Bibliothèque nationale de France
Bibliothèque municipale classée d'Orléans

Ecrire un cahier des charges de numérisation et de conversion en mode texte de collections de presse

Mars 2010

Ministère de la Culture et de la Communication
Comité de pilotage numérisation
Département de la Recherche de l'Enseignement supérieur et de la Technologie

Le guide *Écrire un cahier des charges de numérisation et de conversion en mode texte de collections de presse* a été rédigé par **Marie-Elise Fréon**, chef du service numérisation, Département de la conservation, Bibliothèque nationale de France et **Catherine Mocellin**, Directrice adjointe et chargée de la bibliothèque numérique, Bibliothèque municipale à vocation régionale d'Orléans. Nous les remercions d'avoir accepté de partager leur expérience professionnelle avec la diffusion de ce guide, destiné à tous les professionnels concernés par la numérisation de collections de presse.

Il a bénéficié de relectures de la part des représentants des directions du Ministère de la Culture et de la Communication dans le cadre des travaux du comité de pilotage numérisation :

- Jean-François Moufflet, Service interministériel des Archives de France
- Michel Jacobson, Service interministériel des Archives de France
- Claire Sibille-de Grimouärd, Service interministériel des Archives de France
- Edmond Fernandez, Archives nationales d'outre-mer
- Thierry Claerr, Service du Livre et de la Lecture
- Patricia Le Galèze, Service du Livre et de la Lecture
- Sonia Zillhardt, Département de la Recherche, de l'Enseignement supérieur et de la Technologie

Ce guide s'inscrit dans les travaux du comité de pilotage numérisation du Ministère de la Culture et de la Communication. Il vient compléter la collection des guides techniques lancée en 2008. Le premier guide *Écrire un cahier des charges de numérisation* (février 2008, Direction des archives de France) auquel il est fait référence dans ce document est celui concernant la numérisation de documents reliés, manuscrits, plans, dessins, photographies, microformes. Un guide spécifique sur la numérisation de documents sonores, audiovisuels et filmiques est en ligne depuis l'automne 2009. Est également en cours de préparation un guide sur la numérisation 3D.

SOMMAIRE

1 Introduction

1.1 Le contexte

Pour répondre à des besoins d'orientations techniques, ce guide s'adresse aux bibliothécaires, archivistes, responsables de la conservation et de la valorisation de collections de presse qui souhaitent rédiger un cahier des charges de numérisation de ces collections. La presse est ici entendue comme les titres de presse ancienne nationale, régionale, locale, libre de droits, à publication durable ou éphémère et pour un large public. Sont exclues de ce champ la presse récente et les publications périodiques destinées à la communauté scientifique.

La numérisation de la presse peut répondre à des besoins multiples :

- diffusion de contenus très utiles, pour un public vaste et varié, les chercheurs, ainsi qu'un public de curieux et d'amateurs trouvent dans la presse des sources d'information très précieuses, souvent difficilement accessibles. Il faut donc pouvoir répondre à la demande très forte de communication pesant sur les collections ;
- reconstitution virtuelle de collections réparties en plusieurs établissements : la numérisation et l'échange de métadonnées permettent de reconstituer pour un vaste public des collections lacunaires ou dispersées ;
- la presse est un support d'information d'actualités grand public ; sa numérisation permet aussi au public actuel de s'approprier davantage son histoire et son patrimoine national ou régional.

Ce guide s'appuie sur le document « Écrire un cahier des charges de numérisation – Guide technique » en ligne sur le site des Archives de France (rubriques « Gérer les archives » puis « Numérisation »).

Il fournit les points d'attention à prendre en compte en fonction des particularités de la presse. Les recommandations correspondent à une numérisation à des fins de communication, diffusion et sauvegarde à long terme des documents numérisés. Des orientations pour la conversion en mode texte par reconnaissance optique des caractères (OCR) sont également fournies : la presse étant un support au contenu très dense, divers (informations d'actualité, données économiques, feuilletons…), et peu structuré, les modes de recherche possibles diffèrent fortement des monographies ou des périodiques qui contiennent souvent une table de matières. Le mode texte permet de fournir à l'utilisateur des services de recherche et donc un moyen d'accès direct et immédiat aux contenus (par exemple : effectuer une recherche contextuelle sur les évènements, personnes, lieux particuliers…)

La presse est par ailleurs un support très fragile dont la numérisation répond à des besoins de conservation et de sauvegarde à long terme des originaux: du fait des techniques de fabrication, le support se dégrade rapidement et les grands formats sont difficilement manipulables surtout lorsque le papier est fin et cassant. Lorsqu'ils sont anciens, les fascicules de presse sont donc fréquemment incommunicables.

Les rubriques de document sont présentées dans l'ordre qui pourra être repris dans un futur CCTP, après une description des spécificités de la presse ayant un impact sur le montage du projet. Les éléments à mentionner dans le CCAP ne sont pas traités ici, les aspects génériques à tout projet de numérisation ne sont que brièvement développés : on se reportera au document des Archives de France précité.

1.2 Spécificités de la presse : identification et traitements préalables

1.2.1 Étude des collections de presse à numériser

On prendra en compte, notamment :

- l'intérêt et la portée des titres ;
- la demande des lecteurs ;
- la présence d'illustrations (photos, publicités…) ;
- la couverture par les droits d'auteur ;

- la complétude de la collection ;
- la disponibilité d'une collection dans plusieurs autres institutions ;
- les collections numérisées existantes et leurs modalités de consultation ;
- l'existence de supports de substitution : la presse a fréquemment été l'objet de campagnes de reproduction sur microfilms, et ce support peut être très utile pour la numérisation à la place de l'original. En fonction de ses contraintes propres, l'institution pourra choisir de numériser le papier d'origine ou le microfilm. Deux conditions minimales doivent toutefois être remplies pour la numérisation de films : traiter les matrices plutôt que les copies de consultation, et détenir la propriété sur ces matrices.

Outre la prestation de numérisation, on peut envisager d'inclure une prestation de conversion en mode texte par OCR.
La reconnaissance de caractères est un processus automatique qui permet de récupérer un flux textuel de plus ou moins bonne qualité selon l'état et la présentation de l'original. Pour des documents à la mise en page peu complexe, de présentation homogène et avec un papier en bon état (ni tâché ni bruni) les résultats peuvent être très bons. Malheureusement les collections de la plupart des institutions patrimoniales sont dans leur grande majorité dans un état tout autre et donnent des résultats de conversion, par OCR seul, très mitigés.

Selon l'utilisation que l'on souhaite faire du texte obtenu (indexation seule afin de permettre la recherche et/ou affichage du plein texte) le niveau de qualité requis pour la conversion pourra varier. Plus ce taux sera élevé (proche de 100%), plus il sera nécessaire d'intervenir manuellement (ressaisie des mots erronés) pour améliorer les résultats de l'OCR avec un impact important en terme de coûts de réalisation.
Les résultats obtenus en combinant processus automatique et correction manuelle peuvent être bons à condition que le titulaire du marché de numérisation sache évaluer correctement les mots reconnus et les mots douteux ou mal reconnus pour porter son effort de correction à bon escient. Une phase de calage et de tests au début de chaque projet de ce type est indispensable afin de s'assurer que les processus mis en place par le titulaire sont efficients.

Une conversion complètement manuelle est quant à elle complètement inadaptée aux larges volumes caractérisant la presse, d'autant que pour atteindre des niveaux de qualité élevés il faudrait procéder à une double saisie.

Quels que soient le titre et le support à traiter, si le projet comprend un volet d'OCR, on sera vigilant sur des critères de contenu spécifiques :
- la présence de caractères d'alphabets exotiques et non latins (caractères gothiques éventuels, idéogrammes, caractères arabes, grecs, etc.) ;
- la présence de zones manuscrites sur la majeure partie d'un fascicule ;
- une présentation particulièrement complexe des pages, qui tranche avec l'organisation courante des articles dans un fascicule de presse (à savoir plusieurs colonnes avec des illustrations, des encarts, des publicités) ;
- la présence de lacunes ou de déchirures affectant le texte ;
- la qualité d'impression et l'hétérogénéité de celle-ci au sein d'un fascicule d'une part, de la publication d'autre part.

Ces caractéristiques ont un impact sur la complexité de la conversion par OCR, mais ne sont pas bloquantes. On attirera donc la vigilance du prestataire sur ces éléments, en particulier sur le respect de l'ordre de lecture des blocs de texte. Ils seront également déterminants pour faire le choix du niveau de qualité exigé (voir § 6 : « Conformité et suivi de production »).

Pour le montage du marché incluant l'OCR, l'établissement peut soit demander des prestations séparées en lots (l'un pour la numérisation et l'autre pour l'OCR), soit faire les 2 prestations dans un seul lot.
Par ailleurs de nombreux fonds déjà numérisés peuvent constituer un gisement rétrospectif d'images susceptibles d'être converties par OCR. Si le volume concerné est faible, il sera préférable d'inclure cette prestation dans un marché de numérisation incluant de l'OCR. *A contrario* si le volume est important, l'établissement d'un lot de conversion rétrospective seule ou un marché spécifique seront plus adaptés.

1.2.2 Étude de la présentation physique des supports

Une expertise fine sur les supports de presse à traiter conditionne le montage de l'ensemble du projet. Leurs caractéristiques (présentation matérielle, tenue mécanique, qualité du microfilm, etc.) impliquent des précautions et des traitements spécifiques.

L'expertise doit être menée avant toute rédaction du CCTP : la présentation matérielle des documents doit être décrite de manière précise dans le cahier des charges, car elle détermine les choix et les contraintes du projet, parmi lesquels une consolidation préalable, les modalités de manipulation, les appareils et les réglages.
Selon les résultats de l'expertise, des moyens de l'établissement, de la politique de conservation et de communication des originaux, on choisira de numériser la collection telle qu'elle se présente, de la consolider avant ou après traitement, de faire appel à une collection d'une autre institution en meilleur état, ou à des microformes sous réserve de la qualité des *masters* et de leur disponibilité pour la numérisation.

1.2.2.1 Documents papier : présentation et traitements de consolidation

La fragilité de ce support est due tant aux matériaux de fabrication utilisés (papier, encre, puis matériaux de reliure) qu'aux conditions de conservation et de communication. On précisera donc au CCTP les conditions de manipulations par le prestataire.

