OCR à la BnF

Zoom sur la chaîne de traitement OCR
de la Bibliothèque Nationale de France.

Elliot Fabert
18 juillet 2024

Introduction

La Bibliothèque Nationale de France (BnF) a mis en place une chaîne de traitement OCR interne, visant à produire son propre OCR (Reconnaissance optique de caractères) sur ses documents numérisés. Cette initiative permet de pouvoir enrichir les documents numériques avec des données textuelles exploitables, afin de faciliter la recherche ainsi que la consultation. Le processus de l'OCR repose sur différentes étapes clefs passant de la sélection des documents au traitement OCR, et est structuré autour de critères rigoureux et d'une organisation soignées.

Mode de traitement et Critères de sélection

Avant de pouvoir commencer tout traitement OCR, les équipes du BCN, bureau chaine numérique, de la BnF passent par l'application DispoDocNum, afin de sélectionner les documents et de les regrouper selon des critères spécifiques, permettant ainsi de déterminer les priorités et d'optimiser le processus de traitement.

Pour cela, deux principaux modes de traitement existent : le *traitement courant* et le *traitement rétrospectif*.

- Traitement courant : Ce mode de traitement consiste simplement à appliquer de l'OCR sur des documents récemment numérisés par les prestataires de la BnF. On y applique une période d'attente avant de procéder à l'OCR, permettant une gestion fluide et continue des flux de documents.

- Traitement rétrospectif : Ce mode de traitement consiste à ajouter de l'OCR sur des documents déjà numérisés présents dans les catalogues, mais aussi sur des documents bien plus anciens de la BnF.

Un intervalle de traitement est établi manuellement par les équipes pour optimiser la gestion des flux documentaires. Cette approche vise à prévenir toute surcharge ou, à l'inverse, tout ralentissement excessif du traitement. L'objectif principal est de maintenir un flux constant et efficace. Il convient de noter que la Bibliothèque nationale de France (BnF) traite en moyenne environ 5 000 documents par jour, ce chiffre représentant le pic d'activité quotidienne. Cette donnée souligne l'importance d'une gestion rigoureuse du flux documentaire.

La définition de l'intervalle de traitement s'adapte à la nature des prestations. Pour les prestations s'étendant sur plusieurs mois, il est possible de traiter les documents quotidiennement. En revanche, pour les prestations couvrant une ou plusieurs années, un intervalle spécifique est mis en place. Par exemple, pour l'année 2021, on pourrait instaurer un intervalle de trois jours, permettant ainsi de traiter les documents du 1er au 3 janvier, puis du 4 au 6 janvier, et ainsi de suite.

Une prestation de numérisation, identifiée par un numéro unique (par exemple 115, 706, 219), représente un ensemble cohérent de documents à numériser. Cette prestation peut être réalisée soit par un prestataire externe, soit par un atelier interne de la BnF. Chaque prestation correspond à un type spécifique de documents (comme des monographies, des périodiques, ou des manuscrits) et à un prestataire particulier. Cette organisation permet une gestion efficace et un suivi précis des différents projets de numérisation au sein de la BnF.

Dans le cadre de la méthode courante, un délai prudent de 100 jours est observé entre la numérisation d'un document et son traitement OCR. Ce délai sert plusieurs objectifs essentiels :

1. Il permet au service de numérisation d'effectuer d'éventuelles corrections ou réfections si des erreurs sont détectées dans le processus initial de numérisation.
2. Il offre aux utilisateurs de Gallica l'opportunité de signaler des problèmes potentiels dans les documents récemment mis en ligne.

Il est important de noter qu'une fois un document océrisé, toute réfection devient impossible. L'annulation de l'OCR n'est pas une option prévue dans la chaîne de traitement de la BnF. En effet, un blocage est automatiquement mis

en place dans la base de données lorsque le Service Numérisation tente une réfection sur un document erroné et océrisé.

Cependant, ce processus reste largement théorique et s'adapte, en pratique, aux volumes traités. Avec un pic quotidien pouvant atteindre 5 000 documents, le contrôle exhaustif devient humainement impossible. Bien qu'un bureau de contrôle qualité ait été mis en place au sein du service de numérisation, il ne parvient à examiner que 4 à 5% des numérisations journalières. Cette situation souligne le défi constant de concilier le maintien de la qualité et la gestion de volumes importants dans le processus de numérisation et d'océrisation de la BnF.

Pour assurer un suivi efficace et concret de la préparation des documents, la BnF utilise un système de tables de données interconnectées. Une table principale est dédiée à l'enregistrement des documents sélectionnés pour d'océrisation. Cette table est ensuite mise en relation avec la table des livraisons de documents, permettant ainsi une vision globale et détaillée du processus. Cette méthode de croisement des données offre plusieurs avantages comme connaître précisément la date de livraison des documents, savoir quel document a été rejeté au cours du processus et surtout, pouvoir sélectionner les documents à océriser. La table est interrogée par le biais de requêtes SQL afin de sortir les documents océrisables.

Lorsqu'un document est éligible à l'océrisation, il est ajouté à une table dédiée contenant tous les flux de documents en attente de traitement. Chaque document y est enregistré avec les détails de son parcours à travers les différentes étapes de la chaîne de traitement. Le processus commence par une étape de vérification du type de document. Seuls deux types de documents sont éligibles pour poursuivre le traitement[1] :

[1] c.f annexe, figure 1a/b.

- les monographies à partir de la moitié du XVIIIe siècle
- les périodiques.

Si le document appartient à une autre catégorie, comme une carte numérisée, il sera rejeté et ne passera pas dans le moteur d'océrisation. En revanche, une numérisation des *Fleurs du mal* (une monographie postérieure à 1750), par exemple, serait acceptée et continuerait dans la chaîne de traitement.

