10% du volume numérisé (par lot). »

Concernant le contrôle de la qualité des méthodes, les directives techniques FADGI insistent sur un examen continu qui s'étend à toutes les phases d'un projet de numérisation et au-delà, ainsi que sur la complémentarité entre techniques automatisées et manuelles.

> « *Il est moins probable que les techniques automatisées soient aussi efficaces pour évaluer l'exactitude, l'exhaustivité et l'utilité du contenu des métadonnées (selon sa complexité), ce qui nécessitera un certain niveau d'analyse manuelle. L'évaluation de la qualité des métadonnées nécessitera probablement une évaluation humaine qualifiée plutôt qu'une évaluation par machine.* » (Rieger, 2016, p.88).

Pour l'inspection technique des fichiers images, les directives techniques FADGI recommandent l'utilisation de JHOVE, un outil logiciel largement utilisé par les institutions du patrimoine culturel pour valider la conformité aux spécifications techniques des images.

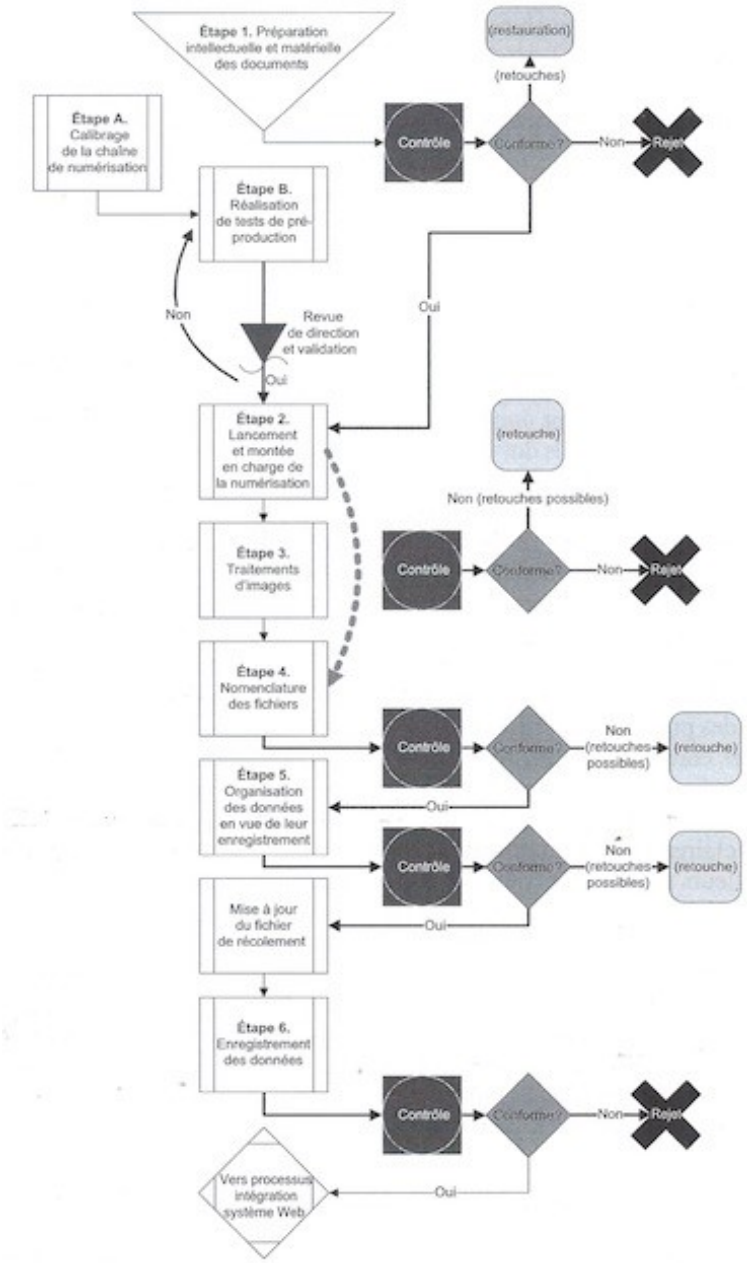
Étape 15 : Export des données

Les images numérisées et les données d'indexation associées sont à exporter vers les applications métiers : PEDAGO pour les données pédagogiques, DDE pour les données de mandat et OLOS pour les données de recherche.

Étape 16 : Restitution des données numériques et documents physiques

La voie et le support de restitution des données numériques et des originaux vont dépendre du choix du client. Nous avons vu que les entreprises interviewées utilisent plusieurs moyens de restitution : disques durs, serveurs NAS, Protocole de transfert de fichiers (FTP), Protocole sécurisé de transfert de fichiers (SFTP)… Dans tous les cas, il faut trouver un compromis qui permet de garantir la sécurité des documents lors de leur transfert.

Figure 11 : Illustration des moments de contrôle qualité dans un processus de numérisation sous forme de diagramme



(Claerr et Westeel, 2011, p.265)

5.3. Proposition d'une méthode de tarification des prestations de numérisation d'ArchiLab

Pour calculer le coût de la numérisation, il faut d'abord évaluer le temps nécessaire pour réaliser chacune des étapes du workflow de la **figure 10** ci-dessus. Pour illustrer cette idée, nous nous appuyons sur un exemple fictif tiré du livre de Claerr et Westeel (2011).

Pour le processus de numérisation fictif ci-après il faut :

- 1 heure de travail pour transporter un lot de 100 documents de 250 pages ;
- 3 minutes de préparation par document ;
- 10 secondes pour numériser une page ;
- 5 secondes par page pour réaliser un contrôle exhaustif des images numérisées et effectuer les post-traitements éventuels de rotation et de détourage ;
- 1 minute pour relire et corriger l'OCR de chaque page ;
- 15 minutes de contrôle-qualité par document.

Rassemblées, ces informations permettent d'obtenir des temps totaux indicatifs pour l'ensemble du processus de numérisation de ce projet fictif (**tableau 29**).

Tableau 29 : Exemple d'évaluation du temps nécessaire pour réaliser chaque étape d'un projet de numérisation fictif

Prestations	**Temps unitaire (s)**	**Quantité**	**Temps total**
Transport	3 600	1 (lot)	3 600
Préparation	180	100	18 000
Scannage	10	25 000	250 000
Contrôle et traitement d'images	5	25 000	125 000
Relecture OCR	60	25 000	1 500 000
Contrôle final	900	100	90 000
Total (secondes)			1 986 600
Total (heures)			552
Total (journées de 7 heures)			79
Total si on considère qu'un opérateur est productif 80% du temps			99

(Claerr et Westeel, 2011, p.139)

A partir de cette planification du temps de travail, et en nous inspirant des grilles tarifaires de l'entreprise C et de la Direction des Archives de France (DAF, 2008, Annexe 2), nous présentons dans le **tableau 30** une grille tarifaire adaptée et incluant les prestations de notre workflow de numérisation que nous proposons pour le laboratoire ArchiLab. Selon la Direction des Archives de France (DAF, 2008, Annexe 2, p.1), si les documents à numériser sont très hétérogènes, il est souhaitable de demander un prix spécifique en fonction de la typologie des documents et des difficultés particulières que présentent ces documents pour le scannage.

Tableau 30 : Grille tarifaire liée au workflow de numérisation d'archives proposé pour ArchiLab

Prestation	**Unité**	**Prix unitaire CHF**	**Rabais en fonction des quantités**	**TVA**	**Prix unitaire TTC (CHF)**
Transport	Aller/retour en fonction de la distance				
Inventaire	Volume (ml)				
Préparation matérielle et intellectuelle des documents, reconditionnement	Forfaitaire				
Tests de pré-scannage	Forfaitaire				
Scannage proprement dit	Page et document				
Rangement	Forfaitaire				
Conversion en TIFF	Page et document				
Traitements d'images (détourage, redressement…)	Page et document				
Contrôle qualité (complétude et lisibilité des images)	Page et document				
Conversion en format PDF avec OCR ou sans OCR	Page et document, Champ (de X caractères)				
Relecture du texte OCR	Page				
Structuration (agrégation) manuelle et nommage des fichiers	Agrégat				
Indexation selon la notice Dublin Core	Champ				
Contrôle final de la qualité et validation	Page et document				
Stockage	Support (disque, clé)				
Restitution des documents numérisés	Support (disque, clé…)				
Restitution des documents physiques	Aller/retour en fonction de la distance				
TOTAL					

6. Conclusions et recommandations

6.1. Principales conclusions

Au terme de cette étude sur les pratiques de numérisation au sein des entreprises privées en Suisse, nous pouvons émettre des conclusions suivantes :

Un projet de numérisation est un projet de production qui mobilise des compétences (domaines) multidisciplinaires : aspects techniques informatiques, aspects juridiques et normatifs, aspects qualitatifs et quantitatifs et aspects managériaux. Mettre en place un projet/service de numérisation demande donc de maîtriser tous ces aspects.

Cette étude montre que les entreprises interviewées et prestataires de numérisation appartiennent à trois catégories quant à la typologie des documents numérisés : entreprises spécialisées en numérisation patrimoniale (Entreprise A, B, C et D), entreprises spécialisées en numérisation de production ou documentaire classique (Entreprise F et G) et une entreprise mixte (Entreprise E).

L'étude montre aussi que ces entreprises utilisent du matériel (scanner et logiciel) varié pour numériser les documents de leurs clients : scanners à plat (largement dominants), scanners verticaux, scanners à chargeur, scanners à tambour, appareil photo numérique, caméras planétaires, scanners à bac et scanners à tapis roulant. Le scanner type Fujitsu à plat est utilisé par une entreprise sur sept.

Les logiciels utilisés sont aussi variés : Kodak capture Pro, HD Ultra, Copy book, Autoscan, Limb capture, Youdoc, Capture One, Silverfast, 4Digitalbooks, I2S, Newcolor 7000, Kofax et DocProStar. Sauf Kodak Capture Pro qui est utilisé par deux entreprises sur sept, les autres logiciels sont utilisés par une seule entreprise.

Le choix de ce matériel et les paramètres techniques de numérisation (résolution, profondeur de couleurs, colorimétrie, formats de fichier, type et taux de compression…) dépend des objectifs du projet (conservation, diffusion, valorisation), de la nature et de l'état physique des documents.

D'une manière générale, les entreprises numérisent les documents avec une résolution minimum de 300 dpi. TIFF non compressé, PDF, PDF/A et JPEG 2000 sont des formats de conservation les plus utilisés par les entreprises sondées. Tandis que Dublin Core en XML/JSON est le format de métadonnées le plus utilisé.

Le projet de numérisation de chaque entreprise commence en amont en écoutant le client, en discutant avec lui afin de définir ensemble et de manière claire les livrables attendus, leurs caractéristiques et les délais de restitution.

Tous les workflows de numérisation des entreprises ayant participé à cette étude comportent une phase de calibrage et de tests (projet pilote) dont les résultats (images numérisées, nommage des fichiers, métadonnées…) sont envoyés au client pour validation.

Toutes les entreprises sont unanimes pour dire que l'étape de préparation matérielle et intellectuelle des documents est l'étape la plus longue, la plus complexe et la plus coûteuse du processus de numérisation. Elle influe de façon significative sur le coût du projet.

