La presse ancienne locale numérisée n°2

Numérisation de l’Indépendant et du Courrier de Tarn-et-Garonne

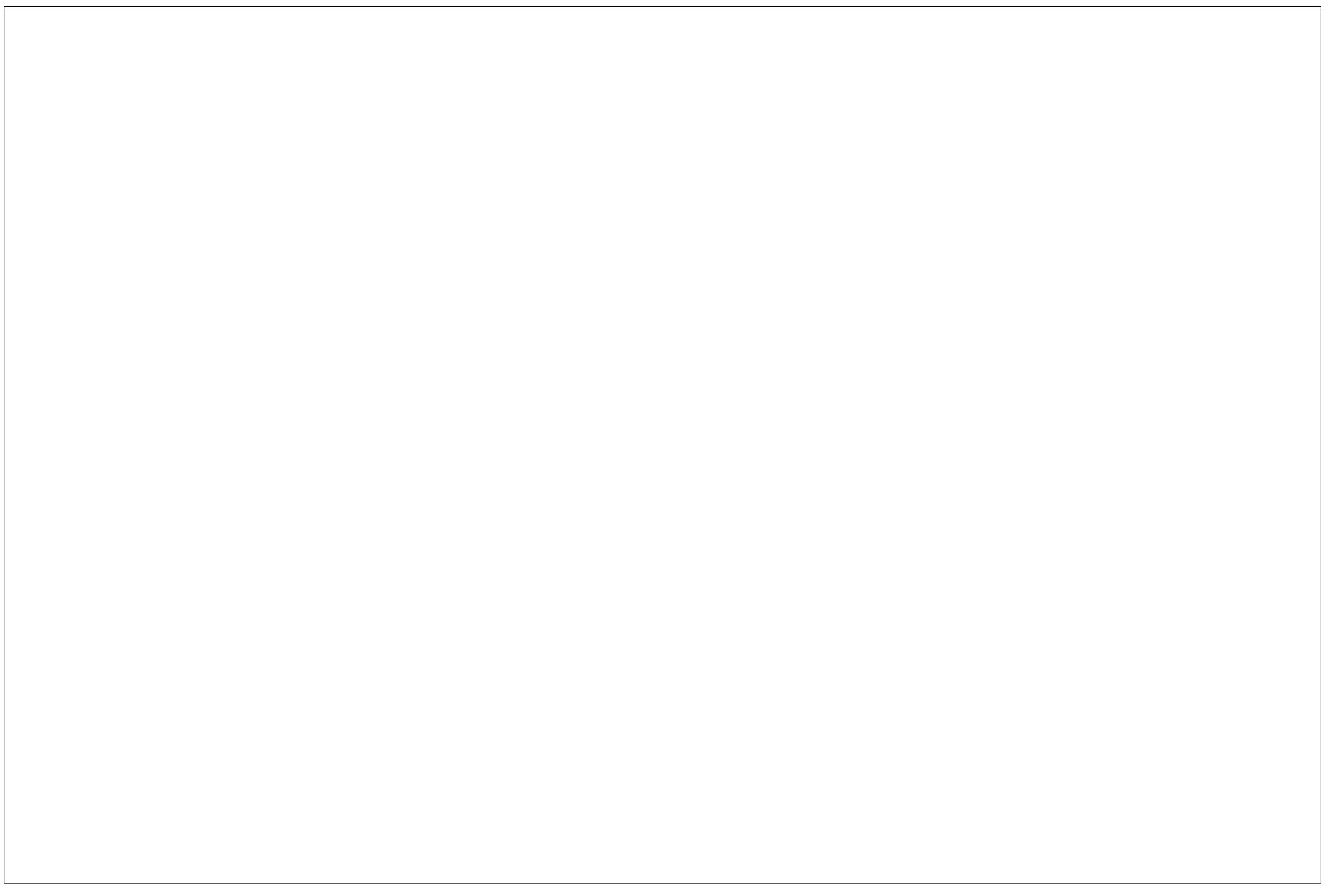
Les différentes étapes de traitement, du papier journal au journal en ligne

Acte 1, la préparation des collections

Acte 2, le contrôle-qualité des fichiers numériques

La phase de contrôle des images numérisées et océrisées est en cours d’achèvement.