Il en est de même avec le code langue. Le moteur OCR doit connaître la langue du document pour proposer une océrisation, car il ne détermine pas de façon autonome la langue d'un document. Pour cela, on interroge l'entrepôt OAI de la BnF afin d'obtenir le code langue du document. L'entrepôt OAI (Open Archives Initiative) est un système de gestion des métadonnées bibliographiques qui permet de centraliser et de standardiser l'accès à ces informations. En interrogeant cet entrepôt, les données récupérées permettent d'identifier si le document est en français, latin, allemand ou occitan, par exemple[2]. Certains documents peuvent avoir plusieurs langues enregistrées dans la métadonnée code langue. Lorsque la langue n'est pas reconnue, le document passe à l'état « abandonné » dans la table, signifiant qu'il y a eu une demande d'océrisation pour ce document mais qui ne sera jamais reprise car le code langue ne concorde pas avec les exigences du moteur OCR. Cela se produit notamment lorsque le document est multilingue.

Concernant le moteur OCR, l'établissement a acquis une licence pour numériser l'équivalent de 20 millions de pages A4 avec le moteur ABBYY, que nous présenterons plus en détail ultérieurement. Ce moteur prend en charge de nombreuses langues comme le français, le grec, le tatar, ou encore l'azerbaïdjanais. Bien que la prise en charge de l'ancien français soit prévue, elle n'est pas encore fonctionnelle pour le moment.

[2] https://www.bnf.fr/fr/les-entrepots-oai-de-la-bnf

Zoom sur SPAR

La Bibliothèque nationale de France (BnF) utilise SPAR (Système de Préservation et d'Archivage Réparti) comme pierre angulaire de sa stratégie de conservation numérique[3]. Lancé en 2010, SPAR est bien plus qu'un simple outil de stockage ; c'est un dispositif complexe visant à pérenniser l'information numérique. Les missions principales de la BnF sont :

1. La conservation du patrimoine
2. La diffusion des connaissances

SPAR s'inscrit dans la mission de conservation, avec l'ambitieux objectif de préserver les documents numériques sur le très long terme. Son rôle va au-delà de l'archivage sécurisé, Il assure la lisibilité à long terme des documents, malgré l'évolution des technologies. Et il maintient la compréhensibilité et la réutilisabilité de l'information, indépendamment des changements d'environnement technique et humain.

Face aux défis de la dégradation numérique et de l'obsolescence technologique, SPAR adopte une approche proactive :

- Veille technologique constante
- Migrations préventives des données

Cette stratégie anticipe les problèmes plutôt que de tenter de réparer des documents déjà endommagés ou devenus illisibles. En constante évolution, SPAR intègre régulièrement de nouvelles collections de la BnF et fonctionnalités, s'adaptant aux besoins croissants de la conservation numérique.

[3] https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti

De plus, le caractère "réparti" de SPAR se manifeste par sa capacité à gérer et stocker de multiples copies des documents sur différents sites géographiques. Cette redondance stratégique vise à minimiser les risques de perte ou de destruction des données, renforçant ainsi la résilience du système de préservation.

Mais pourquoi parler de SPAR ? Tout simplement car SPAR joue un rôle important dans le processus de numérisation et d'océrisation, se positionnant à la fois en amont et en aval de la chaîne de traitement. En entrée, SPAR fournit les documents numériques initiaux destinés à l'océrisation. Une fois le traitement effectué, le système accueille en sortie les versions OCRisées comme de nouveaux documents. Ces versions enrichies coexistent avec les originaux dans SPAR, formant ainsi une sorte de doublon amélioré. De plus, en sortie de chaîne, le processus se divise en deux volets distincts : d'une part, la préservation à long terme dans SPAR, et d'autre part, la diffusion publique via Gallica.

Le processus d'Océrisation

Après avoir examiné le processus de sélection des documents à océriser par la BnF, plongeons au cœur de la chaîne de traitement OCR. La première étape est le passage des documents à la phase de « Demande d'océrisation ». C'est à ce moment que le moteur ABBYY entre en jeu pour effectuer l'OCR sur le document. Avant de nous étendre sur le processus, il est nécessaire de faire un point sur ABBYY.

ABBYY, entreprise américaine leader dans le traitement documentaire, est particulièrement reconnue pour son expertise en OCR. Le « moteur ABBYY » utilisé par la BnF est *ABBYY FineReader*, un kit de développement permettant

aux développeurs de créer des outils capables d'extraire des informations textuelles à partir de diverses sources : documents papier, images, ou même écrans[4].

FineReader se distingue par ses algorithmes avancés de reconnaissance de texte, offrant une extraction d'informations d'une grande précision tout en préservant la mise en page originale et les éléments structurels du document. Son atout majeur réside dans sa capacité à s'intégrer de manière fluide et transparente dans des workflows - ou flux de travail - automatisés et des solutions de gestion documentaire, le rendant ainsi particulièrement adapté aux environnements de travail complexes et variés, comme la BnF. Cette intégration est facilitée par des API (*Application Programming Interface* ou *Interface de Programmation d'Application*) robustes et des outils de développement flexibles. Ces fonctionnalités permettent aux entreprises de personnaliser et d'automatiser leurs processus OCR en fonction de leurs besoins spécifiques, offrant ainsi une solution sur mesure et hautement efficace.

Pour lancer le processus d'océrisation des documents, les équipes de la BnF utilisent une application nommée SchedulingTool. Cette application, que l'on pourrait qualifier de "robinetterie", fonctionne comme un lanceur permettant d'initier des tâches de façon périodique. Par exemple, elle peut traiter un nombre défini de documents toutes les x secondes, minutes ou jours. À la BnF, les documents sont traités à intervalles réguliers de x minutes.

Cette approche permet une gestion efficace de l'espace de stockage. Pour illustrer, imaginons un poulailler abritant 100 poules très productives, pondant chacune 5 œufs par jour. Si la capacité de stockage est limitée à 1000 œufs, il est crucial de gérer quotidiennement la production pour éviter tout débordement, la capacité n'étant pas infinie.