Pour des raisons de compétences, seulement peu d'entreprises proposent les services d'inventaire d'archives (Entreprise B : de façon systématique, Entreprise A : sur demande, Entreprise E : inventaire sommaire).

Toutes les entreprises proposent les services d'océrisation brute, mais seulement trois entreprises sur sept procèdent à une relecture du fichier texte généré. Une entreprise sur sept procède à cette relecture uniquement sur demande du client. D'après les représentants de toutes les entreprises, la majorité des clients se contentent de l'OCR brut ; très peu de clients demandent de l'OCR raffiné pour des raisons liées au coût élevé de la prestation.

Toutes les entreprises déclarent pratiquer une indexation minimum des images numérisées selon les consignes fournies par le client, à qui revient le travail d'indexation complète. Dans les cas les plus simples, l'opération consiste, à l'aide d'un code-barres, à mettre en relation la notice du document fourni par le client (original) avec le document qui vient d'être numérisé. Pour les documents iconographiques, il peut tout simplement s'agir d'une légende à saisir, communiquée par le client.

Toutes les entreprises sondées procèdent à des contrôles quantitatifs et qualitatifs des images numérisées, mais le nombre, les moments et les méthodes utilisées diffèrent légèrement. Au moment où certaines entreprises font deux contrôles (le minimum), d'autres peuvent en effectuer quatre, voire un contrôle à la fin de chaque étape du processus de numérisation.

Toutes les entreprises déclarent respecter les lois et règlements applicables en numérisation, mais ceux-ci différent en fonction de la typologie des documents numérisés et du pays où se déroule le projet de numérisation. Pour les normes, à l'exception de l'entreprise C qui mise sur la confiance et la qualité du travail accompli, les autres entreprises sont certifiées pour un certain nombre de normes de qualité et de sécurité s'appliquant sur le document numérique (sa numérisation, sa diffusion, son archivage, sa sécurité, etc.).

La restitution des données numériques se fait au moyen des disques durs, du protocole de transfert de fichiers (FTP), du protocole sécurisé de transfert de fichiers (SFTP), de serveurs comme NAS et autres. Dans certains cas, la clé USB peut être utilisée, mais seulement pour des faibles quantités de données.

Toutes les entreprises facturent leurs prestations de numérisation soit à la page et au document (dossier ou agrégat), soit à l'unité de temps (heure, jour, semaine) ou au forfait.

Au vu des pratiques de numérisation en vigueur dans les entreprises interviewées, nous pouvons dire que le scanner Fujitsu et le logiciel Kofax Express choisis par ArchiLab, sont en mesure d'effectuer toutes les étapes du workflow complet de numérisation proposé dans cette étude (que ce soit dans le cadre de la formation des étudiant-e-s ou de prestations de service).

Nous voulons terminer notre conclusion en soulignant l'importance des étapes d'avant et d'après scannage des documents dans la réussite d'un projet de numérisation :

> « *La numérisation n'est pas assimilable à une simple reproduction, elle exige de détailler les processus et d'appliquer des normes, elle doit être justement et complètement renseignée (les ″métadonnées″) pour donner une chance aux travaux numériques d'être considérés comme du patrimoine numérique. Ce n'est pas la prise de vue qui prend du temps, c'est tout ce qui la précède et ce qui la suit. Autant dire que toute démarche numérique demande une préparation rigoureuse, une organisation soignée, écrite et solidaire des étapes.* » (ACV, 2016, pp.38-39)

Dans ce travail, nous avons rencontré les difficultés suivantes :

La première difficulté concerne le temps long d'attente de réponse à notre sollicitation d'entretien (cela concerne 3 entreprises sur 7). D'autres entreprises n'ont pas donné suite à notre demande et ce malgré de nombreux rappels (par téléphone et par écrit).

La deuxième difficulté concerne l'analyse des étapes des workflows de numérisation des entreprises ayant participé à cette étude. Seulement deux entreprises sur sept (entreprise F et G) affichent leurs workflows de numérisation sur leurs sites Internet. Pour les autres, nous ne savons pas pourquoi, car nous n'avons posé cette question (elle nous a échappée). Il nous a donc fallu beaucoup de temps pour synthétiser les étapes des workflows de toutes entreprises selon les informations obtenues lors des entretiens. Le travail est si difficile d'autant plus que, pour un workflow de numérisation donné, l'ordre des étapes n'est pas figé.

La troisième difficulté concerne la réticence de certaines entreprises à communiquer leurs prix de numérisation. Sur sept entreprises, seulement une entreprise (l'entreprise D) affiche sur son site Internet les prix de ses prestations de numérisation. Pour les autres, nous les comprenons puisqu'il s'agit de la concurrence. Mais selon l'économie libérale, la transparence des prix fait partie des cinq conditions qui caractérisent un marché de concurrence pure et "parfaite"[34]. Nous rappelons qu' en tant qu'institution publique, la Haute école de gestion de Genève (HEG), n'a pas l'intention, à travers cette étude, d'entrer en concurrence avec les entreprises privées en matière de numérisation d'archives. Dans cette étude, les prix sont demandés à titre informatif. Nous remercions donc les quelques entreprises qui ont accepté de nous communiquer leurs prix (qu'ils soient indicatifs ou non).

Enfin, sept entreprises sont beaucoup comme échantillon à étudier par une seule personne ; surtout si l'étude comme celle-ci inclut des entretiens et porte sur un nombre élevé de paramètres à analyser.

6.2. Recommandations

Au vu de tout ce qui procède, et dans le but de dynamiser l'enseignement archivistique à travers l'atelier de numérisation du nouveau laboratoire ArchiLab, nous formulons les recommandations suivantes :

A. Numérisation dans le cadre de la formation des étudiant-e-s (application pratique) et de prestations de service

Mettre en place un workflow complet de numérisation incluant l'**inventaire**, le **projet pilote** (phase de tests), l'**océrisation**, l'**indexation** (automatique et manuelle) et la **restauration** des documents comme prestations incontournables.

1) Projet pilote

Outre les images (netteté, orientation…), l'exactitude des métadonnées, les formats, Et autres, **la phase de tests (projet pilote) doit servir à mesurer l'espace nécessaire pour l'archivage des documents numériques**.

[34] Atomicité (plusieurs acteurs), homogénéité du produit, fluidité (absence d'obstacles d'entrée sur le marché), transparence des prix et libre circulation des facteurs de production.

2) Formation des étudiant-e-s dans les différentes phases du processus de numérisation (études de cas concrètes et mise en situation réelle professionnelle)

L'initiation des étudiant-e-s aux pratiques basiques de la numérisation des documents devrait prendre en considération les trois workflows de numérisation proposés dans cette étude (le workflow principal et les deux autres workflows présentés en annexe). Ceci afin de mieux les préparer à la participation aux projets de numérisation qui seront effectués dans le cadre de mandats (prestations de service) et de recherche.

3) Inventaire dans Atom et description archivistique selon la norme ISAD(G)

Considérer l'inventaire archivistique comme une des prestations phares du Service de numérisation d'archives de ArchiLab afin de transmettre aux étudiant-e-s les compétences pratiques en traitement et gestion d'archives.

4) Contrôle qualitatif et quantitatif des produits de la numérisation

Effectuer un contrôle qualitatif et quantitatif à la fin de chaque étape du processus de numérisation.

5) Gestion des livrables tout au long du processus de numérisation

Travailler avec le **fichier de récolement** (au format Excel) à compléter au fur et à mesure de l'avancement du projet pour plus d'efficacité.

6) Planification d'un projet de numérisation

Travailler avec la **grille d'évaluation** qui consiste à identifier toutes les tâches à réaliser et à évaluer la **durée estimée de chaque étape** (heures, jours, semaines), en particulier la préparation des documents (étape la plus longue et la plus coûteuse).

7) Restauration éventuelle des documents

Evaluer la nécessité de mettre en place des outils de **restauration des documents** (restauration comme un des services de prétraitements afin de pouvoir récupérer toute l'information portée par un document en fonction de son état physique).

8) Scanner avec caméra planétaire (appareil photo numérique)

S'équiper d'un appareil photo numérique avec caméra planétaire comme mode de numérisation. Ce type de scanner est approprié pour numériser les documents fragiles.

9) **Evaluer les besoins en formation de l'équipe de projet de numérisation**.

B. Evaluation du bilan d'un projet de numérisation porté par ArchiLab

Avant d'initier un nouveau projet de numérisation porté par ArchiLab, il faudra d'abord évaluer le projet précédent afin d'en tirer des enseignements. Voici les critères à prendre en considération pour cette évaluation (BCI, 2014, p.36).

Critères d'évaluation	Description
Pertinence	Dans quelle mesure les objectifs fixés ont-ils répondu aux problèmes identifiés ou aux besoins réels ?
Efficacité	Mesure de l'écart entre les objectifs et les résultats.
Efficience	Est-ce que les objectifs ont-ils été atteints à moindre coût (financier, humain, organisationnel) ?
Impact (quantitatif ou qualitatif)	Retombées du projet de numérisation à moyen et long terme (degré de satisfaction des bénéficiaires).
Durabilité	Est-ce que les effets du projet perdureront après sa clôture ?
Délai imparti	Temps de réalisation du projet en regard du temps prévu au départ.
Ecarts budgétaires	Ecarts positifs (coûts moindres que prévus) ou écarts négatifs (coûts supérieurs à ceux prévus). Quelles en sont les causes ?

Références bibliographiques

4DIGITALBOOKS, 2022. 4Digitalbooks [en ligne]. [Consulté, le 12 avril 2022]. Disponible à l'adresse : www.4digitalbooks.com

AGENCE DE MUTUALISATION DES UNIVERSITES ET ETABLISSEMENTS (AMUE), 2010. *La gestion des archives au sein d'un établissement d'enseignement supérieur et de recherche* [en ligne]. [Consulté, le 18 juillet 2022]. Disponible à l'adresse: https://www.amue.fr/fileadmin/amue/documents-publications/amue/Gestion_archives_web.pdf

AGENCE RHÔNES-ALPES POUR LE LIVRE ET LA DOCUMENTATION (ARALD), 2010. *Lexique du livre numérique* [en ligne]. [Consulté, le 18 juillet 2022]. Disponible à l'adresse : https://www.enssib.fr/bibliotheque-numerique/documents/57091-lexique-du-livre-numerique.pdf

ARCHIVES D'ETAT DE GENEVE (AEG), 2010. *Glossaire* [en ligne]. Genève : Groupe de travail "Records management et archivage définitif des documents électroniques". [Consulté, le 30 novembre 2021]. Disponible à l'adresse : https://ge.ch/archives/media/site_archives/files/imce/pdf/procedures/glossaire_rm_archdef_v1_0.pdf

ARCPLACE, 2020. Arcplace [en ligne]. [Consulté, le 21 juillet 2022]. Disponible à l'adresse : https://www.arcplace.ch/fr/solutions/technologies/input-management/

ARKHÊNUM, 2022. Arkhênum [en ligne]. [Consulté, le 12 avril 2022]. Disponible à l'adresse: www.arkhenum.fr

ASSOCIATION DES ARCHIVISTES SUISSES (AAS), 2009. *Directives suisses pour l'application de la norme générale internationale de description archivistique ISAD(G)* [en ligne]. [Consulté, le 22 juin 2022]. Disponible à l'adresse : https://archiv.vsa-aas.ch/wp-content/uploads/2015/06/Richtlinien_ISAD_G_VSA_f.pdf

ASSOCIATION NATIONALE DES DIRECTEURS ET DIRECTEURS-ADJOINTS DES CENTRES DE GESTION DE LA FONCTION PUBLIQUE TERRITORIALE (ANDCDG), 2018. *Les archives, un atout pour la modernisation de l'administration territoriale* [en ligne]. [Consulté, le 18 juillet 2022]. Disponible à l'adresse: https://www.cdg84.fr/wp-content/uploads/2018/07/2018_Guide_archives_web.pdf

ASSOCIATION SUISSE DES PROFESSIONNELS DE LA ROUTE ET DES TRANSPORTS (VSS), 2013. *Normalisation et droit-Le statut juridique des normes* [en ligne]. [Consulté, le 11 juillet 2022]. Disponible à l'adresse : https://www.vss.ch/fileadmin/user_upload/vss/downloads/Normalisation_et_droit.pdf

AUGUIE Katell et VIALLE Coline, 2017. *La gestion de archives. Maîtriser les documents et les données.* Voiron : Territorial éditions, Collection "Dossier d'Experts". 140p. ISBN 978-2-8186-1159-3.