Un travail de longue haleine, fastidieux, puisque chaque page (soit près de 49 000 vues) a été « visionnée » par les équipes dévolues au sein des Archives départementales, de la Mémo patrimoine et d’Occitanie Livre et Lecture (sans disposer des originaux, conservés chez le prestataire jusqu’à la validation finale de l ’opération).

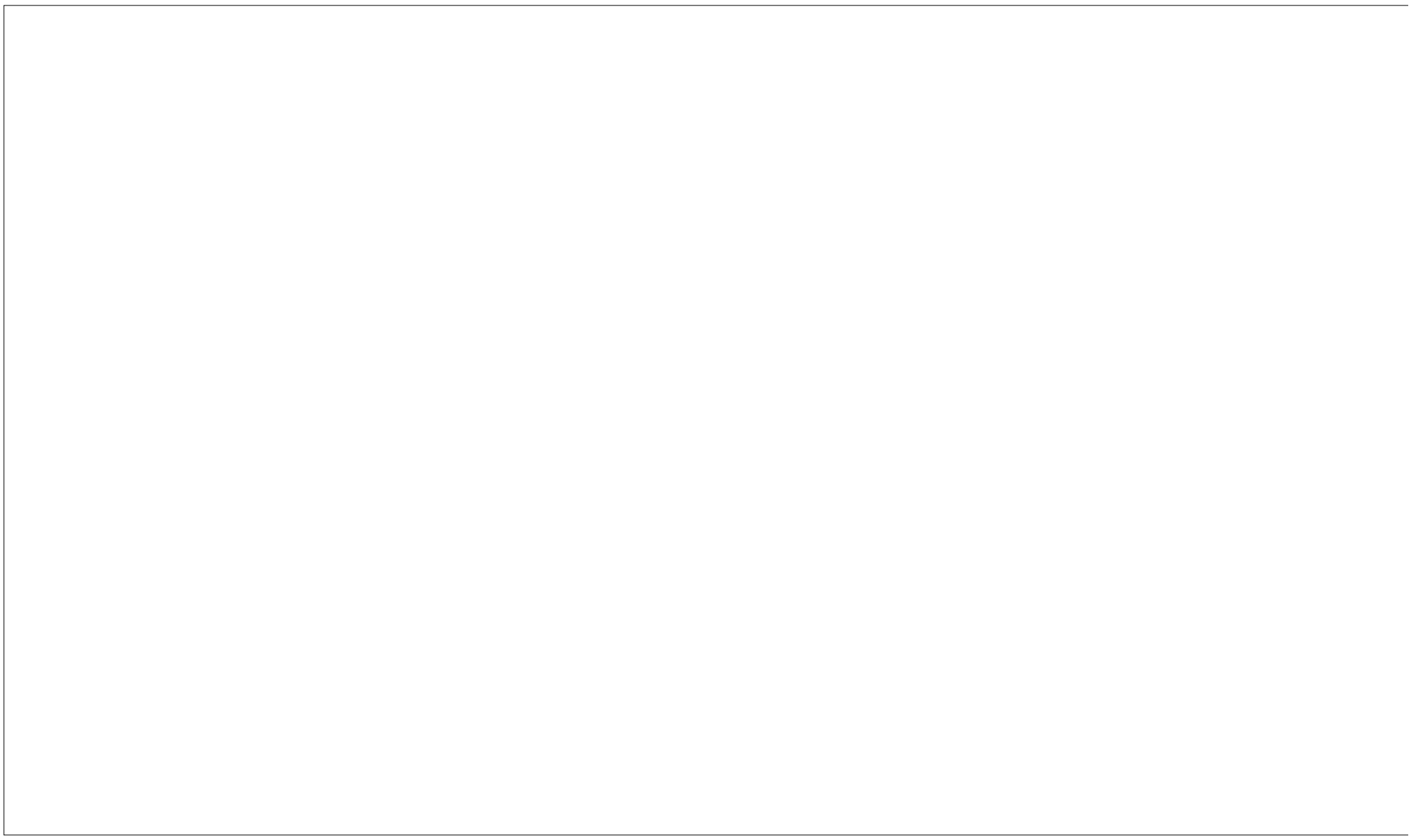


La vérification s’est faite à deux niveaux. Il a d’abord fallu s’assurer qu’aucune page n’avait été oubliée, que les spécifications techniques avaient été respectées (netteté constante, image redressée, marge respectée, contraste satisfaisant, etc.) et que les fichiers avaient été identifiés à la bonne date et dans le bon ordre des

pages au moment de la numérisation des fascicules.

Dans un second temps, le contrôle s’est porté sur la reconnaissance optique des caractères... de quoi s ’agit-il ? Plus connu sous le nom d’ « OCR » (pour Optical Character Recognition), ce traitement informatique automatisé analyse l’image, identifie les