[4] https://www.abbyy.com/fr/ocr-sdk/

Le même principe s'applique au traitement des documents. Si un paquet de données occupe 50 mégaoctets et qu'un million de paquets similaires sont en attente, l'espace total requis atteindrait 50 pétaoctets, soit 50 000 téraoctets. Or, les capacités de stockage de la BnF, bien que considérables, ne sont pas illimitées. De plus, cette gestion minutieuse permet d'optimiser l'utilisation des ressources informatiques, assurant ainsi un flux de traitement continu et efficace.

Mais concrètement, si nous changeons d'échelle, que voit-on ? Tout ceci est visible via l'application SuiviNum, dont nous détailleront le fonctionnement par la suite. Le processus détaillé se déroule comme suit :

1. Prise en compte de la demande d'océrisation
2. Fabrication du paquet de données
3. Récupération des images via les systèmes SPAR et FLR (systèmes internes à la BnF)
4. Traitement OCR proprement dit

Le cœur du système est le GenOCR Manager, qui gère la distribution des tâches aux workers. Ce manager récupère les documents, prend les images, et les envoie aux workers pour traitement. La BnF dispose de 4 serveurs dédiés, chacun avec 32 threads[5], qui tournent continuellement pour optimiser le processus. Ces workers traitent les images en parallèle, page par page, sans notion de document complet. C'est le manager qui reconstitue le document une fois toutes les pages traitées.

Petit point sur les workers : ce que l'on appelle un « worker » est un composant logiciel spécialisé qui exécute des tâches spécifiques. À la BnF, il s'agit d'un programme écrit en JAVA, un langage de programmation. Ce programme permet de gérer les océrisations page par page, traitant chaque image individuellement. Les workers sont supervisés par un « manager », lui

[5] un « fil d'exécution », en français, est une séquence d'instructions que l'OS - système d'exploitation - peut gérer indépendamment. Cela permet à un programme de pouvoir exécuter plusieurs tâches en simultané. Ici, cela signifie donc que chaque serveur peut exécuter 32 tâches de traitement simultanément.

aussi un composant logiciel, qui joue un rôle crucial dans la coordination et la supervision de l'ensemble du processus. Le manager remplit plusieurs fonctions essentielles. Il est responsable de la distribution des tâches, assignant les pages à océriser aux différents workers disponibles. Il surveille en temps réel la progression de chaque tâche attribuée et coordonne l'ensemble pour s'assurer que le système fonctionne de manière fluide et efficace. De plus, le manager veille à ce que les workers puissent récupérer les nouvelles tâches à partir d'une file d'attente centralisée. Il supervise également la gestion des erreurs éventuelles par les workers, assurant ainsi la robustesse du système. Cette architecture permet une parallélisation efficace du traitement, optimisant ainsi les performances globales du système d'océrisation de la BnF. Le manager agit comme un chef d'orchestre, s'assurant que chaque worker joue sa partition au bon moment, tout en maintenant la cohérence de l'ensemble du processus.

Cette approche permet une gestion efficace de l'espace de stockage et des ressources informatiques. Pour une efficacité optimale, le système maintient entre 200 et 400 pages en traitement simultané. Cette gestion minutieuse permet d'optimiser l'utilisation des ressources informatiques, assurant ainsi un flux de traitement continu et efficace. Un point intéressant à noter est la flexibilité du système. En théorie, si la BnF décidait de changer de moteur OCR, passant par exemple d'ABBYY à un moteur open source comme Tesseract[6], cela serait possible en modifiant simplement les workers. Cependant, ABBYY reste privilégié pour ses performances, notamment en matière de pré-traitement des images.

Une fois l'OCR terminé, le document suit le processus normal de la chaîne d'entrée de la BnF. Tout ce processus est visible via l'application SuiviNum, que nous détaillerons par la suite.

[6] Moteur libre développé par Hewlett-Packard dans les années 1980, aujourd'hui maintenu par Google.

La chaîne d'entrée est un processus qui reçoit les documents numérisés, qu'ils proviennent d'un prestataire externe ou interne comme nous l'avons vu précédemment. Il s'agit d'un paquet contenant des images et des fichiers de métadonnées (master). Cette chaîne effectue des vérifications et des contrôles sur les métadonnées afin de s'assurer que tout soit en ordre et que les références concordent avec la notice du catalogue de la BnF. Un contrôle est également réalisé sur les images pour vérifier leurs dimensions et s'assurer que le document n'est pas corrompu. Ensuite, le catalogue est sollicité afin de mettre à jour la notice du document et pour signaler sa numérisation. Le document sera alors envoyé sur Gallica pour la diffusion et sur SPAR pour la conservation.

Cette chaîne d'entrée, que nous avons déjà plus ou moins évoquée, mérite d'être explicitée car un document océrisé (ou « traitement complémentaire ») repasse par le même chemin qu'un document numérisé pour la première fois (une « primo-livraison »). C'est comme si on traitait un nouveau document à part entière. Cette approche assure une cohérence dans le traitement de tous les documents, qu'ils soient nouvellement numérisés ou qu'ils aient bénéficié d'une océrisation ultérieure.

La différence entre les primo-livraisons et les traitements complémentaires, comme l'OCR, réside dans la présence de nouveaux fichiers : les ALTO. Le XML ALTO, acronyme d'Analyzed Layout and Text Object, est un format XML standard utilisé pour décrire la disposition et le contenu textuel des pages numérisées. Dans le cadre de l'OCR à la BnF, les fichiers ALTO jouent un rôle crucial en structurant les informations extraites par le processus d'océrisation. Ces fichiers ne se contentent pas de conserver le texte brut, mais incluent également des données précieuses telles que les positions exactes du texte sur la page, les différents styles typographiques utilisés, ainsi que l'emplacement des images. Cette approche permet de préserver fidèlement la mise en page originale des documents.