- **Qualité du papier**

Pour les titres dont la durée de publication est longue, on signalera l'hétérogénéité de la présentation matérielle des fascicules pour l'ensemble de la période : évolution des formats, de la qualité du papier et de l'impression, accroissement du nombre de caractères moyen par pages au long de la période de publication.
L'hétérogénéité du papier et de l'impression peut caractériser un même objet relié ou un même fascicule.
Par ailleurs le papier est fréquemment acide et s'effrite ou se déchire lors des manipulations ; les fascicules peuvent être lacunaires, comprendre des articles découpés par les lecteurs, des rousseurs et mouillures…
Ces variations conduiront à des ajustements fréquents de la chaîne de production.
Une production de masse pour la presse ne pourra s'envisager qu'en fonction de la qualité du ~~papier~~ support, et de la prise en compte des moyens organisationnels et techniques de l'établissement.

- **Caractéristiques de la reliure ou du façonnage**

Selon la manière dont les fascicules ont été reliés puis conservés, la marge de petit fond des volumes est parfois trop étroite pour obtenir une image de qualité, et/ou le dos et/ou le papier peuvent être très fragiles (du fait de l'acidité, de l'encollage, ou de toute autre raison due aux matériaux utilisés, à la fréquence d'utilisation du titre et aux conditions de conservation). Il ne sera donc pas possible de numériser l'ouvrage en l'état sans l'abîmer : on veillera donc à dérelier l'ouvrage.

On pourra numériser les reliures présentant un intérêt particulier.

- **Préparation physique**

La manipulation des volumes est difficile, surtout lorsqu'ils sont de grand format et épais. Afin de ne pas dégrader le document lors de la numérisation et de fournir des images de qualité, il est recommandé de procéder à des traitements préalables, en particulier :

- débrochage, déreliage, en fonction de l'état de la reliure pour éviter de détériorer l'ouvrage lors des manipulations et obtenir une vue complète de chaque page ;
- dépoussiérage éventuel ;
- consolidation des pages abîmées (comblement de lacunes, repassage pour aplanir les plis, réparations de déchirures…).

Cette étape de préparation et de consolidation doit pouvoir être quantifiée en fonction des ressources humaines et matérielles nécessaires, afin de pouvoir établir des durées de traitement moyennes à intégrer au planning global du projet : il s'agit de veiller à ce que ces traitements et leurs imprévus éventuels n'occasionnent pas de rupture de charge pour le prestataire.

Selon la politique de communication des originaux numérisés, si ces opérations sont jugées nécessaires et qu'il n'est pas possible de les réaliser avant numérisation, on prévoira une consolidation après traitement. Cependant, il faut assumer qu'en présence de papier acide déjà fragilisé, le seul fait de tourner les pages des

volumes peut provoquer une dégradation de l'original qui de fait se verra aussi sur sa reproduction numérique.

Une sensibilisation particulière des prestataires est donc nécessaire sur la rareté et à la fragilité de la presse. Il faut veiller à fournir dans le CCTP des préconisations particulières sur les conditions de stockage des supports (locaux aux normes de conservation) et de manipulation par les opérateurs, pour lesquels on pourra prévoir une formation spécifique sur ce point. En effet toutes les sociétés n'appréhendent pas forcément l'ensemble des contraintes que pose la fragilité des fascicules de presse ancienne.

1.2.2.2 Microfilms de presse

Il est préférable de choisir de traiter des *masters* de très bonne qualité non communiqués aux lecteurs plutôt que des copies. Il est donc nécessaire d'effectuer une visualisation préalable par sondage des collections de films. Lors de cette étape, on prendra particulièrement garde aux points suivants :

- homogénéité de la densité sur une bobine ;
- longueur et solidité du film (on vérifiera la présence ou non de collages) ;
- qualité de l'émulsion ;
- type de reproduction : en noir et blanc ou en modelé continu ;
- état matériel du film : absence de poussières, de rayures ;
- caractéristiques du document reproduit : une reliure très serrée se traduit sur le film par une zone d'ombre pouvant masquer les premiers caractères de chaque ligne et induit des difficultés supplémentaires pour l'OCR. Si cette zone est large et sur de nombreuses pages, il faudra envisager de traiter l'original papier dérelié, s'il est disponible (mais les vues numériques n'auront pas un aspect homogène) ; dans le cas contraire, on veillera à signaler au lecteur que la numérisation est conforme à l'original ;

- propriété des films : elle doit appartenir à l'institution ; on veillera donc à identifier les modalités de reproduction avant de lancer le programme. On notera que les matrices des films reproduits par l'ACRPP sont en général la propriété de l'Association ; il est donc nécessaire d'obtenir son autorisation avant d'entreprendre leur numérisation.

Les scanners de microformes ne permettent pas à ce jour (2010) de numériser tous les formats originaux si la résolution dépasse 300 dpi, si l'on choisit un échantillonnage en niveaux de gris. Pour traiter des documents à une résolution supérieure à cette dernière, il faudra par exemple :
- se limiter dans les formats des originaux (le A3 représente une limite sûre)
- prévoir deux résolutions : 400 dpi jusqu'au A3 et 300 ou moins au-delà

En revanche, si l'on choisit un échantillonnage en noir et blanc, il demeure possible d'effectuer une numérisation de grands formats à 400 dpi ; il faut toutefois prendre en compte les résultats de l'OCR qui sont meilleurs lorsque la numérisation est effectuée en niveaux de gris.

1.2.3 Types d'accès à l'information

Après l'étude physique, on devra s'intéresser au contenu et aux types d'accès offerts aux lecteurs en ligne.

Il est indispensable de différencier les niveaux d'accès entre les volumes originaux et leur version numérisée : tout ce qui ne se présente pas en fascicule doit être analysé pour déterminer l'accès le plus approprié à une consultation en ligne, et donc les règles de numérisation de l'original. Par exemple un volume papier relié ou une bobine de film contiennent plusieurs parutions. Selon les titres de presse, il arrive souvent que ces originaux représentent une période de publication longue (semestre, année), et dans bien des cas plusieurs titres différents (reproduits sur une même bobine de film). Il n'est pas envisageable de conserver ce découpage pour la publication en ligne car il est bien plus difficile d'atteindre un numéro précis si ce volume est numérisé tel quel dans son intégralité : l'utilisateur est face à un grand nombre d'images numériques dans lesquelles il doit naviguer longuement avant de s'y retrouver.

Un niveau d'accès plus fin que l'original doit donc être prévu selon une périodicité suffisamment courte pour fournir des accès en ligne confortables : le jour ou la semaine par exemple.

À cet effet, on veillera à intégrer une phase de récolement des originaux par titre afin de fournir un récapitulatif des entités numériques à produire selon le niveau d'accès choisi. Le fichier de récolement devra contenir les éléments de répartition du contenu en document numérique, par exemple par date (année-mois-jour) permettant de reconstituer un affichage par calendrier annuel présentant l'ensemble des mois puis des jours (voir les détails en annexe), voire à l'article en fonction des moyens disponibles (voir ci-dessous).

Ce récolement peut être réalisé par l'institution ou par le prestataire. Dans tous les cas on fournira au CCTP l'ensemble des éléments bibliographiques à inclure dans le fichier de dépouillement qui constitue la source des métadonnées du document numérique.
Ces métadonnées permettront, en fonction de l'outil de gestion de la bibliothèque numérique, de fournir des services avancés aux lecteurs : calendrier, fonctions de kiosque (informations parues dans plusieurs titres pour un jour donné).
Cependant pour les parutions irrégulières, ou les changements fréquents de périodicité, il n'est pas toujours possible de proposer un accès par calendrier ; on peut alors fournir une liste des fascicules numérisés classés par ordre chronologique de parution.

Certaines institutions étrangères proposent un accès à l'article, très utile pour les lecteurs mais complexe à mettre en œuvre du fait de l'absence de table des matières dans la presse. Par ailleurs il est rare de pouvoir faire une indexation manuelle de ces articles, c'est pourquoi une conversion en mode texte peut être intéressante surtout lorsqu'elle est faite par des processus automatiques de reconnaissance, éventuellement améliorés par des corrections manuelles. Dans ce cas, il faut demander une qualité de conversion suffisante afin qu'un assez grand nombre de mots soit reconnus pour répondre aux requêtes des utilisateurs. Une structuration précise des titres d'articles peut aussi être demandée pour rendre exploitable un accès à l'article, à condition là aussi, de demander une qualité suffisante qui peut-être ciblée sur les titres d'articles (leur reconnaissance automatique pouvant être aléatoire à cause de leur typographie).

Choix des accès pour la presse régionale (variantes d'éditions locales)

Les éditions multiples d'une parution de quotidien, avec des variantes entre les informations générales et les informations locales propres à chaque édition et, selon le cas, des éditions particulières pour certains jours de la semaine voire pour des rubriques spéciales, sont à prendre en compte lors de la numérisation.

Lors de la reproduction sur microfilm de la presse régionale, des pilotes[1] ont souvent été constitués. On signalera au CCTP les caractéristiques des bobines, la structuration du contenu du fonds (années, thématiques, éditions différentes…) ainsi que la manière de restituer les documents numériques finaux. Les modalités de signalement des titres dans le catalogue devront être instruites par l'institution avant la numérisation, afin d'ajuster les métadonnées à créer pour ces cas complexes.

Plusieurs solutions de numérisation peuvent être envisagées :

- Numériser une seule fois les pages générales, puis dans chaque édition locale ne numériser que les pages qui lui sont propres.

  Avantages et inconvénients :

  - Numériser en supprimant les doublons d'informations générales peut s'avérer difficile et nécessite de garder un historique des parties numérisées afin de connaître ce qui a déjà été fait lorsque l'on reçoit une édition d'une même parution. Cela peut être ardu sur une collection rétrospective. La numérisation ne peut donc se faire sans suivi dans une base de production afin de savoir si une parution donnée a déjà été numérisée : en effet la règle est de numériser les parties générales du premier numéro reçu en numérisation, et de ne plus le faire sur les numéros des autres éditions qui peuvent arriver en traitement ultérieurement.
  - Assurer une numérisation exhaustive n'est pas toujours facile.