zones de texte et détecte les symboles reconnaissables (caractères, dictionnaire de mots). Le résultat est une reconstitution du texte imprimé contenu dans une image. Les journaux non seulement sont consultables en ligne, mais deviennent interrogeables dans le texte.

Sur des documents patrimoniaux de type presse, le taux de satisfaction sur la transcription proposée est de l ’ordre de 70 %, ce qui est suffisant pour alimenter un moteur d’indexation en plein texte.



Ce taux peut sembler relativement bas, mais la qualité de reconnaissance dépend malgré tout d’un grand nombre de facteurs liés tant au document original qu’à la numérisation elle-même. Ainsi :

- le support doit être en très bon état de conservation : le texte pris dans des pliures ou des courbures, ou dont l’encre a migré, donnera de mauvais résultats ;

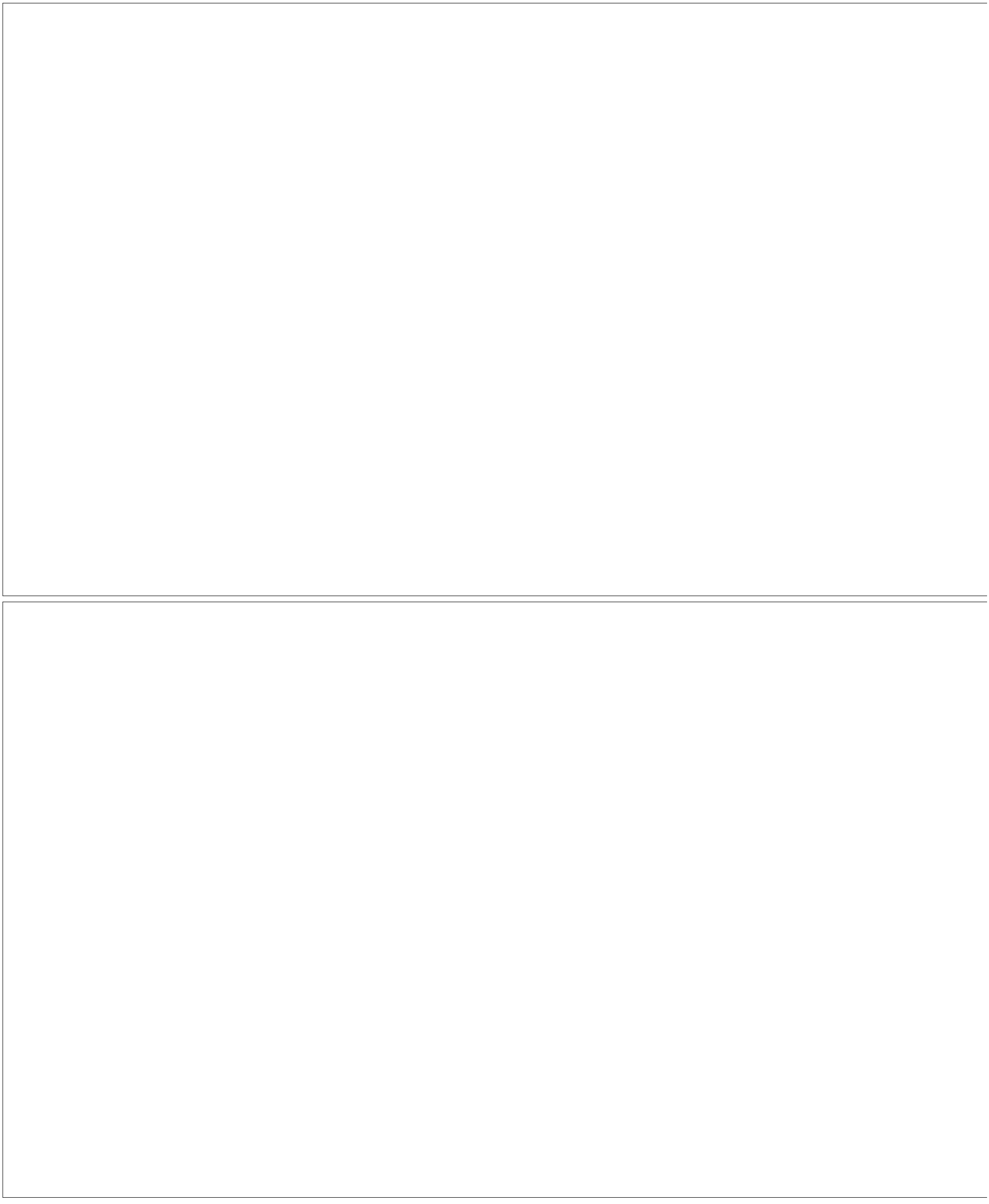
- le texte doit être bien lisible : les caractères trop empâtés, les polices « originales », trop petites, trop grandes sont difficilement traitables ;