BIBLIOTHEQUE ET ARCHIVES NATIONALES DU QUEBEC (BAnQ), 2019. La numérisation des documents administratifs. Méthodes et recommandations [en ligne]. [Consulté, le 11 mars 2022]. Disponible à l'adresse :

https://www.banq.qc.ca/documents/archives/archivistique_gestion/aide_conseil/Guide_numerisation_documents_administratifs_VF(2019-06-17).pdf

BIBLIOTHEQUE ET ARCHIVES NATIONALES DU QUEBEC (BAnQ), BIBLIOTHEQUE NATIONALE DE FRANCE (BnF) et MUSEE CANADIEN DE L'HISTOIRE (MCH), 2014. *Recueil de règles de numérisation* [en ligne]. [Consulté, le 5 août 2022]. Disponible à l'adresse :https://bibliopiaf.ebsi.umontreal.ca/bibliographie/3F228PD9/download/IYJV8C4Y/Anctil%20et%20al.%20-%202014%20-%20Recueil%20de%20règles%20de%20numérisation.pdf

BIBLIOTHEQUE NATIONALE DE FRANCE (BnF), 2021. *Formats de données pour la préservation à long terme : la politique de la BnF. Version initiale pour appel à commentaires* [en ligne]. [Consulté, le 25 mars 2022]. Disponible à l'adresse : https://www.bnf.fr/sites/default/files/2021-04/politiqueFormatsDePreservationBNF_20210408.pdf

BIBLIOTHEQUE NATIONALE DE FRANCE (BnF), 2018. *Référentiel de numérisation des documents opaques* [en ligne]. [Consulté, le 5 août 2022]. Disponible à l'adresse : https://www.bnf.fr/sites/default/files/2018-11/ref_num_doc_opaques_v2.pdf

BIBLIOTHEQUE NATIONALE DE FRANCE, 2015. *Référentiel OCR* [en ligne]. [Consulté, le 07 janvier 2022]. Disponible à l'adresse : https://www.bnf.fr/sites/default/files/2018-11/ref_num_ocr_v2.pdf

BIBLIOTHEQUE NATIONALE DE FRANCE (BnF) ET BIBLIOTHEQUE MUNICIPALE CLASSEE D'ORLEANS (BMCO), 2010. *Écrire un cahier des charges de numérisation et de conversion en mode texte de collections de presse* [en ligne]. [Consulté, le 1er mars 2022]. Disponible à l'adresse : https://francearchives.fr/file/9180b7a6e40c2a25b6757174a71702258efae66e/static_4115.pdf

BIBLIOTHEQUE NATIONALE DE FRANCE (BnF) ET BIBLIOTHEQUE MUNICIPALE CLASSEE D'ORLEANS (BMCO), 2010. *Écrire un cahier des charges de numérisation et de conversion en mode texte de collections de presse. Annexes* [en ligne]. [Consulté, le 3 août 2022]. Disponible à l'adresse : https://francearchives.fr/file/c4fcb863fa4e28acbd072c0a5a3a2ffe9b28f80d/static_4113.pdf

BIBLIOTHEQUE NATIONALE SUISSE (BN), 2017. *Ligne de conduite pour la numérisation* [en ligne]. [Consulté, le 11 mars 2022]. Disponible à l'adresse : https://www.nb.admin.ch/dam/snl/fr/dokumente/e-publikationen/publikationen/digitalisierungsleitlinienb.pdf.download.pdf/ligne_de_conduitepourlanumerisationbn.pdf

BIBLIOTHEQUE NATIONALE SUISSE (BN), BIBLIOTHEQUES CANTONALES DE SUISSE ROMANDE ET RESEAU DES BIBLIOTHEQUES DE SUISSE OCCIDENTALE (RERO), 2007. *Un patrimoine en ligne. Digitaliser les collections historiques et contemporaines de la presse suisse pour en garantir la pérennité et les rendre accessibles. Recommandations* [en ligne]. [Consulté, le 17 juin 2022]. Disponible à l'adresse : https://www.digicoord.ch/images/e/ef/PresseSuisse_numerisation_journaux_-_version_def.pdf

BRAULT Chloé, 2021. *Numérisation et qualité. Guide de gestion et de contrôle de la qualité dans un projet de numérisation patrimoniale. Version 1.1-septembre 2021* [en ligne]. *Bruxelles : ADOCHS.* [Consulté, le 23 août 2022]. Disponible à l'adresse : https://www.cegesoma.be/sites/www.cegesoma.be/files/Version%20fr_compressed_1.pdf

BUREAU DE COOPERATION INTERUNIVERSITAIRE (BCI), 2014. *Guide de gestion d'un projet de numérisation* [en ligne]. [Consulté, le 11 mars 2022]. Disponible à l'adresse : https://www.enssib.fr/bibliotheque-numerique/documents/64628-guide-de-gestion-d-un-projet-de-numerisation.pdf.

CLAERR Thierry et WESTEEL Isabelle (dir.), 2011. *Manuel de la numérisation*. Paris : Cercle de la Librairie, "Collection Bibliothèques". 310p. ISBN 978-2-7654-0983-0.

Code de procédure civile Suisse du 18 décembre 2008 (CPC). L'Assemblée fédérale de la Confédération suisse [en ligne]. 18 décembre 2008. Mise à jour le 1er juillet 2022. [Consulté, le 15 juillet 2022]. Disponible à l'adresse : https://www.fedlex.admin.ch/eli/cc/2010/262/fr

CONFERENCE DES RECTEURS ET DES PRINCIPAUX DES UNIVERSITES DU QUEBEC (CREPUQ), 2013. *Guide de gestion d'un projet de numérisation de documents. Référentiel technique* [en ligne]. [Consulté, le 25 février 2022]. Disponible à l'adresse : https://www.lccjti.ca/files/sites/105/2014/10/Guide_Gestion_Projet_Num_Ref_Technique_v0200.pdf.

DIGICOORD, 2002. Digicoord [en ligne]. [Consulté, le 4 avril 2022]. Disponible à l'adresse : https://www.digicoord.ch/index.php/Accueil

DIRECTION DES ARCHIVES DE FRANCE (DAF), 2008. *Écrire un cahier des charges de numérisation du patrimoine. Guide technique. Documents reliés, manuscrits, plans, dessins, photographies, Microformes* [en ligne]. [Consulté, le 11 mars 2022]. Disponible à l'adresse : https://francearchives.fr/file/bf50d8fa5f554586dbf18fdc862d25970a1da0a7/static_4132.pdf

DIRECTION DES ARCHIVES DE FRANCE (DAF), 2008. *Écrire un cahier des charges de numérisation du patrimoine. Guide technique. Documents reliés, manuscrits, plans, dessins, photographies, Microformes. Annexe 2 : Exemples de bordereaux des prix unitaires* [en ligne]. [Consulté, le 14 juillet 2022]. Disponible à l'adresse : https://archives.var.fr/_depot_ad83/datas/ark_cms/_depot_arko/articles/13825/siaf-num-annexe2_doc.pdf

DIRECTION GENERALE DES SYSTEMES D'INFORMATION (DGSI), 2015. *Guide de la dématérialisation* [en ligne]. Genève : République et canton de Genève, Département de la sécurité et de l'économie. [Consulté, le 18 mars 2022]. Disponible à l'adresse : https://ge.ch/archives/media/site_archives/files/imce/pdf/procedures/20151221_guide_dematerialisation_v2_21.pdf

DOCUTEAM, 2022. Docuteam [en ligne]. [Consulté, le 6 avril 2022]. Disponible à l'adresse : www.docuteam.ch

DUFEY Jean-Philippe, 2018. *De la capture à la capture intelligente* [en ligne]. [Consulté, le 21 juillet 2022]. Disponible à l'adresse : (https://www.ictjournal.ch/articles/2018-09-24/de-la-capture-a-la-capture-intelligente)

DUNANT GONZENBACH Anouk et DROZE Pascal, 2016. *Directive transversale. Dématérialisation des documents à valeur probante* [en ligne]. Genève : République et canton de Genève, Département présidentiel. [Consulté, le 18 mars 2022]. Disponible à l'adresse :https://ge.ch/archives/media/site_archives/files/imce/pdf/procedures/ege-08-02_v1.pdf

EQUIPE VITAM, 2020. *Programme Vitam-Extraction des métadonnées techniques, version 2* [en ligne]. [Consulté, le 25 mars 2022]. Disponible à l'adresse : https://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/20200131_NP_Vitam_preservation-extraction-MD-v2.0.pdf

ESSEVAZ-ROULET Baptiste, 2016. *La numérisation d'archives. Des fondamentaux techniques aux programmes de numérisation*. Voiron : Territorial éditions, Collection "Dossier d'Experts". 117p. ISBN 978-2-8186-1042-8.

EVERIAL, 2020. Everial [en ligne]. [Consulté, le 13 avril 2022]. Disponible à l'adresse : www.everial.com

EVERIAL, 2020. *Everial-plaquette-corporate-1* [en ligne]. [Consulté, le 14 juin 2022]. Disponible à l'adresse : https://www.everial.com/wp-content/uploads/2020/03/EVERIAL-plaquette-corporate-1.pdf

GRASSIN Geoffroy, 2008. *Numérisation des fonds patrimoniaux. Fiche pratique Enssib* [en ligne]. [Consulté, le 24 juin 2022]. Disponible à l'adresse : https://www.enssib.fr/bibliotheque-numerique/documents/21198-numerisation-des-fonds-patrimoniaux.pdf

GROUPE-T2I, 2022. Groupe-T2i [en ligne]. [Consulté, le 18 mars 2022]. Disponible à l'adresse : www.groupe-t2i.com

GUEIT-MONTCHAL Lydiane et BOULLY Vincent, 2020. *Les archives à l'âge courant et intermédiaire*. In : GUEIT-MONTCHAL Lydiane (dir.). Abrégé d'archivistique. Principes et Pratiques du métier d'archiviste. 4ème édition refondue et augmentée. Paris : Association des archivistes français, pp.47-95. Collection Archivistes français formation (aff). ISBN 978-2-900175-09-5.