[1] Pilotes de reproduction qui reconstituent les différentes parutions d'une même date de publication en regroupant les parties générales et les différentes éditions locales d'un journal, évitant ainsi de reproduire plusieurs fois les pages générales communes à toutes les parutions.

- Le fac-similé numérique ne représente plus l'original dans son intégralité.
- L'application de consultation, lorsqu'elle donne la liste des éditions d'un titre donné, doit montrer clairement dans laquelle se trouvent les pages communes d'informations générales et, pour chaque édition locale, faire le lien vers le fascicule qui contient les pages générales. À défaut, la conversion en mode texte devient indispensable afin de permettre une requête qui pourra diriger l'utilisateur vers les pages concernées, qu'elles soient générales ou locales.
- Le principal avantage de ce procédé est d'éviter d'offrir des pages doublons au lecteur. Pour l'établissement, cette solution représente aussi une économie financière.

- Numériser chaque édition en l'état en générant ainsi des « doublons d'articles » pour les pages d'informations générales

  Avantages et inconvénients

  - Numériser toutes les éditions d'un titre en l'état permet une sauvegarde exhaustive d'un fascicule donné, et une reproduction exacte qui peuvent être utiles pour diverses utilisations ultérieures que l'on pourrait en faire.
  - Le processus de traitement s'avère plus facile tant en terme de production que de suivi.
  - Cependant il est nécessaire de prendre en compte le poids non négligeable des images produites en plusieurs exemplaires et d'évaluer l'économie de stockage générée en éliminant les informations redondantes ou de pallier les volumes produits en appliquant une compression, même minimum, sur les fichiers.
  - Par ailleurs les doublons ainsi présents peuvent parasiter la consultation, en particulier si le texte est converti, mais la forme numérique permet une navigation aisée pour passer rapidement les pages inutiles.

Recommandations
Si l'on numérise à partir d'un microfilm qui a suivi les pilotes de fabrication, donc en ne gardant qu'une fois les informations générales, il n'est pas intéressant de reconstituer les fascicules originaux. Il faudra donc suivre la même « architecture » lors de la numérisation. De plus, si des originaux complètent un titre microfilmé il faudra suivre le même procédé afin d'obtenir une homogénéité de présentation pour un titre donné.

Pour les établissements qui ne possèdent qu'une édition, il faut numériser chaque fascicule dans son intégralité à moins de se concerter avec les partenaires régionaux qui possèdent les autres éditions. Une politique régionale est alors à mettre en place non seulement pour la numérisation, mais aussi pour la consultation et la sauvegarde des fichiers.

2 Liste des prestations demandées

Cette liste est détaillée dans les chapitres suivants. Son objectif dans le CCTP est double : permettre à l'institution de lister l'ensemble des tâches, et au prestataire d'avoir rapidement une vue globale sur les opérations à réaliser.

En préambule, on indiquera le lieu précis du traitement : tout ou partie chez le commanditaire ou chez le prestataire. En fonction de différents points (espaces, volumes, coûts…), le cahier des charges peut en effet prévoir une numérisation sur place plutôt que dans les ateliers du prestataire.

Il s'agit de permettre aux candidats d'embrasser rapidement l'étendue du projet :
- Si besoin, tests préalables à la remise des offres sur une portion représentative du contenu à traiter, afin d'aider l'institution à départager les candidats (dans ce cas il faut fournir un test représentatif du titre ou des collections à traiter ; le même jeu de test doit être mis à disposition de tous les candidats).
- Visite et consultation des collections par les candidats dans les locaux de l'établissement, dans le délai de réponse suivant la publication du marché
- Éventuellement, récolement préalable – Cf chapitre 5.1.
- Numérisation et post production – Cf chapitre 5.2.
    - *contrôle à réception des documents ;*

    - *production d'images en niveau de gris et/ou en couleur des divers types de documents à numériser ;*
    - *compléments d'identification, et organisation des métadonnées ;*
    - *création des tables de correspondance pages physiques/pages électroniques ;*
    - *saisie des informations nécessaires à la création de l'exemplaire numérique.*
- Conversion par OCR (transcription du texte au format ISO 8859-1) de l'intégralité des images numérisées – Cf. Chapitre 5.3 et Annexe 2. Cette conversion comprend :
    - *contrôle à réception des images à convertir ;*
    - *la segmentation physique de la page afin de faire correspondre le texte issu de la conversion à l'image par transparence ;*
    - *la structuration du texte selon un schéma XML fourni ;*
    - *la préparation de l'ensemble des données (images et autres) en vue de leur livraison conformément aux type, structure et formats décrits dans le CCTP.*
- Liste des livrables à fournir par le prestataire – Cf. Chapitres 4.3. et 5.4.
- Contrôle qualité et fourniture des rapports correspondants - Cf. Chapitre 6.
- Livraison de documents numériques – Cf. Chapitre 7.
- Stockage dans les locaux du prestataire titulaire d'une sauvegarde de l'ensemble de ses prestations jusqu'à l'admission complète des prestations.
- Reprise gratuite des prestations rejetées.
- Éventuelles opérations préalables à charge de la société (pesage des caisses, par exemple).
- Le matériel éventuel à mettre à disposition du commanditaire sur son site (caisses et scellés, ou poste de contrôle calibré et régulièrement vérifié, etc.).

Les originaux remis par l'institution lui sont retournés dans leur conditionnement d'origine (boîte, pochette, caisse) selon le planning défini en commun avec le prestataire.

Le cahier des charges pourra demander aux candidats de proposer un planning de réalisation incluant les délais de retour des originaux. Ces éléments sont utiles pour l'évaluation des offres, dans la mesure où ils révèlent la compréhension du besoin et des contraintes de l'établissement par le candidat, ainsi que la qualité de l'organisation proposée.

3 Présentation et répartition des collections à numériser

Le Cahier des charges doit décrire les documents à numériser avant les prestations à effectuer.

Il s'agit de fournir une répartition par type de support, permettant au candidat de connaître l'étendue du projet, les collections à traiter, et de choisir le matériel le plus approprié. Il peut ainsi établir son offre en fonction des cadences de la chaîne de production à prévoir. Elles engagent en retour l'établissement sur les sélections (type, volumes).
La répartition peut être fournie en nombre de pages par type de numérisation à faire et/ou par taille (format ISO A4-A3, A2-A1…) pour un support donné, et comprendre une marge d'ajustement pour offrir au commanditaire une souplesse dans ses sélections.
Cette répartition doit être ajustée au budget dévolu au projet, et définie par l'institution selon des moyennes de prix pour des prestations similaires.

Filière	formats	%
Papier	≤ A4	90% - plus ou moins 5 %
	A3 jusqu'au A2	10% - plus ou moins 5 %
Microformes	35 mm	70%
	16mm	10%
	microfiche	20%

Il est possible aussi d'exprimer les volumes à traiter en nombre de pages :

Période	Format	Pages
1764-1828	A4 - un peu plus petit que A4	16 420
1829-1842	Entre A4 et A3 - plus proche du A3	5 970
1843-1852	Entre A3 et A2 - plus proche du A3	5 600
Total		27 990

Il est également possible de ne pas s'engager sur une répartition très précise par type de document, mais d'indiquer que le commanditaire souhaite numériser le plus d'ouvrages possible par support (papier, papier relié, microfilms) en fonction des prix qui lui seront proposés. Le candidat s'appuie alors sur le CCAP qui précise les montants minimum et maximum du marché.

Pour la numérisation, on distinguera deux filières en fonction du projet. Donner les volumes à titre ~~sont~~ indicatif et en pourcentage sur le nombre de pages à produire et pour la totalité du marché.

filière	**%**
Microformes	10 % - plus ou moins 5 %
Papier	90 % - plus ou moins 5 %

Cette répartition des supports doit inclure une description précise des caractéristiques essentielles des originaux à traiter, et signaler les points induisant une complexité dans la chaîne de production.

On pourra par exemple reprendre les éléments suivants, fournis à titre indicatif :

> *L'ensemble du programme concerne des fascicules et volumes de presse qui ont des caractéristiques physiques très différentes en termes de formats, de contraste (qualité d'impression, couleur de support variables) et de tenue mécanique.* ***Ces variations valent entre les documents d'un même lot d'une part, et au sein d'un même ouvrage, voire d'un même fascicule, d'autre part.*** *Le titulaire tiendra compte de cette variété dans la mise en œuvre de ses systèmes et processus de numérisation, et en particulier dans les réglages des scanners.*
>
> *L'impression, recto/verso, est de densité variable sur une page et d'une page à l'autre. Le recto et le verso des pages parasitent la lecture. Marginalement, quelques documents peuvent contenir en plus du texte des dessins et des photographies en noir et blanc ou en couleur.*

Les documents sont mis à disposition sous plusieurs formes :

- **Documents papier**

Fascicules débrochés (feuilles volantes), en pochettes, en boîtes ou rassemblés à l'intérieur de leur reliure d'origine.
Fascicules brochés ou volumes reliés (dans ce cas plusieurs fascicules sont assemblés dans un même volume).
Le format des pages est compris entre A4 et A0. Le cas échéant, quelques titres peuvent avoir des fascicules au format supérieur à A0, qui seront numérisés en plusieurs vues A0 ou inférieures.

- **Documents sur microfilm**

Les dimensions maximales des originaux reproduits sur les microfilms sont A3.

En règle générale, l'échelle centimétrique figure dans la vue ; elle sera utilisée pour appliquer le rapport d'agrandissement à la taille de la microvue, afin de ramener l'image finale à l'échelle 1. Dans le cas contraire, il faudra ajuster la résolution de manière à restituer tous les détails. (Cf. chapitre 5.2.3 : « Règles de numérisation »).

Il convient d'indiquer le rapport d'agrandissement entre la taille de la microvue et celle du document original reproduit et souhaité en sortie de numérisation.

Le tableau ci-dessous indique à titre indicatif la répartition prévisible des procédés de production à mettre en œuvre.

	type	Traitement à prévoir	Répartition par type
PAPIER 90 %	Niveaux de Gris		90 à 95%
	Niveau de gris + Couleur		5 à 10%
MICROFORMES(1) Microfilm35mm **10 %**	Noir&Blanc : 5 à 10 %	Traitement auto	66 %
		Traitement manuel (2)	44 %
	Niveaux de Gris : 90 à 95 %	Traitement auto	66 %
		Traitement manuel (2)	44 %

Note 1 : les divers types de microformes peuvent être d'aspect positif ou d'aspect négatif, ce dernier cas étant majoritaire[2].
Note 2 : certaines microformes ont des caractéristiques densitométriques irrégulières nécessitant des ajustements fréquents et précis des paramètres de numérisation. Le traitement sur numériseur automatique est à réserver aux microfilms de qualité constante[3].