De plus, l'OCR interne ajoute une difficulté supplémentaire : la création d'un nouveau manifeste qui renseigne l'océrisation. Comme nous l'avons déjà mentionné, le fichier d'un document numérisé se compose des images et d'un fichier de métadonnées, appelé manifeste. Dans le cas de l'OCR, on ajoute non seulement les fichiers ALTO à ce dossier, mais on doit également créer un nouveau manifeste qui indique qu'une océrisation a été effectuée sur le document, ce nouveau manifeste venant compléter les informations du document initial. À la suite de cela, LIVATEL, un nouveau badge, va chercher les paquets océrisés qui se trouvent sur un espace de stockage puis va les livrer à la chaîne d'entrée. Pour bien illustrer, il faut imaginer deux espaces bien distincts : d'un côté l'espace de production, où l'on va ajouter les ALTOs au fur et à mesure de leur création, et de l'autre un espace de livraison de paquets, la « PEF », pour Plateforme d'Échange de Fichiers, où la chaîne d'entrée va scruter les paquets qui ont été livrés afin de les intégrer dans le processus d'entrée. Les documents sont distribués dans des dossiers de prestations (un dossier par prestation) et les documents océrisés sont acheminés vers les prestations 560 à 564. Pour comprendre, le 560 représente le courant, le 561 représente le rétrospectif, le 562 représente l'espace esco (les partenaires livrant des images), 563 est pour les ateliers internes de la BnF et le 564 pour l'OCR à la demande (on peut prendre en compte le 565 - Urgences mais il n'a jamais été activé). Ces paquets sont donc distribués dans chaque dossier suivant leur nature et finissent dans un dossier appelé « Liv-externe » qui est le fichier où la chaîne d'entrée va venir chercher les paquets pour les faire entrer ensuite.

Point rapide sur les dossiers de prestation : Pourquoi avoir des dossiers séparés ? Tout simplement car il n'y a pas que de l'OCR dans les dossiers, ces derniers contiennent tous les prestataires de la BnF. Et ces prestataires font exactement la même chose que LIVATEL, après numérisation, ils envoient via un

client FTP, *File Transfert Protocol* ou *Protocole de Transfert de Fichiers*[7], des paquets numérisés afin qu'ils puissent entrer dans la chaîne.

Au coeur de la chaîne d'entrée : SuiviNum.

Avant de rentrer dans la présentation de SuiviNum, il est important d'expliquer comment fonctionne la chaîne d'entrée.

La chaîne d'entrée est un processus qui reçoit les documents numérisés par le prestataire, comprenant des images et des fichiers de métadonnées. On vérifie alors que les métadonnées sont correctes et que les références à la notice du catalogue sont exactes. Les documents sont ensuite envoyés, dans un premier temps, vers la préservation avec SPAR, puis vers la diffusion afin qu'ils soient disponibles sur Gallica. La préservation précède la diffusion afin de pouvoir prendre en compte les éventuels rejets qui peuvent survenir avec SPAR ; le cas échéant, le processus s'arrête à SPAR. La préservation est l'étape la plus importante, car il faut prévenir les problèmes potentiels (crashes informatiques, malveillance, incendie, etc.), d'où l'utilisation de SPAR avant la diffusion sur Gallica.

Un point central dans ce processus est le suivi de la chaîne d'entrée. Pour assurer la supervision, une table de suivi a été mise en place, offrant à tous les intervenants une vue d'ensemble sur l'état d'avancement des documents[8]. Cette table indique avec précision à quelle étape du processus chaque

[7] Protocole permettant de pouvoir transférer des fichiers entre un ordinateur local et un serveur distant via le protocole FTP. Il offre une interface pour naviguer dans les répertoires du serveur, télécharger des fichiers depuis le serveur vers l'ordinateur local (download), et envoyer des fichiers depuis l'ordinateur local vers le serveur (upload). Les clients FTP sont couramment utilisés pour gérer le contenu de sites web, partager des fichiers volumineux, ou synchroniser des données entre différents systèmes.

[8] cf annexe fig 2.

document se trouve et dans quel état il est. L'application SuiviNum joue ici un rôle clef en proposant une interface graphique intuitive qui donne un aperçu clair et détaillé de ce qui se passe au cœur de la chaîne. SuiviNum permet de suivre toutes les chaînes d'entrée de la BnF, pas uniquement l'OCR. Cette table comporte des colonnes essentielles :

- Une colonne pour les documents en attente.
- Les colonnes d'état : Indiquant si le document est en demande d'océrisation ou en cours d'océrisation.
- Les colonnes d'erreurs : Signalant s'il y a eu une erreur fonctionnelle ou technique, ou si les documents ont été abandonnés et sont à relancer.

La table se met à jour en continu pour fournir un suivi en temps réel des flux grâce aux différentes applications vues jusqu'ici. Chaque application informe de l'avancée ou d'une erreur d'un document et le renseigne dans SuiviNum. En cas de plantage d'application ou d'erreur d'accès à un fichier, elle place le document en état 7 « KO Technique ». L'équipe fonctionnelle, le BEA, pour Bureau Études et accompagnement, peut alors le relancer.

Concernant les erreurs, nous avons identifié deux types principaux : l'erreur fonctionnelle et l'erreur technique. La première provient de la prestation elle-même, tandis que la seconde est une erreur interne pouvant avoir diverses origines : une donnée mal renseignée, un « plantage » informatique, etc. Les erreurs fréquentes sont souvent liées à un code langue incorrect, comme mentionné précédemment. Un autre problème rencontré est la « fausse erreur », une erreur fonctionnelle qui n'en est pas vraiment une, souvent à cause de crash serveurs.

SuiviNum s'avère également être un outil précieux pour les différents départements de la BnF. Comme évoqué plus haut, il permet d'avoir une visualisation graphique en temps réel des flux, offrant ainsi la possibilité d'obtenir des chiffres sur ce qui est réalisé et ce qui reste à faire. Par exemple, on peut suivre le nombre de pages, équivalent A4, numérisées au cours de

l'année[9], ou encore obtenir des statistiques sur les x derniers jours par type de documents[10]. Ces informations permettent aux chefs de projets d'avoir une idée précise de l'avancement des différentes prestations dont ils ont la charge. C'est également un outil de reporting indispensable pour l'équipe de numérisation de la BnF.