HIGHENDSCAN, 2022. Highendscan [en ligne]. [Consulté, le 18 mai 2022]. Disponible à l'adresse : www.highendscan.ch

KOFAX, 2019. *Kofax Express. Présentation du produit* [en ligne]. [Consulté, le 10 mars 2022]. Disponible à l'adresse : https://www.kofaxfrance.fr/-/media/Files/Datasheets/FR/ps_kofax-express_fr.pdf

LOCARCHIVES, 2022. *Notre solution : numérisation fidèle* [en ligne]. [Consulté, le 17 juin 2022]. Disponible à l'adresse : https://locarchives.fr/services-documentaires/numerisation-de-documents/notre-solution-numerisation-fidele/

Loi du 9 octobre 2008 sur l'information du public, la protection des données et l'archivage (LIPDA, 170.2). Le Grand Conseil du canton du Valais [en ligne]. 9 octobre 2008. Mise à jour le 1er septembre 2011. [Consulté, le 15 juillet 2022]. Disponible à l'adresse : https://lex.vs.ch/frontend/versions/1589/download_pdf_file?locale=fr

Loi du 11 septembre 2007 sur la protection des données personnelles (LPrD, 172.65). Le Grand Conseil du canton de Vaud [en ligne]. 11 septembre 2007. Mise à jour le 1er janvier 2009. [Consulté, le 15 juillet 2022]. Disponible à l'adresse suivante : https://avenirsocial.ch/wp-content/uploads/2018/12/LV_PrD.pdf

Loi du 5 octobre 2001 sur l'information du public, l'accès aux documents et la protection des données personnelles (LIPAD, A 2 08). Le Grand Conseil de la République et canton de

Genève [en ligne]. 5 octobre 2001. Mise à jour le 17 octobre 2020. [Consulté, le 15 juillet 2022]. Disponible à l'adresse suivante :
https://silgeneve.ch/legis/data/rsg/rsg_a2_08.htm?myVer=1657878933187

Loi du 19 février 1986 sur la protection des données (LCPD, 152.04). Le Grand Conseil du canton de Berne [en ligne]. 19 février 1986. Mise à jour le 1er janvier 2020. [Consulté, le 15 juillet 2022]. Disponible à l'adresse :
https://www.belex.sites.be.ch/frontend/versions/1509/download_pdf_file?locale=fr

Loi fédérale du 9 octobre 1992 sur le droit d'auteur et des droits voisins (Loi sur le droit d'auteur, LDA). L'Assemblée fédérale de la Confédération suisse [en ligne]. 9 octobre 1992. Mise à jour le 1er avril 2020. [Consulté, le 15 juillet 2022]. Disponible à l'adresse suivante :
https://fedlex.data.admin.ch/filestore/fedlex.data.admin.ch/eli/cc/1993/1798_1798_1798/20200401/fr/pdf-a/fedlex-data-admin-ch-eli-cc-1993-1798_1798_1798-20200401-fr-pdf-a.pdf

Loi fédérale du 19 juin 1992 sur la protection des données (LPD). L'Assemblée fédérale de la Confédération suisse [en ligne]. 19 juin 1992. Mise à jour le 1er mars 2019. [Consulté, le 15 juillet 2022]. Disponible à l'adresse suivante :
https://fedlex.data.admin.ch/filestore/fedlex.data.admin.ch/eli/cc/1993/1945_1945_1945/20190301/fr/pdf-a/fedlex-data-admin-ch-eli-cc-1993-1945_1945_1945-20190301-fr-pdf-a.pdf

MAKHLOUF SHABOU Basma, TURNER Cécile, NICOLET Aurèle et NOIRJEAN Aimée, 2021. ArchiLab : *Projet de laboratoire archivistique. Version 5*. Genève : Haute école de gestion de Genève (HEG), 26 mars 2021, 10p.

MASSOT Marie-Laure (coord.), 2020. *Glossaire. Introduction aux humanités numériques* [en ligne]. [Consulté, le 18 juillet 2022]. Disponible à l'adresse :
https://hal.archives-ouvertes.fr/hal-02410396/file/Glossaire_HN_31_01_20.pdf

MEMORIAV.2017. *Recommandations photo 2017* [en ligne]. [Consulté, le 20 juin 2022]. Disponible à l'adresse : https://memoriav.ch/wp-content/uploads/2017/12/Memoriav-recommandations-Photo-2017.pdf

MICALETTI Marie-Angèle, 2001. *Création d'un produit documentaire électronique* [en ligne]. Mont Saint Aignan : Institut Régional des Techniques Documentaires. [Consulté, le 19 juillet 2022]. Rapport de stage. Disponible à l'adresse :https://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_7/divers2/010025437.pdf

MKADMI Abderrazak, 2021. *Les archives à l'ère du numérique, préservation et droit à l'oubli*. London : Éditions ISTE, Série Outils et usages numériques, vol. 6. 210p. ISBN 9 781784 057343.

MOCELLIN Catherine, 2010. *Maîtriser les aspects techniques de la numérisation*. In : CLAERR Thierry et WESTEEL Isabelle (dir.)[en ligne]. Villeurbanne : Presse de l'Enssib, pp. 19-43. [Consulté, le 3 mars 2022]. Disponible à l'adresse :
https://books.openedition.org/pressesenssib/pdf/423

Ordonnance du 24 avril 2002 concernant la tenue et la conservation des livres de comptes (Olico, 221.431). Le Conseil fédéral suisse [en ligne]. 24 avril 2002. Mise à jour le 1er janvier 2013. [Consulté, le 15 juillet 2022]. Disponible à l'adresse :
https://fedlex.data.admin.ch/filestore/fedlex.data.admin.ch/eli/cc/2002/216/20130101/fr/pdf-a/fedlex-data-admin-ch-eli-cc-2002-216-20130101-fr-pdf-a.pdf

ORGANISATION INTERNATIONALE DE NORMALISATION (ISO), 1999. Règles d'échantillonnage pour les contrôles par attributs. Partie 1 : Procédures d'échantillonnage pour les contrôles lot par lot, indexés d'après le niveau de qualité acceptable (NQA) [en ligne]. Genève : ISO 1999. Norme internationale ISO 2859-1 : 1999. [Consulté, le 28 juin 2022]. Disponible à l'adresse : https://cdn.standards.iteh.ai/samples/1141/16a4b4110f6d4790aaeda106510332f1/ISO-2859-1-1999.pdf

PACIFIC REGIONAL BRANCH INTERNATIONAL COUNCIL ON ARCHIVES (PARBICA), 2021. *Boîte à outils d'archivage pour une bonne gouvernance. Directive 12 : Introduction à l'archivage numérique* [en ligne]. [Consulté, le 16 février 2022]. Disponible à l'adresse : https://parbica.org/wp-content/uploads/2021/09/Guideline-12-Introduction-to-Digital-Recordkeeping_FR.pdf

PERROT Patrick, 2016. *Module 9-Section 2 : Numériser les documents* [en ligne]. Support de cours : « Module 9 - Reproduction par microfilmage et numérisation », Portail International Archivistique Francophone (PIAF). [Consulté, le 26 juillet 2022]. Disponible à l'adresse : https://www.piaf-archives.org/sites/default/files/bulk_media/m09s2/section2_papier.pdf

PORTAIL INTERNATIONAL ARCHIVISTIQUE FRANCOPHONE (PIAF), 2015 *Glossaire, version 2* [en ligne]. [Consulté, le 3 août 2022]. Disponible à l'adresse : https://www.piaf-archives.org/sites/default/files/bulk_media/glossaire/glossaire_papier.pdf

PROFILS ICC. Wikipedia : l'encyclopédie libre [en ligne]. Dernière modification de la page le 23 février 2021. [Consulté, le 14 octobre 2022]. Disponible à l'adresse : https://fr.wikipedia.org/wiki/Profil_ICC

PROGRAMME NATIONAL DE NUMERISATION ET DE VALORISATION DE CONTENUS CULTURELS (PARBICA), 2017. *Recommandations techniques pour les métadonnées et les standards. Version n°1 2* [en ligne]. Paris : République Française, Ministère de la Culture, Secrétariat général. [Consulté, le 10 mars 2022]. Disponible à l'adresse : https://parbica.org/wp-content/uploads/2021/09/Guideline-12-Introduction-to-Digital-Recordkeeping_FR.pdf

REGAMEY Liliane, 2022. Re: *Dig_spécifications.docx* [message électronique]. 5 mai 2022.

REPRESENTANT DE L'ENTREPRISE A, 2022. Re : *Entretien du 3 mai 2022 : deux dernières demandes* [message électronique]. 16 août 2022.

REPRESENTANT DE L'ENTREPRISE B, 2022. *Archivage numérique avec docuteam cosmos* [message électronique]. 30 mai 2022.

REPRESENTANT DE L'ENTREPRISE C, 2022. Re: *Entretien du 27 mai 2022: demande de compléments* [message électronique]. 27 juin 2022.

REPRESENTANT DE L'ENTREPRISE E, 2022. *Re : Information* [message électronique]. 8 août 2022.

RIEGER Thomas, 2016. *Technical Guidelines for Digitizing Cultural Heritage Materials*. Creation of Raster Images Files [en ligne]. Washington: FADGI-Federal Agencies Digital Guidelines Initiative. [Consulté, le 23 août 2022]. Disponible à l'adresse : https://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf

SERVICE DU SECRETARIAT GENERAL ET DES COMMUNICATIONS (SSGC), 2019. *Procédure de numérisation des documents (version finale)* [en ligne]. *Québec : Commission scolaire du Pays des Bleuets.* [Consulté, le 25 février 2022]. *Cahiers des écrits de gestion*, première version : 16 avril 2019 (CC-8348-04-19). Disponible à l'adresse : https://www.cspaysbleuets.qc.ca/images/123-01_Procédure_de_numérisation_des_documents_version_finale_16_avril_2019.pdf

SOYEZ Sébastien, 2009. *La numérisation en marche. Les étapes de la dématérialisation des processus de travail* [en ligne]. Bruxelles : Archives générales du Royaume et Archives de l'Etat dans les Provinces. [Consulté, le 1er mars 2022]. Brochures des recommandations et de conseil. Section Surveillance, avis et coordination de la collecte et de la sélection. Disponible à l'adresse : https://arch.arch.be/docs/brochures/la_numerisation_en_marche.pdf

TAVARES Sara, 2013. *La gestion des documents iconographiques patrimoniaux : recommandations pour le projet d'acquisition de la Collection iconographique par la Bibliothèque cantonale et universitaire-Lausanne*. Genève : Haute école de gestion. Travail de bachelor [en ligne]. [Consulté, le 5 août 2022]. Disponible à l'adresse : https://core.ac.uk/download/pdf/20662381.pdf

TESSIER Marc. 2010. *Rapport sur la numérisation du patrimoine écrit* [en ligne]. Paris : Ministère de la culture et de la communication. [Consulté, le 12 juillet 2022]. Disponible à l'adresse : https://www.vie-publique.fr/sites/default/files/rapport/pdf/104000016.pdf

VINSONNEAU Emile, 2015. *La qualité d'image dans le contexte de la numérisation de livres anciens* [en ligne]. Bordeaux : Université de Bordeaux-Ecole doctorale de mathématique et informatique, spécialité Informatique. Thèse de doctorat. [Consulté, le 23 août 2022). Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-01233181/file/VINSONNEAU_EMILE_2015.pdf

WESTEEL Isabelle, 2010. *Indexer, structurer, échanger : métadonnées et interopérabilité*. In : CLAERR Thierry et WESTEEL Isabelle (dir.)[en ligne]. Villeurbanne : Presse de l'Enssib, pp. 99-123. [Consulté, le 3 mars 2022]. Disponible à l'adresse : https://books.openedition.org/pressesenssib/434

Annexe 1 : Glossaire

Sources : AMUE (2019, pp.72-78) ; ARALD (2010 ; ANDCDC (2018, pp.120-133) ; Archives d'État de Genève (2010) ; Auguié et Vialle (2017, pp.127-128) ; BCI (2014, p.21) ; BnF (2021, pp.59-61, 74-75) ; BnF (2015, p.8) ; Claerr et Westeel (2011, p.175, pp.189-190) ; CREPUQ (2013, p. 22) ; DGSI (2015, pp.4-6) ; Essevaz-Roulet (2016, pp.48-49) ; Gueit-Montchal (2020, pp.329-334) ; Huc (2010) ; Makhlouf Shabou (2021, pp.9-10) ; MASSOT(2020) ; PARBICA (2021, pp. 20-23) ; PIAF (2015) ; SSGC (2019, p.2) ; Westeel (2010, p.112). ; Wikipedia.org.