Pour l'OCR : si le projet comprend un volet de conversion seule, on pourra fournir le nombre de pages à OCRiser.

Images numériques à OCRiser	1,5 million de pages

4 Mise à disposition des documents, bordereaux d'accompagnement

Il s'agit ici de préciser les modalités du transfert de responsabilité sur les documents, de transmission des résultats de contrôle et de validation administrative. Enfin, il est utile de lister les livrables fournis par chacune des parties.

4.1 Transfert des documents

Doivent être signalés les éléments suivants :

- **Conditions du transport**

Il s'agit de garantir les conditions de conservation normales des originaux et la protection contre les chocs.
Le transport doit être fait dans un camion respectant les conditions de conservation (température et hygrométries réglées de manière homogène sur l'ensemble de l'espace).
Dans le conditionnement (caisses scellées ou autre) les documents doivent être protégés et disposés de manière à ne pas pouvoir s'entrechoquer.
Si l'on ne dispose pas des conditionnements nécessaires au transport des documents, il faut demander leur fourniture au prestataire en le faisant clairement figurer dans le cahier des charges. Le type de conditionnements requis ainsi que leur nombre doivent être précisés (ou demander au prestataire de faire des propositions en fonction de sa capacité de production et des fréquences d'enlèvements/retours).

[2] Il convient d'être vigilant lorsque la numérisation des microfilms positifs est envisagée : en effet, leur numérisation ne donne pas de bons résultats à ce jour (2010) en matière de netteté.
[3] L'hétérogénéité de certaines microformes rend indispensable un traitement manuel afin d'obtenir des résultats exploitables. Ce traitement est onéreux et doit rester aussi exceptionnel que possible. Il vaut mieux tenter un traitement semi-automatique (il peut être utile de tester différentes solutions de traitement avec le prestataire afin de choisir le meilleur compromis).

- **Signalement et retour des originaux refusés**

Le titulaire peut refuser des volumes qu'il estime non numérisables dans le cadre du marché. Afin de permettre à l'institution d'expertiser et de suivre rapidement ces documents refusés, il faut demander leur retour dans une caisse à part et identifiée comme contenant les refus. La liste des refus doit également figurer dans les fichiers fournis par le titulaire au retour des lots

Pour mémoire, on précisera les points suivants au CCAP :

- **Assurances**

Le titulaire doit souscrire une police d'assurance suffisante le protégeant lors des transports et dans ses ateliers face à toute perte et dégradation des documents. Le début des prestations ne devra commencer que lorsque le titulaire du marché aura fourni les attestations suffisantes.
La valeur des ouvrages pouvant être très variable, il est préférable pour l'institution de ne pas fournir de valeur particulière afin de conserver une marge de manœuvre en cas de problème. La base d'indemnisation pourra se faire à partir de devis de restaurateurs ou de la valeur du document en antiquariat à la date du constat de perte. La couverture en valeur, assurée par le transporteur, doit être suffisante pour le nombre de documents à transporter.

- **Pénalités pour retard ou perte**

Il est indispensable de prévoir des pénalités financières pour dégradation et perte. Peuvent s'y ajouter des pénalités pour retard de livraison et/ou de retour des originaux.
Dans tous les cas, la manière d'appliquer et de calculer les pénalités sera précisée : il est préférable de choisir un montant à calculer en fonction du type d'accident (à partir par exemple du devis de restauration ou du nombre de jours de retard), plutôt que fournir un montant fixe qui peut se révéler inadéquat avec la gravité du problème.

4.2 Documents fournis par le commanditaire

4.2.1 Bordereau d'accompagnement

Un document (constitué à partir du fichier de récolement s'il a été fait au préalable par l'institution), doit accompagner les documents à l'aller et au retour. Il permet le transfert de responsabilité entre le commanditaire et le prestataire.

Il doit comporter au minimum :

- L'identifiant du lot.
- L'identifiant de chaque document.
- Le nombre de volumes/supports correspondants.
- La date d'enlèvement ou de retour des éléments.

Le prestataire est tenu de signaler toute non-conformité du lot avec le bordereau d'accompagnement dès la réception des documents (par exemple dans les 3 jours qui suivent l'enlèvement).

4.2.2 Fichier de récolement

Le fichier de récolement est fait par l'établissement, il doit être envoyé au prestataire soit en une seule fois au début du projet (si l'ensemble du récolement est terminé), soit au fur et à mesure des envois de documents. Le format du fichier et la description des champs utilisés doivent être décrits dans le cahier des charges ; on précisera également si certains éléments de ce fichier doivent être repris dans les métadonnées et dans quel champ cible.

On peut aussi récupérer dans le fichier de métadonnées des informations de description bibliographique ou fournir ces éléments séparément s'ils figurent déjà au catalogue et que l'on peut en fournir une extraction. Dans ce cas il faudra également décrire le format du fichier fourni.

4.3 Liste des documents de suivi à fournir par le prestataire

Les documents de suivi accompagnent la réalisation des prestations et permettent au commanditaire de suivre le déroulement du marché dans de bonnes conditions. Ce sont également autant de traces témoignant des relations avec le prestataire.

Les documents requis peuvent être listés dans le CCTP et présentés dans l'ordre de leur livraison au commanditaire selon la chaîne de production :

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 Informations 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.
Numérisation, OCR, contrôles internes	Rapports de production voire de contrôle, 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 re-livraisons suite à des 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.

Ces documents sont demandés au cahier des charges, affinés avec le prestataire lors du lancement, puis régulièrement validés par l'établissement. Le commanditaire précisera au CCTP leur forme (papier ou électronique et dans ce dernier cas, leur format et les modalités de leur transfert – mail, protocole ftp…) et la périodicité d'envoi.

5 Définition des prestations

Cette partie est à rédiger selon le montage du marché et la part d'opérations sous traitées en fonction des ressources de l'institution et de l'ampleur du projet.

Concernant l'allotissement :

- Le risque est d'avoir deux prestataires séparés pour chaque prestation, ce qui implique de transférer les images numérisées par le premier au second prestataire qui fait la conversion ; il en va de même si l'on sous-traite le contrôle. Ce transfert d'images est lourd à gérer, même s'il se fait directement entre les prestataires car il faut suivre l'exhaustivité et la qualité des données fournies au prestataire suivant dans la chaîne de traitement. Ceci implique d'être en mesure d'arbitrer les responsabilités à tout moment et peut rapidement bloquer la production. Les règles de fonctionnement entre les deux prestataires doivent donc être particulièrement précisées dans le cahier des charges.

- Faire un seul lot suppose de trouver sur le marché un prestataire capable d'assurer toutes les prestations. Cela est de plus en plus fréquent, mais il est aussi possible d'avoir un titulaire qui assure la maîtrise d'œuvre et se porte garant de l'ensemble des prestations.

Les prestations à réaliser seront présentées de manière exhaustive et précise, sans atteindre un niveau de détail tel qu'il pourrait bloquer certaines offres ou conduire les candidats à accroître les prix pour faire face à ce qu'ils estiment être une trop grande complexité.

Pour aider les candidats à comprendre le besoin du commanditaire il est préférable de présenter les prestations dans l'ordre de la chaîne de traitement.

5.1 Prestation de récolement

Cette opération, si elle n'a pas été réalisée en interne, doit figurer au cahier des charges dans les prestations à réaliser.
Les fascicules de presse étant souvent regroupés, en plus ou moins grand nombre, dans des volumes reliés en fonction de la périodicité du titre, il convient de définir un niveau de consultation minimal pour les lecteurs. Ceci implique de définir la granularité du document numérique à produire. Un récolement est en général nécessaire et doit indiquer les informations d'identification pour chaque unité documentaire ainsi que leur état physique. Ces informations pourront être reprises dans les fichiers de métadonnées des documents numériques (voir l'exemple fourni § 1.2.3, p. 7).

En fin de chaîne, ces éléments de récolement permettent également de faciliter les contrôles d'exhaustivité de la production, de conformité des métadonnées, voire de créer les exemplaires numériques dans le catalogue en fonction des outils de l'institution.

S'il est réalisé par le prestataire, on fournira les règles précises de saisie, en prenant en compte :
- le support : éventuelles variations de présentation entre le papier et le microfilm (par exemple : plusieurs titres reproduits sur une même bobine repérables grâce à la présence de pictogrammes Afnor, pages de couvertures rassemblées à la fin, etc.) ;
- les règles de définition de l'unité documentaire numérique finale en fonction de la granularité choisie, en prenant en compte les variations de pagination et/ou la manière dont l'original a été reproduit, en particulier pour les microfilms « pilotes » pour la presse régionale (voir p. 5). Ces règles comprendront les formats de date, de tomaison et volumaison, les abréviations…
- les données bibliographiques à obtenir par document numérique en fonction de la présentation du contenu sur le support (cas d'édition multiples, régionales, thématiques, suppléments, etc.), des besoins pour la consultation, mais aussi de la chaîne de contrôle et de mise en ligne ;
- les données d'état physique si les ouvrages sont déreliés avant numérisation (page manquante, taches, déchirures sur l'original…) ;
- toute autre information particulière utile au projet ou à l'information des lecteurs.

S'il est réalisé préalablement par l'institution, on pourra fournir un descriptif du fichier de récolement envoyé au prestataire, soit en présentant tous les champs, soit en présentant quelques uns seulement mais en annonçant la non exhaustivité.

5.2 Prestation de numérisation

L'ensemble de règles de numérisation et de prise de vue doit être détaillé précisément. Les précautions à prendre pour la manipulation de collections patrimoniales doivent être mentionnées, ainsi que les calibrages et les réglages des scanners, sans oublier l'obligation de mettre en place des processus de contrôle qui permettent au titulaire d'atteindre les niveaux de qualité demandés.
On peut lister globalement ces étapes ainsi :
- manipulation des documents papier de grands formats ou microformes sans dégradation physique ou chimique ;
- identification des images et des fichiers ;

- optimisation des paramètres de numérisation en fonction des caractéristiques physiques des documents ;
- calibrage des matériels de numérisation et de contrôle ;
- procédures de contrôle de qualité de production.