Le réel défi de la BnF

Le numérique est désormais omniprésent dans notre quotidien. Ce que les auteurs de science-fiction prédisaient ou imaginaient il y a quelques décennies est aujourd'hui une réalité : l'intelligence artificielle fait partie intégrante de nos vies. Cependant, cette révolution technologique apporte son lot de défis, notamment en termes de gestion des volumes de données, comme l'illustre l'exploration de la chaîne de traitement OCR.

Le premier défi est d'ordre physique. Les institutions patrimoniales comme la BnF sont confrontées depuis longtemps au stockage d'objets matériels tels que les livres et les manuscrits. Aujourd'hui, elles doivent également gérer le stockage de l'immatériel : données, fichiers et images numériques. Cette nouvelle réalité nécessite la mise en place de data centers, véritables centres névralgiques où sont stockés les paquets OCR, les œuvres numérisées et les données associées. Ces installations requièrent un espace conséquent pour accueillir des racks (ou baies), sortes d'armoires renfermant les appareils de stockage et de réseau. Disposer d'une capacité de stockage de l'ordre du pétaoctet (PB) ou de l'exaoctet (EB) - sachant qu'un EB équivaut à 1000 PB -

[9] cf annexe fig 3.

[10] cf annexe fig 4.

implique des investissements importants en termes d'espace et d'infrastructure. S'ajoutent à cela les coûts financiers liés à l'entretien des serveurs, à la gestion des racks et à leur consommation énergétique.

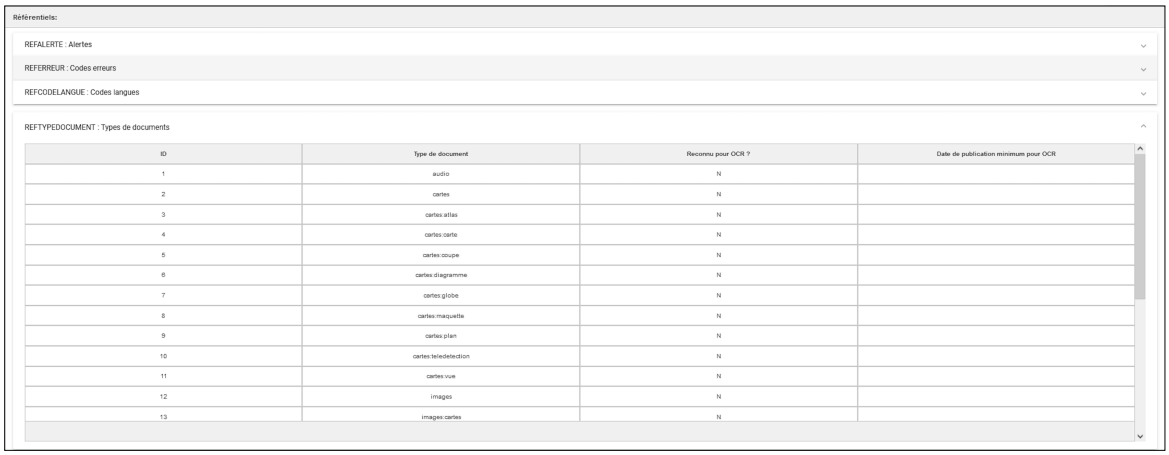
Le second défi concerne la gestion des flux de données. La capacité de stockage de la BnF, bien que considérable, n'est pas illimitée. Il est donc crucial d'optimiser la gestion des flux, tant pour la chaîne de traitement OCR que pour la chaîne d'entrée des documents. L'objectif est de maintenir un fonctionnement automatique et continu des chaînes, en s'assurant qu'elles aient toujours des éléments à traiter. Dans ce contexte, l'OCR joue un rôle de variable d'ajustement. La production quotidienne d'OCR est adaptée en fonction du volume de nouvelles livraisons : elle est augmentée lorsque les entrées sont faibles et réduite en période de forte affluence de nouveaux documents.

Un problème majeur a récemment mis en lumière la complexité de cette gestion des flux. En effet, un même document numérique passe d'abord par la chaîne d'entrée pour ses images, puis une seconde fois pour l'OCR interne. Cette double entrée a provoqué une saturation de la chaîne d'entrée, bloquant tout le circuit d'océrisation interne pendant plus d'un an.

Pour donner une idée de l'ampleur de l'infrastructure nécessaire, la chaîne d'entrée utilise à elle seule 22 serveurs dédiés. Le système SPAR fonctionne sur environ 50 machines virtuelles (VM). L'OCR, quant à lui, mobilise 4 machines dédiées uniquement au calcul. Actuellement, SPAR gère un volume de stockage de 6 pétaoctets, multiplié par 4 car les données sont stockées sur deux zones différentes pour des raisons de sécurité et de redondance.

Cette situation illustre parfaitement les défis techniques et logistiques auxquels font face les grandes institutions culturelles à l'ère du numérique. La gestion de tels volumes de données requiert non seulement des infrastructures conséquentes, mais aussi une planification minutieuse et une capacité d'adaptation constante face aux imprévus.

Annexe



Référentiels:

REFALERTE : Alertes

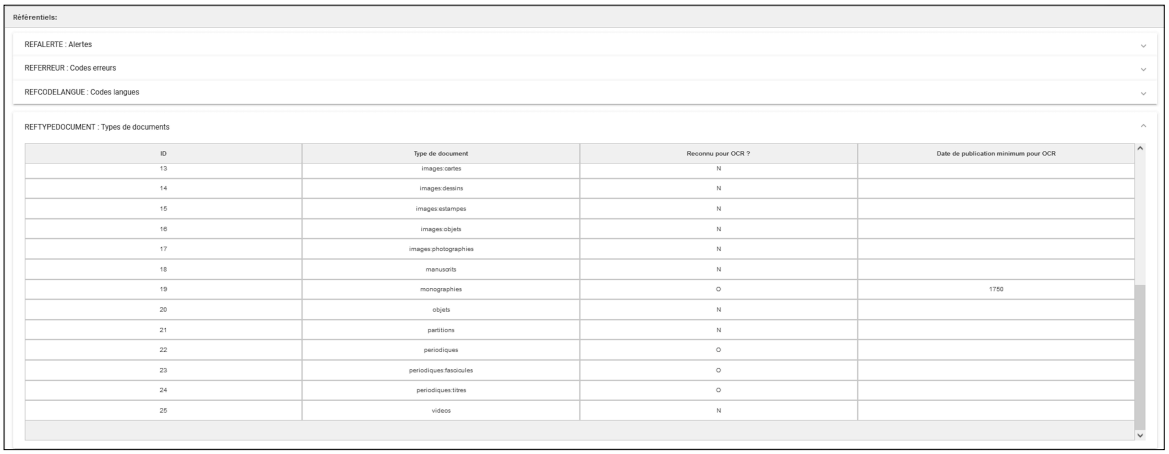
REFERREUR : Codes erreurs

REFCODELANGUE : Codes langues

REFTYPEDOCUMENT : Types de documents

ID	Type de document	Reconnu pour OCR ?	Date de publication minimum pour OCR
1	audio	N	
2	cartes	N	
3	cartes:atlas	N	
4	cartes:carte	N	
5	cartes:coupe	N	
6	cartes:diagramme	N	
7	cartes:globe	N	
8	cartes:maquette	N	
9	cartes:plan	N	
10	cartes:teledetection	N	
11	cartes:vue	N	
12	images	N	
13	images:cartes	N	

Fig 1a : Table des référentiels, aperçu de l'application SuiviNum. Types de documents acceptés ou rejetés pour l'océrisation.



Référentiels:

REFALERTE : Alertes

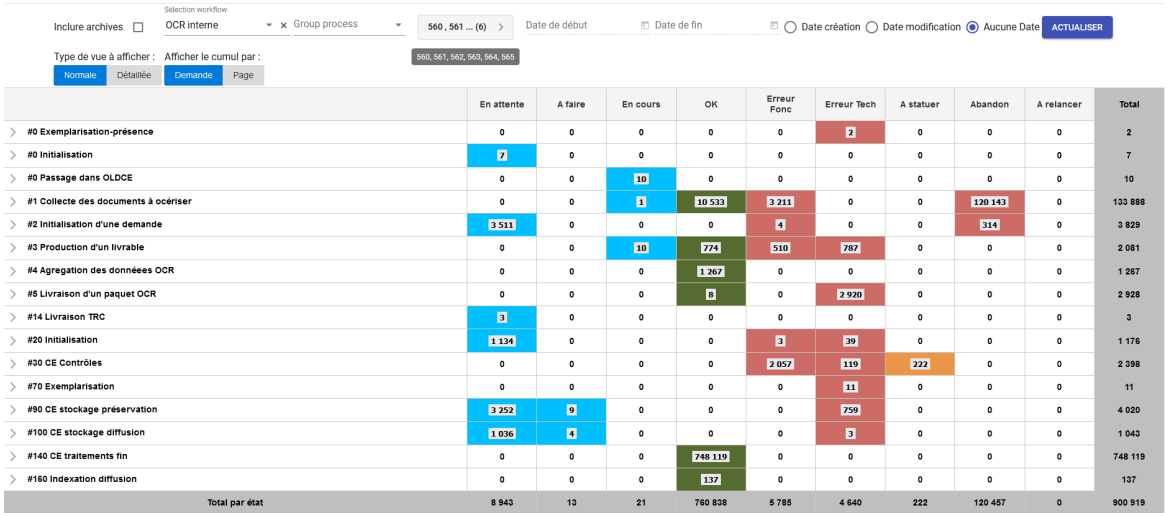
REFERREUR : Codes erreurs

REFCODELANGUE : Codes langues

REFTYPEDOCUMENT : Types de documents

ID	Type de document	Reconnu pour OCR ?	Date de publication minimum pour OCR
13	images:cartes	N	
14	images:dessins	N	
15	images:estampes	N	
16	images:objets	N	
17	images:photographies	N	
18	manuscrits	N	
19	monographies	O	1750
20	objets	N	
21	partitions	N	
22	periodiques	O	
23	periodiques:fascicules	O	
24	periodiques:titres	O	
25	videos	N	

Fig 1b : Suite de la table des référentiels des types de documents. Ici, les documents de types monographies, à partir de 1750, et les périodiques sont autorisés pour océrisation.



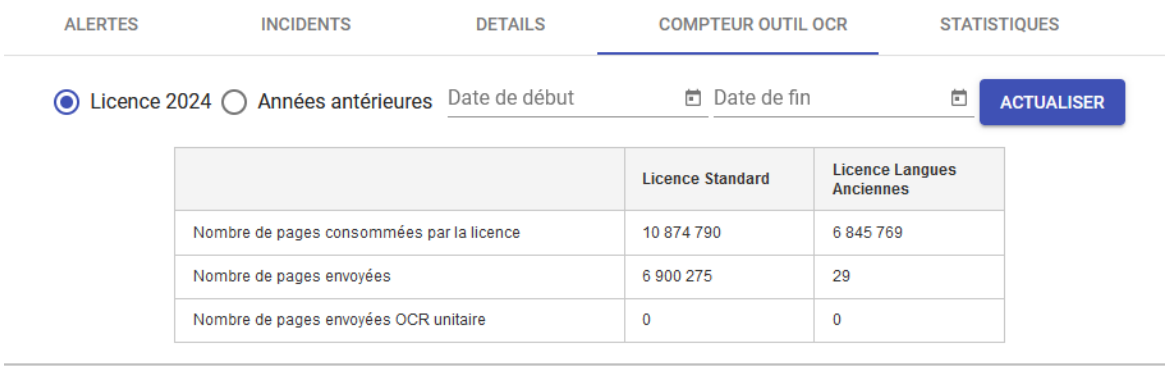
Inclure archives | Sélection workflow: OCR interne | Group process | 560 , 561 ... (6) | Date de début | Date de fin | Date création | Date modification | Aucune Date | ACTUALISER