- la mise en page du document doit être simple : l’OCR ne fonctionnant pas bien sur les mises en page complexes, telles que des colonnes multiples, des encarts publicitaires, des images ou des typographies particulières ;

- les images numériques doivent être assez précises et contrastées pour faciliter le travail de l’applicatif OCR ;

- le logiciel d’océrisation doit être performant sur la reconnaissance de la structuration du document, des symboles, et son dictionnaire le plus riche possible.

Il arrive également que le logiciel ne passe pas de façon uniforme sur du texte simple, laissant ainsi de côté des paragraphes entiers, des lignes ou de façon plus subtile quelques caractères en début ou fin de colonne.



Les Archives départementales ont fait remonter au prestataire les oublis d’océrisation sur le texte, afin qu’il effectue une reprise sur les fichiers défectueux. L’océrisation définitive sera donc plus complète que la première version proposée, dans la limite des contraintes matérielles et informatiques évoquées plus haut.

La diffusion des fichiers et la recherche en plein texte de ces titres de presse se feront via la médiathèque numérique d’Occitanie. Pour le moment, ce site ne centralise que des titres de presse ancienne de l’ancienne région Languedoc-Roussillon. Une mise à jour en début d’année prochaine permettra de consulter les titres déjà disponibles sur le portail Palanca ainsi que le *Courrier de Tarn-et-Garonne* et *l’Indépendant de Tarn-et-Garonne*.

Un lien vers ces ressources ainsi que vers un programme de correction participatif de l’OCR des titres de presse sera proposé sur le site archives82.fr, dans la rubrique Rechercher et consulter / Archives en ligne.

Restez donc connectés, et en attendant, n’oubliez pas de feuilleter la presse libre de droits sur Gallica et de découvrir le catalogue des titres de presse locale, conservés aux Archives départementales de Tarn-et-Garonne ou conservés dans d'autres bibliothèques.

Site "presse locale ancienne" de la BNF.

Catalogue de la presse, Ad82.