Les caractéristiques physiques doivent être présentées précisément pour chaque type de support, en veillant toutefois au bon degré de détail : il faut être précis et le plus exhaustif possible pour couvrir l'essentiel du fonds et mettre en évidence l'hétérogénéité des collections, mais sans atteindre le niveau du cas particulier pour se réserver une latitude dans les sélections.

Il est recommandé de toujours mentionner l'hétérogénéité du fonds.

5.2.1 Documents papier

Il faudra décrire :
- le conditionnement ;
- si les ouvrages auront été consolidés ou restaurés avant transfert au prestataire ;
- le type de document (relié ou en feuilles) ;
- la qualité du papier et les dégradations possibles (émiettement du à l'acidité, déchirures, etc.) ;
- l'existence de taches, pliures, gondoles, etc. ;
- les formats ;
- le degré d'ouverture maximal pour les objets reliés qui poseraient problème.

Il est recommandé également de préciser le niveau d'intervention que le titulaire est autorisé à faire sur les objets (décorner les pages, découronner les éventuels feuillets…).

5.2.2 Microfilms

Il s'agit de décrire le type de reproduction (noir et blanc ou niveau de gris), les formats (16mm, 35mm), la longueur maximale des bobines, la polarité.
Il est important de mentionner si les bobines ont des variations de densitométrie nécessitant des réglages particuliers et la distinction entre un traitement automatique et un traitement manuel.

Toute manipulation des microformes doit être faite avec des gants.

5.2.3 Règles de numérisation

Ces règles doivent être précises et prendre en compte l'éventuelle conversion par OCR qui suivra la numérisation : la qualité de l'OCR dépend dans une large mesure de la qualité de l'image numérique.

Ainsi pour obtenir une qualité de conversion texte maximale, il faut :
- redresser les images pour les rendre verticales (une faible tolérance d'inclinaison peut être admise) ;
- aplanir le plus possible les pages, si besoin avec une vitre en veillant à ne pas les détériorer ;
- éviter au maximum les courbures des lignes de texte ;
- appliquer à bon escient le mode de codage.

On fournira donc dans ce paragraphe les paramètres de numérisation suivants (*a minima*) :
- résolution exprimée en DPI, sans rééchantillonnage, pour un objet original restitué à son échelle exacte au moment de la numérisation ;
- dynamique des couleurs et profondeur d'acquisition : niveaux de gris (8 bits par pixel), couleur (24 bits par pixel) uniquement pour les pages en couleur. Le niveau de gris est préférable au noir et blanc pour la presse, car il permet de restituer la lisibilité du support par nature hétérogène : ainsi l'image numérique reste lisible même lorsque le contraste entre l'encre et le papier est faible, l'impression hétérogène, les pages tachées, légèrement pliées ou abîmées (numériser en noir en blanc, dans ce cas, provoque du bruit sur l'image numérique avec une multitude de points noirs sur le fond qui parasitent la lecture). Ce codage est également bien adapté aux photographies, courantes pour ce type de support. Enfin, il permet de mieux restituer l'esprit de l'original. Si toutefois on décide d'obtenir un

- cadrage, détourage : les images de fascicules de grand formats produites en 8 bits par pixel peuvent être très lourdes, aussi on peut admettre un léger détourage de l'image plein cadre (prise de vue avec les marges recadrée de quelques millimètres à l'intérieur) ;
- fourniture d'un format constant pour tous les fascicules de taille identique (recadrage permettant d'obtenir des pages de taille identique pour un même document numérique) ;
- orientation de l'image numérisée : soit dans le sens de l'original si l'on met à disposition des internautes dans le visualiseur un outil de pivot des images, soit directement dans le sens de la lecture. Si le texte converti doit être affiché sous l'image pour permettre la mise en surbrillance des mots recherchés, l'image devra être fournie dans les sens de la lecture afin de coïncider avec les coordonnées du texte qui auront été calculées avec l'image orientée dans le sens de la lecture lors du traitement OCR ;
- prise de vue ou non des pages spécifiques comme les pages blanches ou de couverture ;
- format des *masters* de livraison et version du format requis (par exemple TIFF V.6) ;
- compression éventuelle en fonction de la profondeur d'acquisition (noir et blanc, niveaux de gris, couleur), et dans ce cas format de compression à fournir ;
- en fonction des outils de consultation et de zoom, livraison de « tuiles »[4] de plus haute résolution, et caractéristiques techniques de ces fichiers.

Ces 3 derniers points sont détaillés ci-dessous au paragraphe 5.2.4 : « Métadonnées des fichiers images et OCR ».

Pour les microformes on ajoutera :

- dynamique des couleurs : on restituera le codage d'origine des vues du film (par exemple : restituer le noir et blanc tels quel si le film est en noir et blanc) ;
- échelle de réduction : présence ou non de cette échelle sur la microforme et modalités de traitement en cas d'absence ; on sera vigilant sur ce taux de réduction en fonction de la résolution choisie pour le document restitué à son échelle originale (1), afin de tenir compte des limites actuelles dans les capacités d'échantillonnage des scanners ;
- mode de traitement : automatique ou manuel ou semi-automatique ce dernier impliquant des réglages fréquents du scanner sur la bobine en cours de numérisation ;
- règles de traitement des vues particulières (logos, couvertures lorsqu'elles sont reproduites en fin de volume…) ;
- présence sur une même bobine de plusieurs titres, et/ou existence éventuelle de deux vues pour une même micro image, et manière de traiter ces cas ;
- règles de traitement des prises de vue doubles dues à des contraintes de production (montages, recollages…) : le prestataire pourra sélectionner lui-même la vue de meilleure qualité (à condition qu'il dispose de l'information concernant la numérotation).

5.2.4 Métadonnées des fichiers images et OCR

Le prestataire doit fournir un fichier d'identification du document numérique. Ce fichier a plusieurs objectifs :

- donner un minimum d'informations bibliographiques permettant d'identifier le document et de l'indexer (dont la numérotation des pages qui peut être reprise à partir du fichier de récolement) ;
- établir la correspondance entre les images numériques et les pages réelles du document physique ;
- fournir les métadonnées suffisantes permettant de conserver à long terme le document numérique ;

[4] Dans certains cas, il peut être intéressant de découper l'image en tuiles (partition d'une image par un ensemble fini d'éléments appelés tuiles). Il s'agit d'un découpage rectangulaire de l'image, découpage à spécifier, qui est généralement utilisé pour compresser des images de grande taille. Les tuiles sont alors un moyen de réduire la complexité mémoire (pour le codeur comme pour le décodeur), en travaillant sur des « sous-images » indépendantes. Par défaut, l'image entière est considérée comme une seule tuile.

L'établissement détaillera donc les éléments à produire par le titulaire et comment trouver l'information (dans le document et/ou dans des fichiers d'accompagnement fournis par le commanditaire). Les règles de saisie des données seront indiquées. Pour une meilleure compréhension par le prestataire, on fournira des exemples ou des schémas d'illustration.

Lorsque le fichier de métadonnées des fichiers images est correctement constitué, il permet de contrôler l'exhaustivité des traitements de conversion OCR sur toutes les images. Ce fichier peut également contenir le signalement de la présence du document en mode texte ; ainsi il gère à la fois le document en mode image et le document en mode texte.

5.2.5 Critères de refus des objets à numériser

Les documents doivent être fournis au prestataire dans un état permettant la numérisation.
Cependant une clause permettant au prestataire de refuser le traitement de certains volumes s'ils sont non conformes aux règles du CCTP et aux indications d'état physique figurant dans les bordereaux d'accompagnement permet de pallier les erreurs de sélection. La liste des cas de refus doit être énoncée et chaque ouvrage refusé doit être expertisé par le commanditaire afin de ne pas déséquilibrer le marché par une production insuffisante ou des dérives vers des refus abusifs, ni surcharger l'établissement en sélections de remplacement.

On demandera au prestataire de signaler les objets concernés, de les restituer à part et avec un motif précis de refus.

Exemples de défauts pour la presse nécessitant de statuer sur la faisabilité de la prestation de numérisation (liste non limitative) :

- pagination incohérente (exception faite de la simple inversion de deux pages). ;
- document incomplet ou présentant de graves lacunes ;
- papier trop acide pour être manipulé sans émiettement atteignant le texte ;
- anomalies graves de présentation ;
- marges étroites pour les ouvrages reliés et les microformes uniquement ;
- variations importantes de densité sur une même série de micro-images ;
- microforme(s) floue(s) ;
- microforme(s) aux caractéristiques densitométriques particulièrement inadaptées ;
- ouvrage contenant trop de pages maculées le rendant illisible.

5.3 Prestation de conversion en mode texte (océrisation)

5.3.1 Règles de conversion

Le cahier des charges précisera la structure des données à produire en fonction du format requis pour la livraison des données:

- structure des fichiers : un seul fichier contenant toutes les pages du document ou un fichier par page ;
- structure des données textuelles : fichier texte au kilomètre ou texte enrichis d'informations permettant des méthodes d'indexation ou de navigation plus sophistiquées : par exemple, repérage des premiers niveaux de titres ou des rubriques de la publication afin de générer une table des matières (qui sera consultable, par exemple, dans un fichier PDF).

Cette structure est très dépendante du format choisi mais doit être précisée dans les règles de conversion afin que le prestataire sache dès la saisie/conversion des données comment organiser l'information alors que cela ne sera pas toujours faisable au moment de la création du fichier final de livraison.

Le format texte permet de présenter le flux textuel converti sans mise en forme. On peut y introduire des identifiants de l'image correspondante afin de pouvoir construire ultérieurement un affichage conjoint. Cependant la superposition n'est pas possible, il faudrait pour cela avoir les coordonnées du texte dans

l'image. Il est ensuite difficile de construire une mise en page facilitant la consultation du contenu surtout sur des pages de presse à la présentation complexe.