Archives

Documents, quels que soient leur date, leur forme et leur support matériel, produits ou reçus par toute personne physique ou morale, et par tout service ou organisme public ou privé, dans l'exercice de leur activité. Le mot « archives » est couramment employé dans le sens restrictif de documents ayant fait l'objet d'un archivage, par opposition aux archives courantes.

Capture

Processus consistant à placer un document ou un objet numérique dans un système de gestion des documents d'activité et à lui affecter des métadonnées pour décrire le document d'activité et le placer dans son contexte, afin de permettre une gestion appropriée du document d'activité dans le temps. Enregistrement, classement, ajout de métadonnées et stockage d'un document dans un système d'archivage.

Code à barres

Séquence de barres fines ou épaisses détectées lors de la numérisation et permettant de générer une transition (nouveau document par exemple).

Coffre-fort numérique

Forme spécifique d'espace de stockage numérique, dont l'accès est limité à son seul utilisateur et aux personnes physiques spécialement mandatées par ce dernier.

Copie numérique fiable

Copie qui a la même force probante que l'original et dont la fiabilité est laissée à l'appréciation du juge. Néanmoins, est réputée fiable la copie exécutoire ou authentique d'un écrit authentique. Est présumée fiable jusqu'à preuve du contraire, toute copie résultant d'une reproduction à l'identique de la forme et du contenu de l'acte, et dont l'intégrité est garantie dans le temps par un procédé conforme à des conditions fixées par décret en Conseil d'Etat (documentation de la numérisation, prise d'empreintes, traçabilité des transferts de support ou de format, conservation dans des conditions de sécurité appropriées).

Cote

Ensembles de symboles (lettres, chiffres, signes) identifiant chaque article d'un service d'archives et correspondant à sa place dans le cadre de classement ou à son adresse dans les magasins.

Dématérialisation

Transformation de documents papier en traitement numérique, soit par le biais d'une opération de numérisation, soit par la révision des processus de production et de gestion de l'information.

Document numérique

Ensemble constitué d'un contenu, d'une structure logique (qui intègre la signature électronique éventuelle) et d'attributs de présentation permettant de restituer une mise en forme intelligible par l'être humain ou lisible par une machine. Le document électronique peut être créé à l'état natif ou obtenu par un processus de transformation d'un document physique, par exemple par numérisation.

Données à caractères personnelles

Toute information se rapportant à une personne physique identifiée ou identifiable, c'est-à-dire qui peut être identifiée, directement ou indirectement, par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale.

Dépôt

Acte juridique confiant la conservation de documents ou de fonds d'archives à la garde d'un service d'archives à titre temporaire et révocable sans transfert de propriété. Par extension, l'ensemble des documents déposés.

Dépôt provisoire

Il permet au propriétaire de déposer pour une période convenue entre les parties, des archives dans un cadre strict de traitement à des fins d'enseignement, de recherche ou de mandat. Les archives restent la propriété du déposant, ce dernier définit le transfert des responsabilités pour la période de traitement entre les parties. Un dépôt provisoire peur impliquer une rétrocession des archives.

Don

Remise de documents ou d'objets à un service d'archives à titre gratuit

DTD

En XML, « définition de type de document » (en anglais document type définition). Ensemble de règles qui déterminent quels sont les éléments que l'on peut employer dans un document XML et où l'on peut les employer. La DTD détermine donc le modèle de document, ou la structure de celui-ci.

Dublin Core

Normalisé en 2003 par l'ISO (15836), le schéma Dublin Core est devenu la base de protocoles d'interrogation permettant l'interopérabilité des métadonnées descriptives issues de ressources documentaires diverses. Toutefois, cette norme internationale ne donne pas d'indications sur la manière de remplir les différents champs, mais conseille pour certains cas d'utiliser des normes ou des listes d'autorités, c'est-à-dire les schémas d'encodage. Le schéma Dublin Core « est constitué de quinze éléments, optionnels ou non, répétables, permettant une description formelle, intellectuelle et juridique du document. » (Tessier, 2003, p. 54).

EAD (Encoded archival description)

Cette DTD est un modèle pour la production en XML d'instruments de recherche archivistiques conformes à la norme internationale ISAD(G).

FTP (Protocole de transfert de fichiers)

Mode de transfert de fichiers utilisant un réseau TCP/IP. Il permet de copier et même de gérer des fichiers d'un ordinateur à un autre. Il s'insère dans un modèle de client-serveur.

SFTP (Protocole sécurisé de transfert de fichiers)

Mode de transfert de fichiers qui permet de vérifier l'identité du serveur et aussi de cryptographier les données.

Fonds d'archives

Ensemble des documents de toute nature qu'une personne physique ou morale a automatiquement produits ou reçus dans l'exercice de ses activités, rassemblés et organisés en conséquence de celle-ci, et conservés en vue d'une utilisation éventuelle. Un fonds d'archive est dit fermé ou clos lorsqu'il n'est plus susceptible d'accroissement, et il est dit ouvert dans le cas contraire.

Gestion électronique de documents (GED)

Système informatisé d'acquisition, de classement, de gestion et de stockage des documents à partir d'applications informatiques dans le cadre normal des activités de l'institution. Les buts principaux d'une GED sont le gain de temps dans la recherche et la diffusion d'une information, ainsi que l'économie de papier en termes de stockage.

Indexation

Opération destinée à représenter par des éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document du document lui-même. Etape (dématérialisation) ou fonction (GED) automatique ou manuelle permettant de renseigner les index, attributs ou propriétés d'un document.

Intercalaire ou séparateur

Page papier spécifique insérée avant ou entre les différents courriers ou documents à numériser contenant un code à barres et constituant une identification de lot, un séparateur de document.

Interopérabilité

Capacité d'une application informatique, d'un système ou d'un schéma de métadonnées à communiquer, fonctionner ou s'interfacer avec un autre.

Inventaire

Instrument de recherche ayant pour niveau de description le dossier ou le document, et donnant pour chacun sa cote, ses dates extrêmes et une analyse (énumération descriptive plus ou moins détaillée).

ISAD(G)

Norme générale et internationale de description archivistique. Elle annonce les grands objectifs et principes de la description archivistique et fournit une liste des éléments de description que l'on peut employer dans un instrument de recherche, avec la définition de ces éléments et des exemples d'utilisation (voir annexe 6).

JPEG (*Joint Photographic Experts Group ; extension:* .jpeg, .jpg ou .jpe.)

Format de fichier image publié mais protégé par des brevets, largement utilisé dans la diffusion d'images sur Internet. Il est défini par la norme ISO /IEC IS 10918-1. Gestion des couleurs : RVB (jusqu'à 24 bits) et Niveaux de gris (8 bits). Ne gère pas les arrière-plans transparents de l'image (impossibilité d'ajouter des arrière-plans transparents). Métadonnées : EXIF, IPTC, XMP. Le format JPEG peut subir une compression avec une perte relative de qualité sans conséquence grave sur la diffusion de l'image.

JPEG2000 (*Joint Photographic Experts Group ;* extension: .jp2)

Format de fichier image publié et ouvert, utilisé pour diffuser des images de résolution variable ("tuiles"). Il est défini par la norme ISO/CEI 15444-1. Gestion des couleurs : jusqu'à 32 bits et supporte plusieurs espaces de couleur et plusieurs profils ICC. Métadonnées : XMP. La Compression JPEG2000 peut être sans perte (lossless) ou avec perte (lossy), mais avec possibilité de définir le taux de compression lors de l'enregistrement.

LAD (lecture automatique de documents)

Ensemble de technologies qui permet de segmenter et d'extraire, par reconnaissance optique de caractères (OCR, ICR et OMR), des informations sur des documents numérisés de type formulaires structurés ou semi-structurés. Les informations ainsi extraites peuvent alors être utilisées comme métadonnées dans un système de GED.

Legs

Disposition de droits ou de biens faite par testament attribuant à un organisme des archives, des objets ou une collection.

Liasse

Unité de conservation matérielle d'archives formée par un ensemble de documents conditionnés ensemble.

Lot

Regroupement physique de documents. C'est avant tout une structure « logique » utile au niveau des opérations de dématérialisation et pour le rapprochement éventuel avec les archives physiques.

Métadonnées

Ensemble structuré d'informations techniques, de gestion et de description attachées à un document servant à décrire les caractéristiques de ce document en vue de faciliter son repérage, sa gestion, son usage ou sa préservation.

Mètre linéaire

Unité de mesure des archives correspondant à la quantité de documents rangés sur une tablette d'un mettre de longueur.

Mire X-RITE

Mire existant dans différentes dimensions, composée de différents blocs colorés, qui permet de vérifier le rendu des couleurs. Certaines de ces mires existent en version « transparente », afin de réaliser des mesures similaires pour les équipements de numérisation de diapositives ou microformes.

Modèle de document

Il permet de caractériser (manuellement ou automatiquement) les documents et leurs traitements associés (reconnaissance, vidéocodage, contrôles via des référentiels).

Norme

Document établi par consensus et approuvé par un organisme reconnu, qui fournit, pour des usages communs et répétés, des règles, des lignes directrices ou des caractéristiques, pour des activités ou leurs résultats, garantissant un niveau d'ordre optimal dans un contexte donné.

Numérisation

Opération qui consiste à obtenir un artéfact numérique à partir d'un document physique. Cette opération fournit un document image que l'on peut traiter à l'aide d'outils informatiques. C'est la conversion d'un signal (vidéo, image, audio, caractère d'imprimerie, impulsion, etc.) en une suite de nombres permettant de représenter cet

objet en informatique ou en électronique numérique. On utilise parfois le terme "franglais" de digitalisation (digit signifiant chiffre en anglais).