Type de vue à afficher : Normale | Détaillée — Afficher le cumul par : Demande | Page

560, 561, 562, 563, 564, 565

	En attente	A faire	En cours	OK	Erreur Fonc	Erreur Tech	A statuer	Abandon	A relancer	Total
#0 Exemplarisation-présence	0	0	0	0	0	2	0	0	0	2
#0 Initialisation	7	0	0	0	0	0	0	0	0	7
#0 Passage dans OLDCE	0	0	10	0	0	0	0	0	0	10
#1 Collecte des documents à océriser	0	0	1	10 533	3 211	0	0	120 143	0	133 888
#2 Initialisation d'une demande	3 511	0	0	0	4	0	0	314	0	3 829
#3 Production d'un livrable	0	0	10	774	510	787	0	0	0	2 081
#4 Agregation des donnéees OCR	0	0	0	1 267	0	0	0	0	0	1 267
#5 Livraison d'un paquet OCR	0	0	0	8	0	2 920	0	0	0	2 928
#14 Livraison TRC	3	0	0	0	0	0	0	0	0	3
#20 Initialisation	1 134	0	0	0	3	39	0	0	0	1 176
#30 CE Contrôles	0	0	0	0	2 057	119	222	0	0	2 398
#70 Exemplarisation	0	0	0	0	0	11	0	0	0	11
#90 CE stockage préservation	3 252	9	0	0	0	759	0	0	0	4 020
#100 CE stockage diffusion	1 036	4	0	0	0	3	0	0	0	1 043
#140 CE traitements fin	0	0	0	748 119	0	0	0	0	0	748 119
#160 Indexation diffusion	0	0	0	137	0	0	0	0	0	137
Total par état	8 943	13	21	760 838	5 785	4 640	222	120 457	0	900 919

Fig 2 : Table de suivi des chaînes de traitement de l'application SuiviNum. Ici la chaine OCR.



ALERTES | INCIDENTS | DETAILS | COMPTEUR OUTIL OCR | STATISTIQUES

Licence 2024 | Années antérieures | Date de début | Date de fin | ACTUALISER

	Licence Standard	Licence Langues Anciennes
Nombre de pages consommées par la licence	10 874 790	6 845 769
Nombre de pages envoyées	6 900 275	29
Nombre de pages envoyées OCR unitaire	0	0

Fig 3 : Compteur outil OCR de SuiviNum.

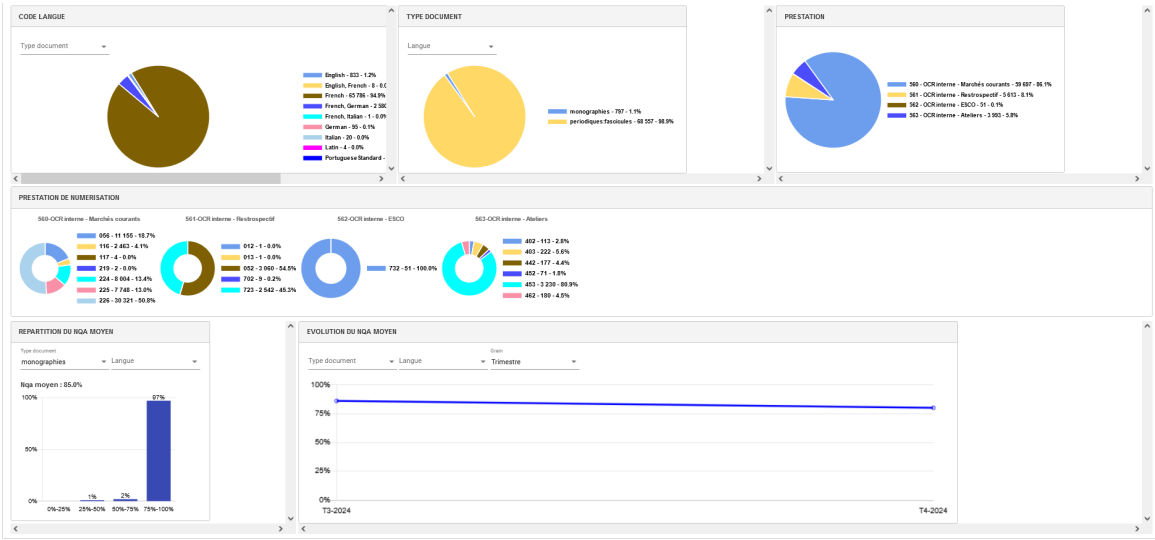


Fig 4 : Statistiques sur 30 jours des documents océrisés avec la langue, le type et la prestation.

