les **bons outils** pour traiter un **vrac numérique**

Développés par des acteurs privés, mais aussi par la puissance publique, les outils dédiés au traitement des vracs numériques intègrent progressivement l'IA pour améliorer leurs performances.

Dresser la liste des risques liés aux vracs numériques ressemble à un inventaire sans fin : non-conformité, responsabilité en cas de fuite des données, infraction envers le RGPD, difficultés d'accessibilité aux fichiers, baisse de la productivité, impact environnemental… De tels enjeux ne sont pas passés inaperçus : plusieurs éditeurs se sont positionnés sur le marché des solutions dédiées au traitement des vracs numériques.

Parmi eux, on trouve des acteurs privés bien connus dans le domaine du traitement documentaire, mais il faut aussi compter sur la présence de la puissance publique, qui propose, depuis plusieurs années, ses propres solutions développées par le Service interministériel des archives de France (Siaf). Deux types d'éditeurs et deux philosophies se distinguent : les premiers développent des solutions propriétaires, alors que les seconds recourent à l'open source.

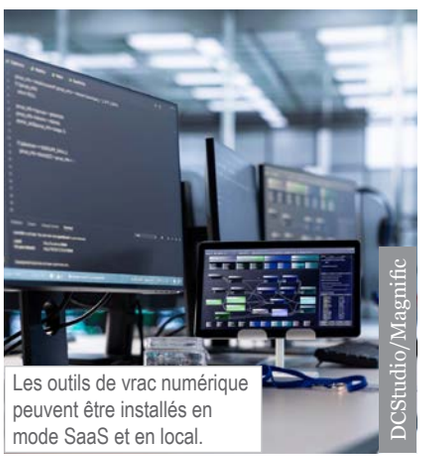
Le tableau comparatif *(pages suivantes)* fait apparaître une dizaine d'outils. Un premier constat s'impose : contrairement à d'autres logiciels qui sont désormais uniquement proposés en mode SaaS, ces outils offrent plus de souplesse et peuvent être installés localement *(on-premise)*.

identifier les fichiers en doublons

Plusieurs dizaines de critères ont été retenus pour mesurer l'étendue des capacités des différentes solutions. Mais comment s'y retrouver ? *« Dans un premier temps, il convient de prioriser des fonctionnalités permettant de réaliser des actions à mener avant l'entrée des archives dans le système d'archivage électronique (SAE) »*, souligne-t-on au Siaf. *« Il s'agit par exemple de la possibilité d'identifier les fichiers en doublons ou de tagger les archives éliminables et de les supprimer au cours des travaux de traitement, afin de ne pas prendre en charge des archives dont la DUA (1) serait échue. »* Selon le Siaf, l'identification de formats non reconnus permet également de réaliser des migrations avant l'entrée dans le SAE, conformément à la stratégie de pérennisation de l'institution. *« Dans un deuxième temps, l'ajout de métadonnées de gestion est structurant pour la gestion du cycle de vie des archives dans le SAE. Enfin, les métadonnées de description sont intéressantes et à utiliser aux niveaux les plus appropriés. »*

intégration progressive de l'IA

Du côté de l'éditeur Office Gemini, qui commercialise la solution Dokmee, on met en avant l'infrastructure elle-même : *« Quand une entreprise veut se doter d'une solution de gestion de vracs numériques, la première question à se poser est celle de la sensibilité des données »*. Selon la réponse, on optera pour un hébergement cloud, sur site ou hybride. *« Ensuite, il faut regarder de près les fonctionnalités indispensables : OCR, automatisation des processus, intégration avec les outils existants ou encore la conformité réglementaire »*, affirme l'éditeur. *« Et surtout, ne pas oublier l'humain : une solution, aussi performante soit-elle, ne vaut rien si les équipes ne savent pas l'utiliser. La formation et l'ergonomie sont donc tout aussi cruciales que la technique. »*


DCStudio/Magnific

Les outils de vrac numérique peuvent être installés en mode SaaS et en local.

Sans surprise, certains éditeurs recourent désormais à l'IA pour améliorer les performances de plusieurs fonctionnalités, comme l'ajout manuel d'éléments aux lots importés, l'identification de répertoires semblables ou l'édition de rapports d'analyse, par exemple. Et ce n'est qu'un début. Plusieurs éditeurs indiquent en effet travailler à l'intégration de l'IA dans un nombre toujours plus grand de fonctionnalités : création de nouveaux répertoires, normalisation en masse des noms de fichier, enrichissement de la description, entre autres. Quant aux IA utilisées par les éditeurs, quelques noms bien connus apparaissent : OpenAI, Mistral, Claude, Gemini…

Au-delà des éditeurs, les intégrateurs apparaissent également comme des acteurs du vrac numérique *(voir article 37)*. ■

(1) DUA : durée d'utilité administrative.