Le format XML ALTO permet de stocker dans un seul fichier le contenu textuel, les coordonnées du texte dans l'image et des éléments d'informations complémentaires permettant des utilisations spécifiques des données. Il permet des réutilisations multiples du contenu ou de parties du contenu. Cependant il n'est pas possible d'obtenir un seul fichier ALTO contenant toutes les pages du document. Le format restitue le contenu des pages image par image, gardant ainsi un référence à sa source et permettant de les exploiter conjointement ou séparément. On veillera à demander le regroupement de tous les fichiers ALTO d'un document dans un même répertoire (cf. paragraphe 7.1 : « Structure de la livraison »). Voir en annexe des informations complémentaires sur ce format.

Le format PDF multicouche peut être un bon compromis, il permet de gérer un fichier multi-pages contenant à la fois le texte et l'image ou le texte seul mis en page et structuré à condition de faire figurer les règles de structuration au cahier des charges (repérage de la mise en page, des niveaux de titres, liens éventuels vers des suites d'articles, etc.)

L'inconvénient du fichier PDF réside dans la difficulté à récupérer le contenu textuel pour des utilisations futures, ce qui n'est pas le cas du fichier ALTO qui non seulement le permet, mais offre également les mêmes niveaux de structuration. De plus il permet la construction de différents types de fichiers de diffusion y compris le PDF multicouche. Enfin c'est un format pérenne.

On se reportera en annexe pour plus de détail sur les formats textes.

5.3.2 Conversion seule d'images numériques : typologie des images à traiter

Il convient de décrire les caractéristiques des images numériques de presse à OCRiser :

- support d'origine (microformes ou papier) et caractéristiques générales (variations de contraste, de densitométrie…) ;
- description des originaux : formats, présence d'illustrations, désignation des alphabets, qualité du papier, transparence, présence de tâches, mise en page en colonnes ;
- format, version ;
- résolution ;
- compression éventuelle, taux, algorithme ;
- mode de codage (niveaux de gris, noir et blanc, couleur) et variations possibles au sein d'un même document numérique ou d'un même lot ;
- règles générales de prise de vue (présence de pages blanches, de pages avec logos…).

Dans cette partie, on fournira également la structure des documents numériques transmis au prestataire, ainsi que les modalités pratiques :

- nommage et organisation des fichiers et des répertoires ;
- type de fichier de métadonnées jointes aux documents numériques à traiter avec, le cas échéant, le schéma XML utilisé qui doit être fourni en annexe ;
- caractéristiques des données présentes dans le fichier de métadonnées ;
- fréquence de fourniture des lots.

5.4 Création des fichiers numériques

5.4.1.1 Fichiers images

Il convient de préciser dans le cahier des charges les types de fichier que doit fournir le prestataire, en vue de l'archivage et si besoin en vue de la diffusion, selon les besoins propres à chaque établissement. Si l'on désire recevoir plusieurs types de fichiers on précisera les formats et version de chacun d'eux et éventuellement le taux de compression admis.

- **Fichiers d'archive**

Les pages de presse, souvent produites en grand format en particulier pour la presse quotidienne, peuvent générer des images lourdes. Produire un *master* sans compression en vue de la conservation est l'idéal, cependant cela impacte très fortement les moyens de stockage à mettre en place et ce d'autant plus que l'on aura numérisé en haute résolution et le cas échéant en niveaux de gris ou en couleurs.
Un compromis est à trouver en fonction des choix de résolution et de dynamique adaptés aux utilisations que l'on veut faire à plus ou moins long terme de la presse et le taux de compression admissible. Si la résolution permet des traitements que l'on pourrait faire dans un futur assez proche avec un résultat juste admissible (par exemple conversion par OCR) on évitera la compression afin de garder le plus de détails possible. En revanche si la résolution donne un niveau optimum pour des traitements ultérieurs on pourra se permettre une compression légère.

Sur les images en noir et blanc la compression IUT GrIV permet un gain de place très appréciable tout en garantissant un réversibilité complète à la décompression (tous les pixels supprimés peuvent être restitués à leur emplacement d'origine grâce à un codage prédictif utilisant des lignes de référence).

Pour les images en niveaux de gris ou en couleur le choix est devenu plus aisé grâce aux nouvelles méthodes de compression utilisées par l'algorithme JPEG2000 qui offre des capacités énormes de compression avec peu ou pas de perte.

Exemple de rapports de compression pour des fichiers en couleur

Format	TIFF (non compressé)	TIFF LZW	TIFF ZIP	PNG	JP 2000 (compression sans perte)
Taille en octets	10 759 617	7 076 277	6 343 873	5 659 929	4 893 189
Taille en Mo	10,26 Mo	6,7 Mo	6 Mo	5,4 Mo	4,6 Mo
Rapport qualité après compression	100%	66,7%	59,8%	53,3%	46,1%

Exemple de rapports de compression fichiers en niveaux de gris

Format	TIFF (non compressé)	TIFF LZW	TIFF ZIP	PNG	JP 2000 (compression sans perte)
Taille en octets	12 599 232	7 805 917	7 154 621	6 764 053	5 038 368
Taille en Mo	12 Mo	7,4 Mo	6,8 Mo	6,4 Mo	4,8 Mo
Rapport qualité après compression	100%	62%	56,8%	53,6%	40%

Voir, dans l'annexe, les recommandations sur l'utilisation de ces compressions.

- **Fichiers de diffusion**

Le cahier des charges précisera les formats et dimensions des fichiers de diffusion par exemple : fichiers au format vignette, fichiers basse résolution 72 dpi pour visualisation rapide, fichier de consultation haute définition compressé…

Les images de presse correspondant à des pages de grand format ne s'affichent pas en entier lorsqu'elles sont en haute définition, même compressées à un taux qui permet une lecture confortable. Il peut être intéressant de demander des images compressées grâce à l'algorithme de compression jpeg 2000 qui permet de diviser l'image en tuiles de différentes résolutions et est capable de donner un accès aléatoire à différentes parties de l'image (par exemple par colonne). Cependant le nombre de tuiles requises doit être évalué à bon escient afin d'éviter un poids accru de fichiers (de par leur nombre trop important) qui reviendrait au même qu'un fichier d'archivage non compressé. Le but est de faciliter la navigation dans le document et de rendre lisibles les textes en petite police. Il peut-être utile de faire des tests préalables avec

des fournisseurs de logiciels de traitement d'images et avec son système d'information avant de rédiger le cahier des charges.

Pour plus de détails sur les formats se reporter à l'annexe : « Formats d'images »

5.4.1.2 Fichiers texte

Comme pour les fichiers images il est possible d'obtenir :

- d'une part des fichiers d'archivage dans un format pérenne, si possible en xml (voir détails du format ALTO dans l'annexe : « Formats de fichiers en mode texte ») ou au format texte, mais ce dernier ne pourra contenir aucune structuration hormis des retours à la ligne marquant les paragraphes ni aucune coordonnées du texte dans l'image limitant ainsi les utilisations futures que l'on pourrait en faire sauf à mettre en œuvre des moyens de transformation ;
- d'autre part un format de diffusion, soit en PDF, soit en format texte, soit en html. Dans le premier cas, le texte et l'image peuvent être superposés permettant une recherche dans le texte, mais seule l'image est visible ; dans les deux autres cas le texte est visible sans l'image sauf si l'on fait des liens vers les fichiers images. Cependant le fichier texte devra être mis en page pour permettre la consultation sur Internet, c'est pourquoi il peut être intéressant de demander un fichier html prêt pour l'affichage.

Dans tous les cas, on précisera le codage des caractères (par exemple ISO 8859-1, UTF8 avec ou sans entités caractères…)

5.4.1.3 Fichiers de métadonnées

Le format du fichier à produire devra être décrit, s'il est en XML le schéma utilisé sera fourni. Le codage de caractères sera également précisé.

Pour plus de détails sur les formats de métadonnées se reporter au document « Écrire un cahier des charges de numérisation – Guide technique » cité en introduction et aux exemples fournis en annexe du présent document.

6 Conformité et suivi de production

Cette partie détaille la méthode suivie par le commanditaire pour contrôler et valider les prestations, en fonction de la qualité requise et exprimée dans la définition des prestations. Elle doit être particulièrement soignée car elle détermine les processus de contrôle que le titulaire mettra en place et la marge de manœuvre du commanditaire pour rejeter les livraisons.

6.1 Organisation générale et suivi du projet

On prévoira au minimum deux phases du marché, pour lesquelles les règles de contrôle et les modalités de reprise seront détaillés :

- **phase de test** (après la réunion de lancement) : en fonction du CCTP, et à partir d'un jeu de documents tests représentatifs du fonds à traiter, il s'agit pour le titulaire d'ajuster les chaînes de production, ses procédures, d'atteindre la qualité requise des images et de la conversion, de former ses opérateurs, de préciser le planning et les indicateurs de contrôle et de production. Cette phase à durée variable selon le projet et l'expérience du prestataire doit être limitée dans le temps. Elle s'achève par l'accord entre les deux parties sur les procédures, les réglages fins des chaînes et des appareils, et la rédaction d'un cahier de procédures (ou plan assurance qualité) décrivant l'ensemble des points de la prestation et incluant les chartes techniques. Durant cette phase, on pourra avoir recours à des contrôles exhaustifs sur des parties du lot test, et/ou des contrôles par échantillonnage à un niveau renforcé.

- **phase de production courante** : elle ne commence qu'à la validation de l'ensemble des tests (unitaires et globaux, *i.e.* par type de prestation au fur et à mesure que le prestataire adapte ses processus de production, puis pour l'ensemble des prestations fournies pour un document). Elle prévoit une montée en charge. Les contrôles mis en œuvre par le commanditaire sont détaillés pour chaque type de prestation.

Durant la phase de test, et en cas de difficulté récurrente en production courante, on pourra passer à un niveau de contrôle renforcé.

Le CCAP précisera l'articulation de ces contrôles avec la gestion administrative du marché : contrôles valant pour admission partielle, puis pour admission complète par exemple.

Veiller à la bonne gestion du projet par le titulaire est un moyen d'anticiper sur les dérives et de faciliter la réussite du marché. On détaillera alors dans le CCTP les modalités mises en œuvre pour s'assurer de ce suivi (audits, visites sur site, rédaction et respect de procédures dans les ateliers…) et les vérifications portant sur les livrables fournis régulièrement, par exemple :

- qualité et exhaustivité des rapports de production, de l'éventuel fichier de dépouillement à fournir ;
- exhaustivité de la livraison par rapport à la commande ;
- exhaustivité des retours d'objets originaux ;
- respect des délais.