PDF/A (*Portable Document Format/Archives*, extension : .pdf)

C'est une série de déclinaisons standardisées du format de publication PDF version 1.4, 1.7 et 2.0 normalisées ISO 19005-1. Ce format est utilisé pour une conservation à long terme. Il comporte plus de restrictions que le format PDF, mais permet une indépendance des fichiers par rapport aux plateformes sur lesquelles il est utilisé. Gestion des couleurs : les espaces de couleurs doivent être spécifiés de manière indépendante. Son développeur, l'entreprise Adobe, bien que disposant de brevets portant sur le format, accorde une licence gratuite pour la visualisation et l'édition de ces fichiers par des logiciels tiers (CPO-LIB). Il existe quatre versions du format PDF/A :

Date	Version PDF/A	Version PDF de base	Norme ISO
2005	PDF/A-1	PDF 1.4	ISO 19005-1
2011	PDF/A-2	PDF 1.7 (ISO 32000-1)	ISO 19000-2
2012	PDF/A-3	PDF 1.7 (ISO 32000-1)	ISO 19000-3
A venir	PDF/A-4	PDF 2.0 (ISO 32000-2)	ISO 19000-4 en cours

Partenariat

Il permet une relation privilégiée, basée sur la confiance entre les deux parties ; il implique un échange de bons procédés sur une période longue : plusieurs prestations proposées par le laboratoire et effectuées par les étudiant-e-s tout au long de leur scolarité, inclus la notion de prestation globale. Le partenariat implique le partage des connaissances entre les deux parties : mutualisation des ressources et des résultats produits (ex : partenariat avec des archives cantonales, communales ou institutionnelles).

PNG (*Portable Network Graphics, extension : png)*

Format de fichier image ouvert, flexible, robuste, indépendant des plateformes et polyvalent (affichage sur écrans, images web, logos, graphiques, illustrations). Il est standardisé par le W3C et défini par la norme ISO/IEC 15948. Gestion des couleurs : RVB (jusqu'à 48 bits), Niveaux de gris (8 ou 16 bits). Il permet de gérer la transparence d'une image (possibilité d'ajouter un "fond" derrière l'image sans que celle-ci ne soit masquée). Métadonnées : EXIF, IPTC, XMP. . Le PNG peut être compressé sans subir aucune perte de qualité.

Profil ICC

Fichier numérique d'un format particulier (extensions .icc ou .icm) décrivant la manière dont un périphérique informatique restitue les couleurs. Il a été créé par l'*International Color Consortium (ICC)* afin de permettre aux professionnels de la publication assistée par ordinateur de maîtriser la gestion de la couleur tout au long de la chaîne graphique. Un fichier ICC contient les données permettant de convertir les couleurs depuis un espace colorimétrique source (généralement lié à un périphérique) vers un espace colorimétrique indépendant (L*a*b* ou XYZ). Lorsque le périphérique informatique est un écran, un ordinateur ou un appareil photographique numérique, le profil ICC est un profil RVB (rouge, vert, bleu). Lorsque le périphérique informatique est une imprimante, un traceur ou une presse offset, le profil ICC devient un profil CMJN (cyan, magenta, jaune, noir).

Pratiquement, le profil ICC permet d'identifier les couleurs atteintes par le périphérique et la manière dont elles sont atteintes, ce qui permet d'optimiser et d'harmoniser les rendus colorimétriques d'une chaîne graphique (entre scanner, écran, traceur et presse). Le profil ICC contient de tables de correspondances entre les valeurs numériques et les couleurs réelles correspondantes mesurées en L*a*b*.

RAW

Premier fichier image (Raw veut dire "brut") obtenu après scannage qui contient toutes les données brutes du capteur, ainsi que les paramètres nécessaires à sa transformation en fichier image visible sur écran. Pour l'utiliser, il faut d'abord le convertir en autre format. Ce format est propriétaire et varie selon les constructeurs dans sa structure et dans son contenu.

Reconnaissance automatique de documents (RAD)

Technique permettant de distinguer un type de document d'un autre à partir d'une image du document. Cette identification permet de mettre en place des tris électroniques d'images afin de les classer, évitant ainsi de trier les documents avant la numérisation. Une fois regroupées, les images peuvent être envoyées vers des corbeilles de traitement adapté ou bien classées dans un système de GED.

Récolement

Vérification systématique, lors de la prise en charge d'un service d'archives ou à date fixe, de ses fonds et collections, consistant à dresser dans l'ordre des magasins et des rayonnages la liste des articles qui y sont conservés ou qui manquent par rapport aux instruments de recherche existants.

Reconnaissance optique de caractères (ROC, en anglais OCR)

Opération informatique qui consiste à analyser le contenu d'une image pour y repérer des formes qui correspondent à des lettres, et ainsi reconstituer le texte que l'on trouve dans l'image. Cette opération est habituellement effectuée après une opération de numérisation et permet de gérer l'information textuelle avec des outils plus performants que si l'on avait qu'une image du document.

Rétrocession

Dans le cadre d'un dépôt provisoire, ou d'un partenariat, les archives seront rétrocédées selon la convention préalablement signée par les parties, les frais de matériel liés au conditionnement seront facturés.

Signature électronique

Procédé cryptographique permettant de garantir l'intégrité d'un document ou d'un message électronique et l'identité de son signataire. Pour cela, les conditions suivantes doivent être réunies : authentique, infalsifiable, non réutilisable, inaltérable, irrévocable.

TIFF (*Tagged Image File Format*, extension: .tiff, .tif)

Format de fichier image source généralement utilisé par les logiciels de capture pour la numérisation des documents (avant conversion vers les formats de diffusion ou conservation tels que PDF/A), l'archivage à long terme, l'impression et l'édition en raison de la haute qualité d'images. C'est un format conteneur, propriétaire (société Adobe) et publié (pas de licence d'utilisation) défini par la norme ISO 12639. Il est extensible et capable d'embarquer des métadonnées internes dans différents formats (EXIF, IPTC, XMP). Gestion des couleurs étendue : jusqu'à 24 bits, supporte plusieurs espaces de couleur et plusieurs profils ICC. Le format TIFF est doté de différents algorithmes de compression sans perte de qualité dont LZW et CCITT groupe 4.

Typage

Étape automatique (RAD : reconnaissance automatique de documents) ou manuelle permettant d'associer un modèle à un document. Chez certains logiciels, cette étape est appelée « classification. »

Valeur probante

Qualité des documents d'archives qui leur permet de servir de preuve.

Vidéocodage

Fonction permettant de saisir manuellement les index qui n'ont pas été lus correctement par l'OCR.

Workflow ou flux de travail

Automatisation de tout ou d'une partie d'un processus de travail, au cours duquel les documents, l'information ou les tâches sont transmis d'un participant à l'autre, pour action, en application des procédures préétablies.

XML (*Extensible markup language*)

Norme qui permet de structurer de l'information de manière hiérarchique en imbriquant des éléments. XML permet de définir des formats (les DTD) que devront respecter les documents qui seront créés. XML est aujourd'hui utilisée dans toutes les sphères de l'informatique et est omniprésente en informatique documentaire.

XML-ALTO

ALTO (*Analyzed Layout and Text Object*) est un schéma XML standardisé, qui permet de stocker les informations relatives à la structure physique et au texte extrait par OCR (*Optical Character Recognition* : reconnaissance optique de caractères) d'une page d'un document numérisé. Très adapté à la conservation à long terme de ces données, ALTO a été adopté par de nombreuses institutions (*Library of Congress*, Université de Harvard, les Bibliothèques nationales du Danemark, de la Finlande, de la France, de la Nouvelle-Zélande, des Pays-Bas, du Singapour, etc.) dans leur processus de conversion en mode texte de documents numérisés. Ces principales qualités sont : universalité, facilité de création, d'édition, d'archivage et compacité. De plus, il est possible de générer un fichier PDF à partir des images et des fichiers ALTO alors que l'inverse n'est pas possible.

Annexe 2 : Guide d'entretiens

Conception d'un service de numérisation pour ArchiLab

PARTIE I

Introduction : présentation rapide du projet de mandat et son contexte

1°) Mot de remerciements à ou aux personne(s) participant à l'entretien :
2°) Me présenter :
3°) Rappeler les objectifs du projet de mandat :

A. Réaliser un état des lieux :

- Identifier les bonnes pratiques et les principaux services en matière de numérisation d'archives papier ;
- Analyser les prestations des principaux services en Suisse (romande et alémanique).

B. Concevoir un portefeuille de services possibles pour le laboratoire ArchiLab :

- Proposer un flux de travail (workflow) complet pour chacun de ces services ;
- Établir une tarification (pricing) en fonction des différents services proposés.

4°) Rappeler l'objet de la séance d'entretien : en en apprendre plus sur :

- Vos pratiques de numérisation des documents : typologie des documents numérisés et leurs formats, services offerts, étapes de la chaîne de numérisation, matériel de numérisation (scanners et logiciels) et paramétrage ;
- Aspects juridiques et normatifs de la numérisation ;
- Différentes méthodes de facturation des prestations proposées ;
- Difficultés rencontrées en numérisation et solutions envisagées.

5°) Rappeler le déroulement de l'entretien : enregistrement et traitement confidentiel des données.

PARTIE II

Identification de l'entreprise et profil de la/du répondant-e

a) Nom de l'entreprise :

b) Date de création :

c) Zone géographique d'intervention :

d) Domaines ou branches d'activité principaux :

e) Différents services proposés par votre entreprise :

f) Nature de votre clientèle :

- Entreprises privées ?
- Entreprises publiques ?
- Dans quelle proportion ?

g) Nom et prénom de la/du répondant-e :

h) Fonction dans l'entreprise :

i) Années d'expériences dans l'entreprise :

- Années d'expérience en numérisation/dématérialisation :

PARTIE III
Documenter le processus de numérisation

3.1. Documents numérisés et formats

- Quels types de documents numérisez-vous ?
- Quels types de formats ?

3.2. Matériel de numérisation (scanners et logiciels)

a) Scanners utilisés :

- Type (marque) :
- Nombre :
- Dimensions (largeur/hauteur/profondeur) :
- Mode de numérisation (à plat, verticale, caméra planétaire ou appareil photo numérique) :
- Format de papier :
- Capacité du chargeur :
- Volume de numérisation par jour :
- Vitesse de numérisation (ppm) :
- Résolution optique du scanner (en dpi) :
- Numérisation en mode recto-verso : oui/non ?
- Autres caractéristiques importantes du scanner :

b) Logiciels de numérisation et d'identification de la typologie des documents

- Nom du logiciel :
- Est-il associé ou non au scanner ?
- Critères de choix d'un bon logiciel ?
- Formats offerts pour l'importation et l'exportation des données ?

c) Logiciels de reconnaissance optique de caractères (OCR)

- Nom du logiciel :
- Est-il associé ou non au scanner ?
- Critères de choix d'un bon logiciel ?
- Formats offerts pour l'importation et l'exportation des données ?

3.3. Lieu et objectifs de numérisation

a) Où s'effectue la numérisation de vos documents ?

- Chez le client ?
- Dans les locaux de votre entreprise ?

b) Parmi les objectifs suivants de numérisation, lesquels sont largement évoqués par vos clients ?