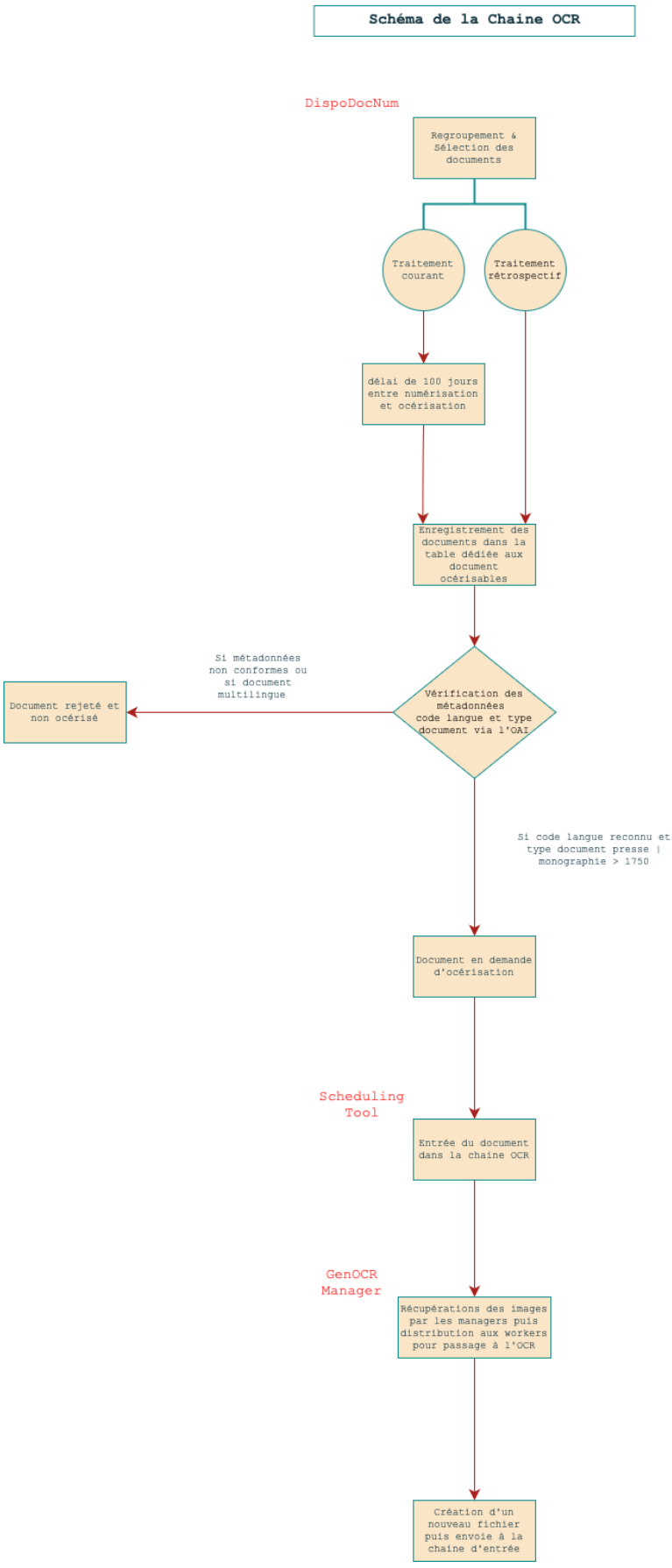


Fig A : Schéma de la chaine OCR

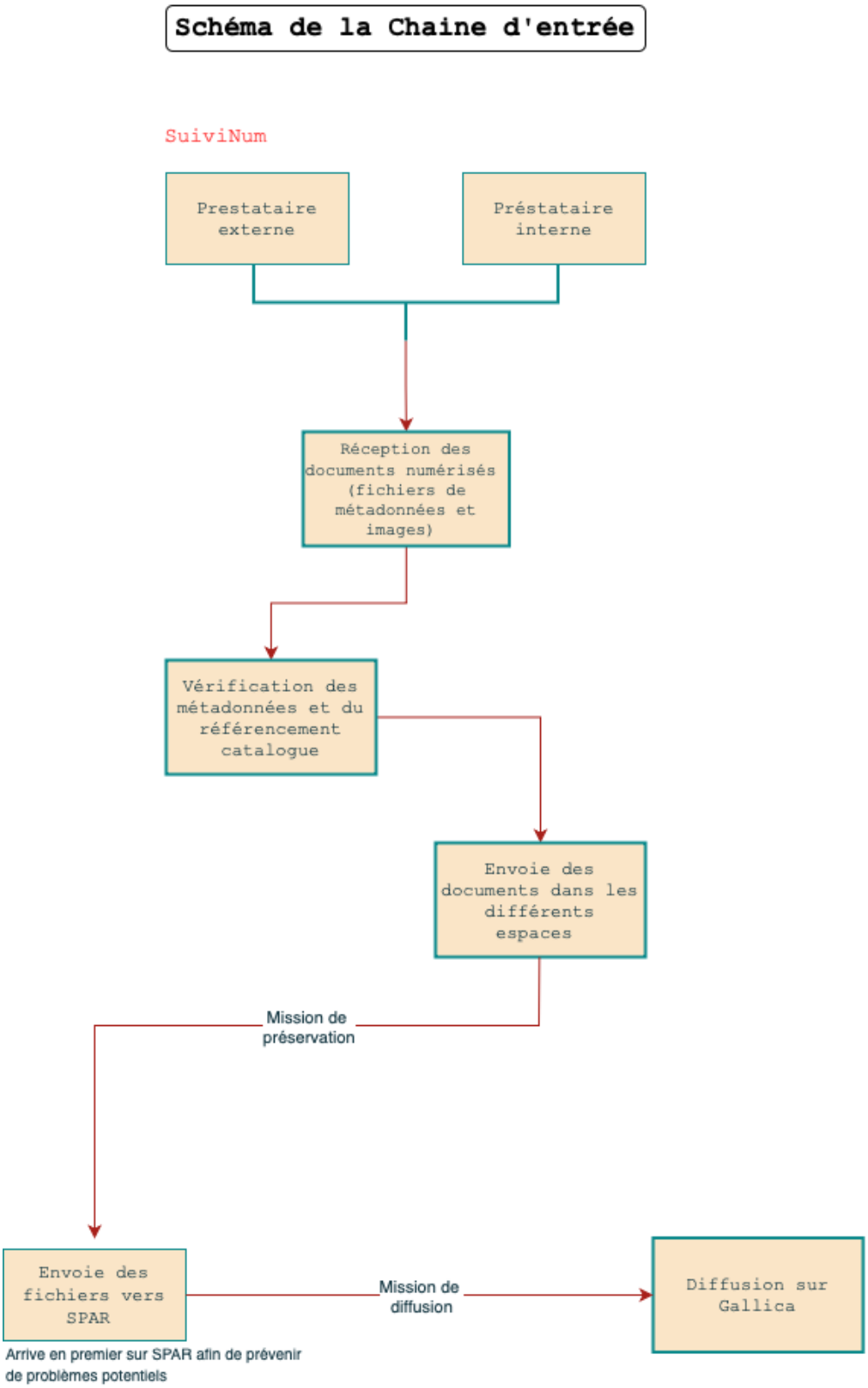


Fig B : Schéma de la Chaine d'entrée des documents numérisés