6.2 Conformité de production

6.2.1 Modalités de contrôle

On prévoira des contrôles quantitatifs et qualitatifs. Plusieurs modalités existent et peuvent être cumulées :

- contrôle exhaustif automatique (par un logiciel, sur les données techniques et de structure) ;
- contrôle visuel avant mise en ligne (contrôle manuel par opérateur sur un échantillon);
- ou à réception des livraisons, contrôle uniquement de structure (conformité technique du document numérique), puis après mise en ligne, contrôle visuel par échantillonnage et en s'appuyant sur les remarques des internautes. Un délai administratif de validation du contrôle visuel est à prévoir au marché pour ce type d'organisation ;
- contrôle par audits des chaînes et procédures du prestataire : le CCTP précisera la fréquence, les apports demandés au prestataire, le contenu global des audits, les résultats (rapports, actions correctives à mettre en œuvre par le titulaire) et leur impact contractuel. Les suites données à chaque audit (actions correctives et leurs conséquences) doivent être vérifiées lors des audits suivants et des contrôles techniques et visuels effectués sur les livraisons. Chaque audit devra suivre la même méthodologie ;
- vérification de la pertinence des choix techniques par les indicateurs de production fournis régulièrement ;
- recours à un prestataire de contrôle (en remplacement ou en complément des contrôles décrits aux trois premiers points ci-dessus), avec les avantages et inconvénients suivants :

Avantages	Inconvénients
Fiabilité du contrôle et du respect de la norme d'échantillonnage (compétence de la société possédant les outils et la capacité de développement de logiciels nécessaires).	Fiabilité réelle seulement si l'institution a les moyens effectifs de contrôler le prestataire de contrôle et si des règles de production et de contrôle précises ont été rédigées et acceptées par toutes les parties.
Possibilité de traiter des masses plus importantes du fait de l'allègement de la charge de contrôle pour le commanditaire (les échantillons à contrôler sur les livraisons du prestataire de contrôle sont de petite taille).	Flux importants de documents à suivre (envoi au prestataire de contrôle les documents originaux numérisés, puis au prestataire de numérisation des ouvrages à retraiter suite aux rejets si ces derniers ont été validés par le commanditaire).

	- Lourdeur de gestion due à la complexité du montage du marché, du suivi de la production et des plannings, de l'arbitrage nécessaire en cas de désaccord sur les résultats de contrôle. - Délais de réalisation et de payement très fortement allongés. - Si les deux marchés sont réalisés en même temps, risques financiers et de production pour un prestataire en cas de difficultés pour l'autre prestataire ; la gestion du projet par l'institution est alors très difficile.

6.2.2 Contrôle et reprise des livraisons

Il est important de mentionner au CCTP que le titulaire est tenu d'assurer lui-même la qualité des prestations fournies : pour cela il mettra en place les outils et procédures lui permettant d'évaluer la qualité de ses prestations et produira les justificatifs permettant au commanditaire de donner son accord sur la validation ou le rejet des livraisons.
Par ailleurs, lorsque l'on demande plusieurs prestations (numérisation, métadonnées, conversion en mode texte…), la qualité de l'une est tributaire de celles des autres (par exemple le résultat de la conversion OCR est meilleure si l'image exploitée est de bonne qualité).

Chaque prestation demandée au marché devra être contrôlée. Comme il est peu fiable à long terme ni envisageable de faire un contrôle exhaustif par opérateur sur de gros volumes, on aura recours de préférence à l'échantillonnage. Les exigences de qualité devront donc être cohérentes avec les moyens de contrôle dont dispose l'institution, en particulier pour l'OCR en l'absence de logiciels de contrôle de l'OCR sur le marché.

Pour l'ensemble des contrôles, on précisera le lot (le document numérique, la livraison, un répertoire contenant plusieurs documents numériques, etc.), et en cas d'échantillonnage l'utilisation ou non de la norme ISO 2859-1. Si la norme n'est pas utilisée, il conviendra de fournir le taux de qualité par type de prestation (niveaux de qualité acceptables ou NQA) et les modalités de comptage des erreurs.
Ces règles de contrôle pourront différer selon le type de prestation et de livrable.

6.2.2.1 Contrôles techniques

Ils porteront sur :

- qualité des supports de livraison eux-mêmes, des livrables, et conformité par rapport au CCTP ;
- respect strict des normes et schémas fournis au CCTP, tant pour la numérisation que pour la conversion OCR ;
- respect des caractéristiques techniques des images numériques (format, résolution, codage, etc.) ;
- structure de livraison : respect de l'organisation et des nommages des répertoires et des fichiers ;
- exhaustivité et lisibilité des fichiers demandés pour la numérisation et la conversion ;
- présence, conformité et cohérence des métadonnées, en fonction des schémas et des fichiers d'accompagnement (fichier de dépouillement, données d'état physique, etc.) ;
- pour l'OCR, conformité des règles de segmentation et de structuration. Niveau de qualité de reconnaissance compris entre 96% et 98,5% (afin de permettre au moins une indexation correcte).

Selon les moyens de l'établissement, ces contrôles techniques pourront être faits automatiquement par un logiciel et seront donc exhaustifs.

6.2.2.2 Contrôles visuels

Ils porteront sur :

- qualité visuelle des images (redressement, détourage, netteté, artefacts, luminosité, qualité de la binarisation éventuelle, non troncature de l'image, correction des données de pagination…) ;

- pour la conversion, qualité de la segmentation et de la reconnaissance du texte, avec des NQA moyens par fascicule similaires à ci-dessus. Il est important de préciser que les parties illisibles ne seront pas comprises dans le comptage du taux qualité, et les erreurs non comptabilisées (ponctuation par exemple).

Concernant la qualité de la conversion en mode texte, voir en annexe l'exemple de cahier des charges de la BnF.

6.2.2.3 Modalités de reprise et de relivraison

Pour faciliter le suivi du projet, il est préférable de demander la livraison des réfections dans des lots séparés de ceux fournis pour de nouveaux documents, régulièrement au fil de la production courante plutôt qu'en fin de marché.
Il faut demander que la reprise des réfections ne retarde pas les rythmes de production courante et veiller au respect de cette exigence.

7 Livraison des fichiers numériques

Cette partie traite des modalités pratiques de livraison. Il convient de spécifier les éléments suivants :

- supports de livraison et modalités : les CD, les DVD, et actuellement la transmission par ftp, sont à exclure pour la presse dont la numérisation en niveau de gris génère des fichiers très lourds. On choisira plutôt les disques amovibles, dont la capacité est de plusieurs centaines de Go voire jusqu'à 3 To, qui se prêtent bien à ce type de projet[5]. Il est de la responsabilité du titulaire de vérifier la qualité d'enregistrement de ses disques et de leur connectique, à mesure de leur utilisation.
- Composants d'une livraison : fichiers d'accompagnement de la livraison, fichiers des documents convertis, fichiers de suivi et contenus…
- Structure détaillée de la livraison des documents numériques et/ou convertis : arborescence et nommage expliqué des répertoires et des fichiers, articulation entre elles des différentes prestations (livraison conjointe ou non de la numérisation et de l'OCR). Il sera utile de fournir un exemple permettant de faciliter la compréhension du besoin par les candidats.
- Structure du document numérique et/ou du document converti en OCR : un schéma clair est souhaitable. En annexe pourront être fournies des informations complémentaires.
- Formats de fichier d'image : spécifications techniques détaillées (métadonnées internes spécifiques à insérer dans les en-têtes de fichiers comme pour le format TIFF par exemple)

En annexe pourront être données des précisions techniques sur les formats d'images et le cas échéant les schémas XML à utiliser pour les métadonnées et le mode texte.

7.1 Structure de la livraison

La structure à élaborer sera plus ou moins simple selon le nombre et le type de prestations demandées

La règle générale est de créer un répertoire par document numérique (c'est-à-dire la plus petite unité numérique produite lorsque, par exemple, on numérise séparément chaque fascicule d'un même volume papier). Pour des besoins de suivi de la couverture numérique des originaux (et si aucune application ne le permet) on peut choisir de regrouper dans un même répertoire tous les «répertoires documents numériques» d'un volume.

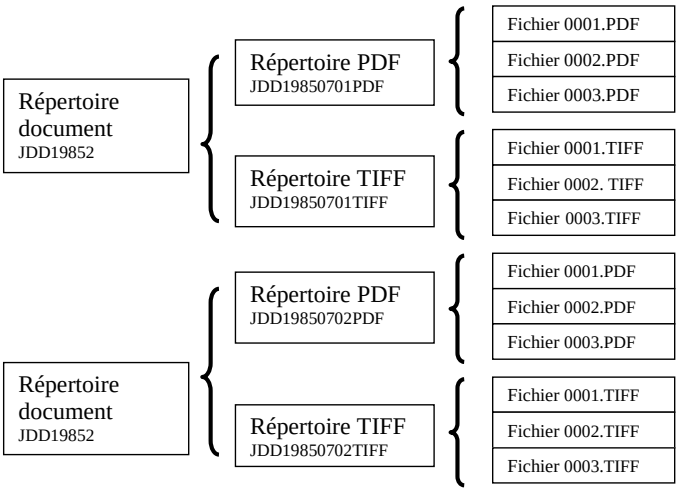
Enfin si l'on demande plusieurs formats de fichiers (par exemple fichier de diffusion compressé et fichier pour l'archivage, voire fichier texte), on peut les regrouper dans un même « répertoire document numérique » afin de s'assurer que toutes les prestations ont été faites pour un même document. Cela permet aussi de faciliter les traitements ultérieurs de mise en ligne et de stockage des documents. Il est

[5] Pour mémoire, on précisera qu'une image dont l'original est en A3 (29.7*42 cm) numérisée à 300 dpi en niveaux de gris, en format TIFF v.6 monopage non compressé, a un volume d'environ 17 Mo. Une image A2 (42*59 cm environ) de mêmes caractéristiques pèse environ 33 Mo.