- Préserver les documents à risque :
- Sauvegarder les documents essentiels :
- Réduire la masse documentaire en substituant le format papier par le format numérique :
- Diffuser (valoriser) le document :
- Je ne sais pas :

3.4. Étapes du processus de numérisation

a) Quels sont vos paramètres de numérisation ?

- Résolution de sortie (en dpi ou ppp) :
- Profondeur de codage ou de couleurs (en bits)
- Mode image :

- Espace colorimétrique
- Formats de fichiers d'images numérisées (stockage, conservation, diffusion) :
- Type et niveau de compression :
- Supports de conservation d'images numérisées :

b) Quelles sont les étapes de votre workflow de numérisation ?

c) Méthode de contrôle de la qualité des images numérisées et problèmes fréquents rencontrés ?

d) Pratiquez-vous de la RAD (reconnaissance automatique de documents), c'est-à-dire typage ou classification des documents ?

- Non/Oui :
- Si oui, quelle approche utilisez-vous ? Graphique (formes, données précisément localisées, etc.), syntaxique (mots-clés, codes-à-barres, etc.) ou les deux ?

e) Après la numérisation, que font vos clients de leurs documents ?

- Destruction :
- Conservation :
- Je ne sais pas :

3.5. Reconnaissance optique des caractères (ROC ou OCR)

a) L'océrisation a-t-elle lieu :

- Pendant la numérisation ?
- Ou après numérisation ?

b) Quelles sont les principales étapes de votre océrisation ?

c) Quels sont les formats de sortie de l'OCR ?

d) Quels sont les problèmes fréquents rencontrés lors de l'océrisation ?

3.6. Description-indexation

a) Votre indexation est-elle manuelle ou automatique ?

b) Si automatique, quels outils logiciels utilisez-vous ?

c) Quels schémas de métadonnées utilisez-vous (XML, Dublin Core, …) ?

d) Quels types de métadonnées utilisez-vous ?

- Métadonnées descriptives :
- Métadonnées techniques :
- Métadonnées administratives :

3.7. Aspects juridiques et normatifs de la numérisation

a) A quels lois ou textes règlementaires êtes-vous soumis en tant que prestataire de numérisation des documents ?

b) Idem pour les normes de qualité et de sécurité ?

PARTIE IV
Méthodes de facturation des prestations de numérisation

4.1. Comment facturez-vous vos clients ?

a) Par page et en fonction du format et de la couleur du document ?

Qualité du document	Format	Prix
Noir & Blanc	A4	
	A3	
	A2	
	A1	
	A0	
Couleur	A4	
	A3	
	A2	
	A1	
	A0	

b) Ou avez-vous une autre méthode de facturation ?

4.2. Poids en volume des étapes de numérisation

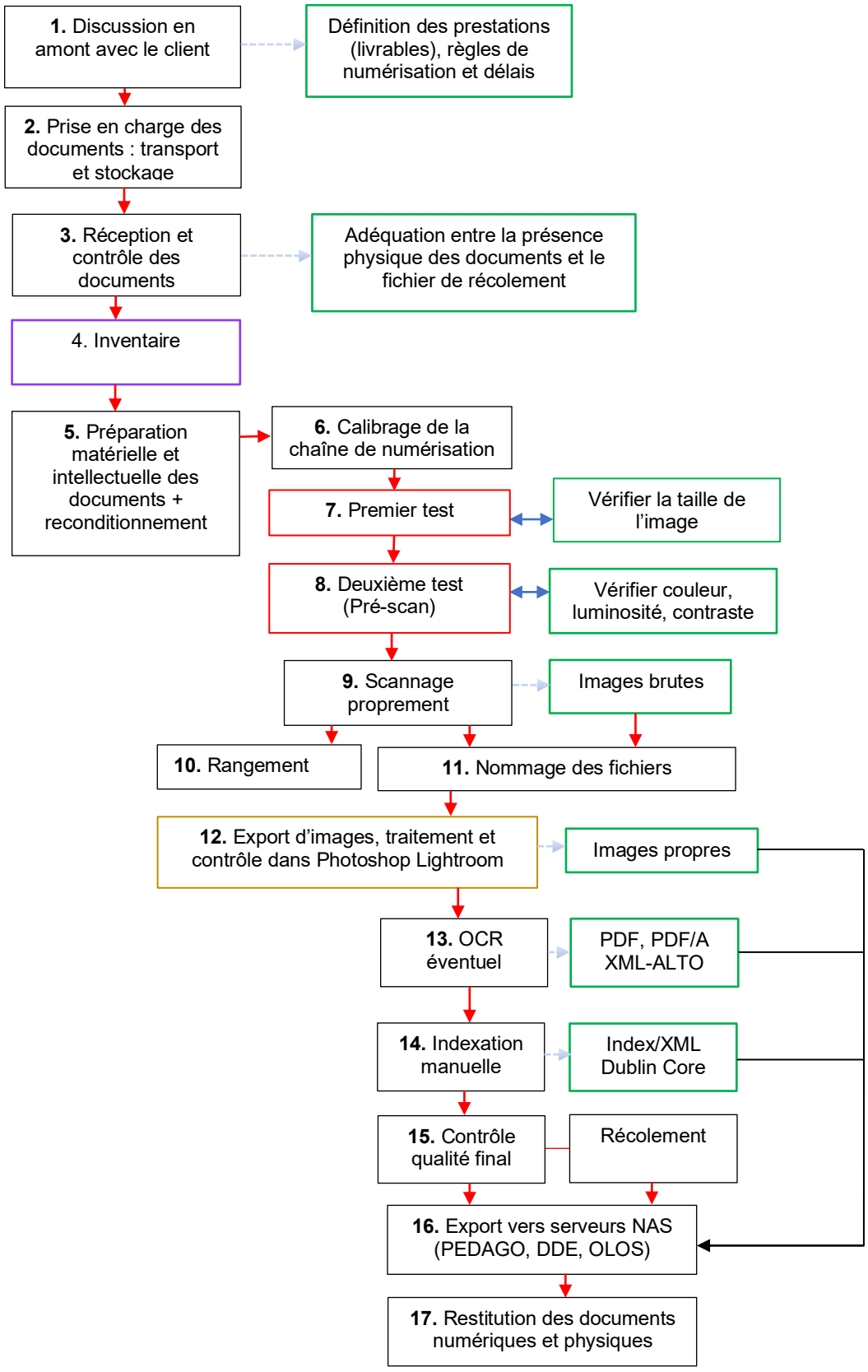
Quel est le temps nécessaire (en minutes) pour réaliser les tâches suivantes :

- Préparation matérielle et intellectuelle d'un document à numériser, format A4 :
- Préparation matérielle et intellectuelle d'un document à numériser, format A3 :
- Scannage d'une page au format A4 :
- Scannage d'une page au format A3 :
- Contrôle et traitement d'une image numérisée, format A4 :
- Contrôle et traitement d'une image numérisée, format A3 :
- Relecture et correction OCR d'une image numérisée, format A4 :
- Relecture et correction OCR d'une image numérisée, format A3 :
- Contrôle final d'un document :

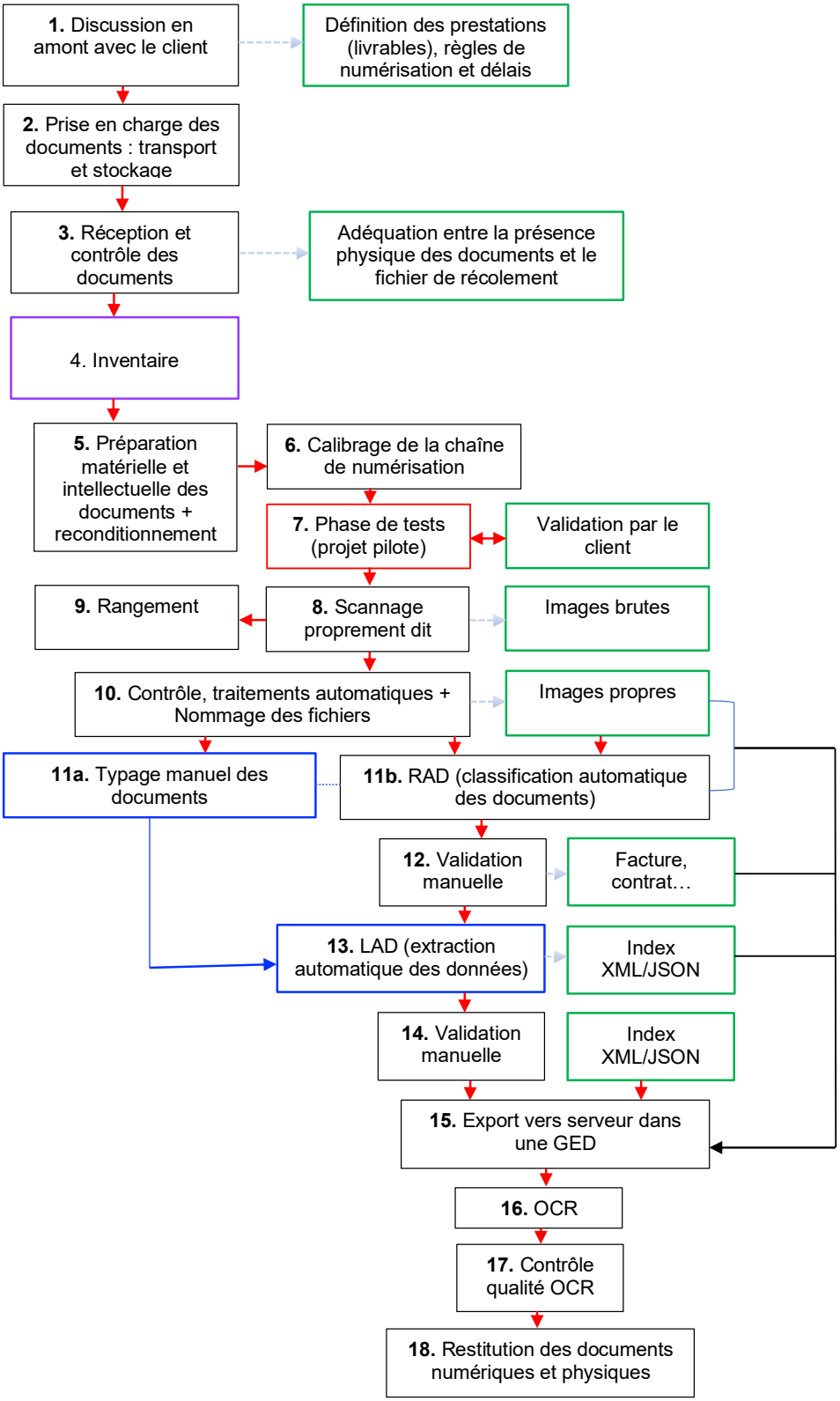
CLOTURE DE L'ENTRETIEN

1) Remerciements ;
2) Informations sur la suite du traitement des données des entretiens : envoi à la/au répondant-e d'un fichier de synthèse au format Word pour validation ;
3) Retour du fichier Word validé.

Annexe 3 : Proposition d'un workflow de numérisation de documents iconographiques



Annexe 4 : Proposition d'un workflow de numérisation mixte (patrimoniale et documentaire)



Annexe 5 : Prix indicatifs des prestations de numérisation appliqués par les entreprises interviewées

Pour des raisons de confidentialité demandée par les entreprises, le contenu de cette annexe n'est pas rendu public.

Annexe 6 : Niveaux et éléments de description obligatoires ou recommandés par la norme ISAD(G)

Éléments de description (Éléments obligatoires, éléments recommandés)		Niveaux de description				
		Service d'archives	Fonds	Série organique	Dossier	Document
1.	**Identification**					
1.1	Référence	o	o	R	o	o
1.2	Intitulé	o	o	o	o	o
1.3	Dates	o	o	R	o	o
1.4	Niveau de description	o	o	o	o	o
1.5	Importance matérielle et support (quantité, volume ou dimension)	o	o			
2	**Contexte**					
2.1	Nom du producteur	o	o			R
2.2	Histoire administrative/Notice biographique		R			
2.3	Historique de la conservation		R			
2.4	Modalités d'entrée					
3	**Contenu et structure**					
3.1	Présentation du contenu		R	R	R	R
3.2	Évaluation, tri et éliminations, sort final		R			
3.3	Accroissements					
3.4	Mode de classement		R	R	R	R
4	**Conditions d'accès et d'utilisation**					
4.1	Conditions d'accès	R	R	R	R	R
4.2	Conditions de reproduction					
4.3	Langue et écriture des documents					
4.4	Caractéristiques matérielles et contraintes techniques				R	R
4.5	Instruments de recherche	R	R			
5	**Sources complémentaires**					
5.1	Existence et lieu de conservation des originaux					
5.2	Existence et lieu de conservation des copies		R			R
5.3	Sources complémentaires		R			
5.4	Bibliographie	R				R*
6	**Notes**					
6.1	Notes					
7	**Contrôle de la description**					
7.1	Notes de l'archiviste		R			
7.2	Règles ou conventions		R			
7.3	Date(s) de la description		R			

*Uniquement recommandé pour les documents diplomatiques

(AAS, 2009, p.12)

Annexe 7 : Eléments du schéma de métadonnées Dublin Core

Libellé en anglais	Libellé en français	Description
Description intellectuelle		
Title	Titre du document	Il s'agit du titre principal du document
Subject	Sujet et mots-clés	Mots-clés, phrases de résumé, ou codes de classement. Il est préférable d'utiliser des mots-clés choisis dans le cadre d'une politique de classement.
Description	Description du document	Résumé, table des matières, ou texte libre.
Source	Ressource dont dérive le document	Le document peut découler en totalité ou en partie de la ressource en question. Il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI.
Language	Langue du document	Il est recommandé d'utiliser un code de langue conforme au format RFC4645.
Relation	Lien vers une ressource liée	Il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI.
Coverage	Portée du document	La portée inclut un domaine géographique, un laps de temps, ou une juridiction (nom d'une entité administrative). Il est recommandé d'utiliser des représentations normalisées de ces types de données, par exemple TGN (Thesaurus of Geographic Names, un dictionnaire de noms de lieux), ISO 3166, Point ou Box pour la portée spatiale, Period ou W3CDTF pour la portée temporelle.
Description matérielle		
Date	Date d'un événement dans le cycle de vie du document	Il peut s'agir par exemple de la date de création ou de la date de mise à disposition. Il est recommandé de spécifier la date au format W3CDTF (AAAA-MM-JJ).
Type	Nature ou genre du contenu	Grandes catégories de document. Il est recommandé d'utiliser des termes clairement définis au sein de son organisation. Par exemple, le Dublin Core définit quelques types dans le vocabulaire DCMITypes.
Format	Format du document	Format physique ou électronique du document. Par exemple, type de média ou dimensions (taille, durée). On peut spécifier le matériel et le logiciel nécessaire pour accéder au document. Il est recommandé d'utiliser des termes clairement définis, par exemple les types MIME.
Identifier	Identificateur non ambigu	Il est recommandé d'utiliser un système de référencement précis, par exemple les URI ou les numéros ISBN.
Description des aspects juridiques		
Creator	Créateur du document	Nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document.
Contributor	Contributeur au document	Nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document.
Publisher	Publicateur (éditeur) du document	Nom de la personne, de l'organisation ou du service à l'origine de la publication du document
Rights	Droits relatifs à la ressource	Permet de communiquer des informations sur le statut des droits du document, par exemple la présence d'un copyright, ou un lien vers le détenteur des droits. L'absence de cette propriété ne présume pas que le document est libre de droits.

(CREPUQ, 2013, p.17) et (TAVARES, 2013, p.45)

Annexe 8 : Contrôle qualité des images numérisées selon le plan d'échantillonnage de la norme ISO 2859-1

Les deux tableaux suivants permettent de définir la taille des échantillons à prélever sur les lots, le nombre acceptable d'objets défectueux dans l'échantillon et le nombre d'objets défectueux à partir duquel le lot sera refusé.

Détermination de la taille de l'échantillon en fonction de la taille du lot (extrait du tableau 1 de la norme)

Taille du lot	2 à 8	9 à 15	16 à 25	26 à 50	51 à 90	91 à 150	151 à 280	281 à 500	501 à 1200	1201 à 3200	3201 à 10'000	10'001 à 35'000	35'001 à 15'000	150'001 à 500'000	500'001 et plus
Taille de l'échantillon	2	3	5	8	13	20	32	50	80	125	200	315	500	800	1250

Détermination du nombre d'individus défectueux acceptables en fonction de la taille de l'échantillon et du NQA (extrait du tableau 2-A de la norme)

NQA	0.015%	0.025%	0.04%	0.065%	0.1%	0.15%	0.25%	0.4%	0.65%	1%	1.5%	2.5%	4%	6.5%	10%
2	-	-	-	-	-	-	-	-	-		-	-	-	0	-
3	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-
5	-	--	-	-	-	-	-	-	-	-	-	0	-	-	1
8	-	-	-	-	-	-	-	-	-	-	0	-	-	1	2
13	-	-	-	-	-	-	-	-	-	0	-	-	1	2	3
20	-	-	-	-	-	-	-	-	0	-	-	1	2	3	5
32	-	-	-	-	-	-	-	0	-	-	1	2	3	5	7
50	-	-	-	-	-	-	0	-	-	1	2	3	5	7	10
80	-	-	-	-	-	0	-	-	1	2	3	5	7	10	14
125	-	-	-	-	0	-	-	1	2	3	5	7	10	14	21
200	-	-	-	0	-	-	1	2	3	5	7	10	14	21	-
315	-	-	0	-	-	1	2	3	5	7	10	14	21	-	-
500	-	0	-	-	1	2	3	5	7	10	14	21		-	-
800	0	-	-	1	2	3	5	7	10	14	21	-	-	-	-
1250	-	1	2	3	5	7	10	14	21	-	-	-	-	-	-

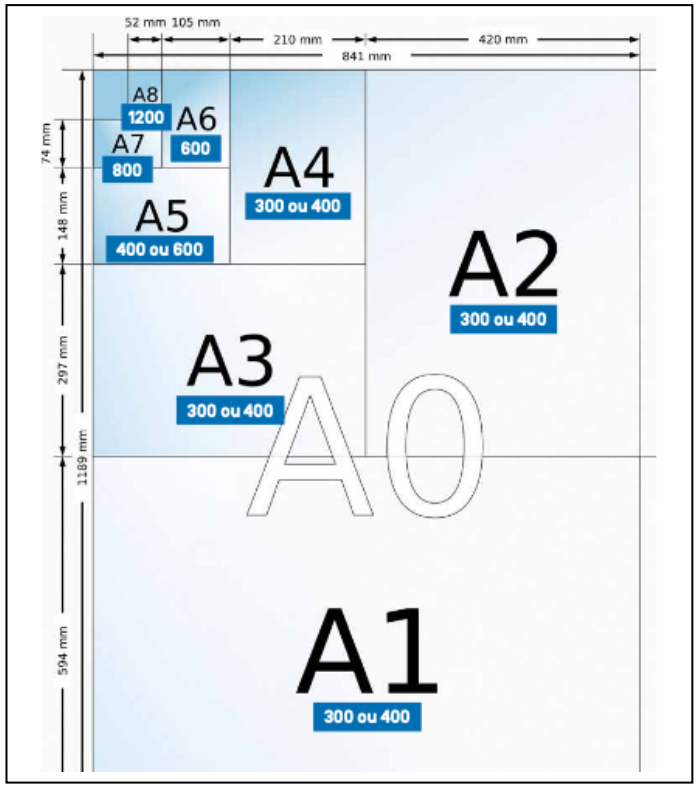
(DAF, 2008, p. 48)

Exemple de lecture des tableaux : Supposons un lot de 3000 images numérisées et que le client et le prestataire s'accordent sur un NQA de 0.65%. Dans ces conditions, la taille de l'échantillon à contrôler est de 125 images, car 3000 se trouve dans l'intervalle 1201-3200 du premier tableau. Le nombre d'images défectueuses autorisées dans cet échantillon (**2 images**) se lit dans le deuxième tableau, à l'intersection de la taille de l'échantillon (**125 images**) et du **NQA (0.65%)**. Autrement dit, le lot de 3000 images numérisées sera considéré comme conforme si le nombre d'images défectueuses est inférieur ou égal à 2.

Annexe 9 : Résolution optique recommandée en fonction du format A, selon la norme ISO 216

Format	Dimensions (mm)	Côté le plus long	
		Cm	po
A0	841 x 1189	118.9	$46^{3/8}$
A1	594 x 841	84.1	$33^{1/8}$
A2	420 x 594	59.4	$23^{3/8}$
A3	297 x 420	42.0	$16^{5/8}$
A4	210 x 297	29.7	$11^{7/8}$
A5	148 x 210	21.0	$8^{1/4}$
A6	105 x 148	14.8	$5^{3/4}$
A7	74 x 105	10.5	$4^{1/8}$

po: pouce (1 pouce = 2.54cm)



(BCI, 2014, p.119)

Annexe 10 : Projet de numérisation : liste des documents de suivi à fournir par le prestataire

Prise en charge des documents au départ, contrôle à réception des originaux	Accusé de réception mentionnant toute information permettant de tracer ce qui est réceptionné, par exemple : o Numéro identifiant le lot et les originaux. o Information sur l'état physique de chaque ouvrage par filière (nature du support) et type de traitement. o Date d'enlèvement et de réception. o Volume réceptionné (nombre d'ouvrages et/ou de caisses). o Liste des documents manquants par rapport au bordereau d'accompagnement. o Numéro du bon de commande. o Coordonnées de l'interlocuteur du commanditaire.
Récolement éventuel	Fichiers de récolement selon le formalisme donné au CCTP (Cahier des clauses techniques particulières).
Numérisation, OCR, contrôles internes	Rapports de production, voire de contrôles statistiques, etc.
Livraison des documents numériques	Bons de livraison des supports (disques) permettant le suivi de la production : o N° du bon de commande éventuel. o N° identifiant le lot et les originaux. o Filière et type de traitement fait. o Date de livraison. o Nombre de pages envoyées. o Si besoin, volume en GO. o Mention séparée des relivraisons à la suite de précédents rejets du commanditaire.
Retour des originaux	Bon de retour listant les caisses et les volumes/supports originaux.
Signalement des refus	Liste contenant les données d'identification et les motifs de refus des originaux.

(BnF, 2010, p.14)