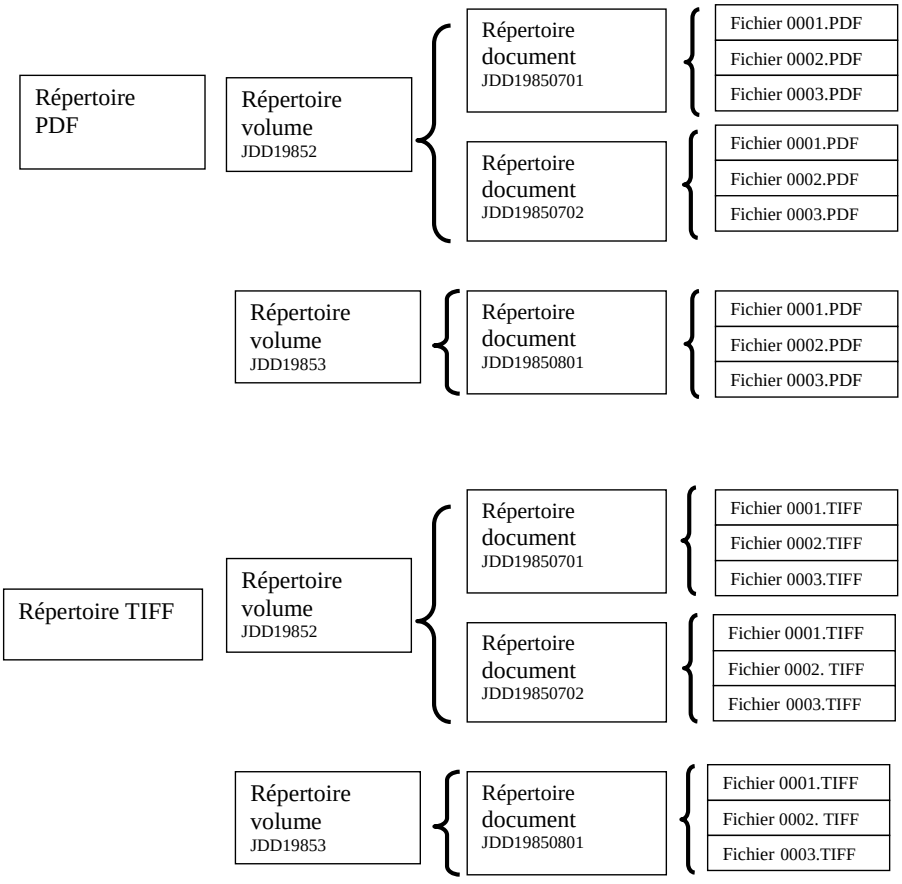
aussi possible de faire l'inverse : regrouper les documents numérique par prestation, l'arborescence choisie devant correspondre aux besoins de traitement de la chaîne d'entrée mise en place par chaque établissement.

Ainsi on pourra obtenir une structure arborescente plus ou moins complexe :

Arborescence au document



Arborescence à la prestation et au volume



7.2 Règles de nommage des fichiers

On peut choisir pour l'identification des noms de répertoire significatifs ou non.
Exemples :

« Les fichiers livrés seront organisés de la manière suivante :
Pour un numéro de la revue, les images océrisées seront regroupées dans un répertoire au format PDF. Un répertoire correspondra à un numéro de revue, il contiendra donc le texte et les planches s'y rapportant.
Les identifiants procéderont de la façon suivante :
- Pour un répertoire, la forme sera : XXXX_AAAA_NNN
- Pour un fichier image, la forme sera : XXXX_AAAA_NNN_EEE
- Pour une planche hors-texte, la forme sera : XXXX_AAAA_NNN_PIII

XXXX : Code de la revue
AAAA : année de publication
NNN : numéro de la revue
EEE : numéro de fichier image texte
PIII : P : lettre fixe indiquant que le fichier est une planche. III sera un nombre indiquant le numéro de la planche dans ce numéro. »

« Soit le numéro 2 de l'année 1907 de la Construction moderne contenant 10 pages et 2 illustrations séparées :
Le répertoire sera COMO_1907_002, il contiendra :

- *des fichiers de texte (COMO_1907_002_008) ;*
- *des fichiers planches se rapportant aux pages de texte(COMO_1907_002_P002).*

Tous ces noms seront précédés d'un préfixe propre à la Cité de l'architecture et du patrimoine qui est FRAPN02. »

Exemple similaire au précédent utilisé en Rhône-Alpes avec pour noms de fichiers les éléments suivants : la référence du titre de presse, la date du jour concerné, le numéro de la page du fichier concerné : 01JOURNALAIN_18090118_P_0001.pdf. A vérifier.

8 Stockage et conservation des documents numériques

8.1 Garantie de stockage et de restitution par le prestataire pendant le projet de numérisation

Il peut-être utile de prévoir au cahier des charges une prestation de stockage des données produites par le prestataire pendant toute la durée du projet et même pendant plusieurs mois après les dernières livraisons afin de s'assurer que l'on a bien terminé le processus de transfert/exploitation des données sans aucune perte. Le prestataire sera tenu de relivrer les données réclamées par le commanditaire dans un délai et des modalités qui seront précisés.

8.2 Règles générales de préservation pour l'établissement

L'établissement doit mettre en œuvre des moyens de stockage lui permettant d'exploiter et de conserver les données numériques dans de bonnes conditions. Il doit pour ce faire évaluer suffisamment tôt avant le démarrage du projet les volumes des fichiers qui seront produits afin d'acquérir la capacité et les outils de gestion du stockage en conséquence. Si l'établissement souhaite sous-traiter le stockage et la mise en œuvre des moyens de préservation à un tiers, il doit monter un cahier des charges requérant un minimum d'opérations et de moyens à mettre en œuvre.

Quel que soit le support choisi (bande, disque dur, support optique), il convient de mettre en place des processus de surveillance et de sauvegarde afin d'éviter la perte ou la corruption des données.

Les principales actions à mener sont :

- **Caractériser des profils de documents**

Afin de mieux cibler les actions à mener, la première étape consiste à lister les différents types de données à conserver :

- formats de fichier ;
- durée de vie des documents, gestion des versions, gestion des remplacements/suppressions) ;
- informations à pérenniser (type de données pour un même document : fichiers numériques, fichiers textes, métadonnées, liens entre les différents types de fichiers d'un même document…) ;
- type de supports utilisés pour le versement dans l'archive et éventuellement moyens réseaux ou système d'échange à mettre en place si le producteur des données est distant ;
- type d'accès à gérer et fréquence d'utilisation : gestion des droits (droits de propriété intellectuelle sur les documents, niveaux d'habilitations pour les différents types d'usagers et selon le type d'utilisation, i. e. consultation, impression, copie), type d'utilisateurs, temps d'accès selon les types de documents.

- **Assurer la conservation des documents numériques**
  - Le stockage de préservation doit être redondant et répliqué dans au moins une deuxième localisation distante ;
  - par mesure de sécurité et pour pallier les déficiences des matériels, les supports utilisés pour la préservation doivent être différents de ceux utilisés pour la consultation : on pourra choisir des supports plus lents et de plus grande capacité d'enregistrement pour la préservation (bandes par exemple) et si les documents sont très consultés, des supports rapides pour la consultation (disques RAID). Dans un cahier des charges il n'est pas utile d'indiquer le type de support, mais il faut préciser de manière détaillée les besoins d'utilisation (temps d'accès, fréquence des accès, temps minimum de restitution d'un document : pour la consultation, pour fournir des copies, etc.) ;
  - une surveillance régulière des supports d'enregistrement doit être mise en place pour anticiper les failles et mettre en œuvre des recopies préventives ;
  - des contrôles de consistances et d'erreurs doivent être faits : l'intégrité des données doit être contrôlée lors du versement dans l'archive, puis régulièrement (contrôles de redondance cycliques CRC) ;
  - une veille technologique doit être menée afin d'anticiper l'obsolescence des matériels, des logiciels et des formats. L'établissement pourra s'appuyer sur des structures ou organismes de références assurant cette veille. Cela suppose de connaître parfaitement le contenu de l'archive et les changements de technologies qui peuvent survenir à plus ou moins long terme. Les actions à mener pour prévenir les défaillances de support et anticiper les obsolescences, sont : la duplication de support, la migration des versions de formats, la migration d'un format dans un autre. On peut distinguer deux types de migrations :
    - les migrations légères qui sont des opérations internes au système de stockage (le rafraîchissement est la recopie d'un support sur un support de même type sans impact sur l'architecture matériel ; la réplication est une recopie avec éventuellement changement de support et donc impact sur l'architecture matérielle. Le contenu des données reste inchangé).
    - les migrations lourdes (réempaquetage et transformation) qui ont un impact sur l'ensemble du système d'archive.
  - Gérer le cycle de vie des documents (par exemple définir une durée de vie par type de document qui sera gérée automatiquement par le système) Mettre en place les règles de mise à jour/suppression :
    - demande de destruction par l'utilisateur ou à expiration d'un délai de garde ;
    - validation de la destruction par l'administrateur, le document n'apparaît plus mais sa présence est conservée ;
    - destruction complète du document au delà du délai de garde de suppression. Dans le cas des bandes et des supports réinscriptibles, les données ne seront donc pas migrées lors du prochain retraitement.

- Fournir une interface technique stable permettant divers services (restitution, nommage pérenne, création d'arborescence virtuelle,...).

- **Assurer l'accès**
  - Rechercher les données qui intéressent les utilisateurs (sélection via les métadonnées) ;
  - sélectionner et récupérer les données qui correspondent à leur besoin ;
  - transformer si besoin les données avant de les fournir à l'utilisateur ;
  - fournir les données *via* le réseau ou sur le support adéquat.

9 Conclusion

Très consultées par le public, les collections de presse méritent d'être préservées et mises en valeur grâce à la numérisation et la conversion OCR. Les spécifications du cahier des charges permettent de mettre en œuvre la première étape qui consiste à produire les fichiers numériques de manière à assurer leur exploitation en vue de la consultation et de la sauvegarde des données numériques.
L'étape suivante doit prendre en compte les besoins de son public pour effectuer la mise en ligne des documents et assurer leur pérennité pour des utilisations futures.
Les collections originales très fragiles ont alors un substitut représentatif de leur richesse qui devient un véritable master de sauvegarde, qui peut être exploité de diverses manières grâce aux images et au texte converti. La version numérique doit continuer à vivre et à évoluer avec les techniques afin de rester accessible au cours du temps ; les actions de surveillance et de maintien des fichiers, supports et moyens de lecture devront être assurées.