Bruno Texier

Panorama des solutions de traitement de vracs numériques

● oui ✖ non

PRÉSENTATION	Calico France (JLB Informatique)	ELP - Édition de Logiciels Professionnels	EverteamSoftware
Année de création	1982	1997	1990
Nom de la solution	Gedzilla	Hyperdhoc Archive	Everteam.discover
Mode d'accès	SaaS	SaaS, On-premise	SaaS, hébergé, On-premise
Solution open source	✖	✖	✖
FONCTIONNALITÉS			
Import d'un ensemble de fichiers	● (avec IA)	● (sans IA)	● (avec IA - File System, M365-SharePoint, M365-Exchange, S3, Google Drive, Google Mail, base de données, données structurées)
Import des contenus de messagerie	● (avec IA)	✖	● (avec IA - e-mail, pièces jointes)
Ajout manuel d'éléments aux lots importés	● (avec IA)	● (sans IA)	● (sans IA)
Navigation dans l'arborescence	● (avec IA)	● (sans IA)	● (avec IA - arborescence dynamique)
Identification de répertoires semblables	● (avec IA)	● (sans IA)	● (sans IA - identification des répertoires identiques ou vides).
Identification de fichiers vides	Sur la feuille de route	✖	● (sans IA - fichiers vides ou répertoires vides)
Ouverture des documents	● (sans IA)	● (sans IA)	●
Alerte sur les formats non identifiés	● (sans IA)	● (sans IA)	● (avec IA - par extraction de métadonnées)
Renommage d'un intitulé	● (avec IA)	● (sans IA)	● (sans IA - renommage, chercher/remplacer, suppression des caractères spéciaux)
Normalisation en masse des noms de fichier	Sur la feuille de route	● (sans IA)	● (avec IA - renommage, chercher/remplacer, suppression des caractères spéciaux)
Marquage des répertoires ou des fichiers à supprimer	● (sans IA)	● (sans IA)	● (avec IA)
Dédoublonnage automatique des objets identiques	Sur la feuille de route	✖	● (avec IA)
Création de nouveaux répertoires	Sur la feuille de route avec IA	● (sans IA)	✖
Transfert du contenu d'un répertoire dans un autre	● (avec IA)	● (sans IA)	● (avec IA - possibilité de déplacer ou de copier)
Taggage des éléments	● (avec IA)	✖	● (avec IA)
Enrichissement de la description	● (avec IA)	● (sans IA)	● (avec IA - possibilité de créer des métadonnées personnalisées)
Enrichissement par un tiers	● (sans IA)	● (sans IA)	● (avec IA - via API ou synchronisation via connecteurs dédiés)
Visualisation des informations relatives à l'import	● (sans IA)	● (sans IA)	● (sans IA - interface de suivi et de supervision)
Édition de rapport d'analyse	✖	●	● (avec IA - tableau de bord personnalisable via l'usage du module SEDA everteam.archive ou autre)
Production du SIP au format SEDA 2.1	✖	✖	● (sans IA - via l'usage du module SEDA everteam.archive ou autre)
Modèle IA utilisé	OpenAI/LM-Kit	NC	●
DIVERS			
Références	Prea Gianca, Lidle CSE, Happyvet, Intercarat, Artebat	Bouygues Immobilier, Mairie du Lamentin, Mairie de Ste Rose, Agglomération de Dreux, Securdhoc	Conseil économique et Social, Conseil régional d'Île-de-France, Fred, Bouygues Construction, SDED
Coût d'un projet (à partir de...)	39.90 € HT/mois	1 600 € HT (hors développement, paramétrage, formation, coût spécifique, etc.)	10K €

● oui ✖ non

Hyland	Office Gemini	Programme Vitam (ministère de la Culture)	Programme Vitam (ministère de la Culture)
1991	2006	2018	2015
Nuxeo	Dokmee	Archifiltre	Vitam, Module de collecte
Sur site / PaaS	SaaS, On-premise, cloud	Outil bureautique	SaaS (VaS), On-premise
●	✖	●	●
● (avec IA)	● (sans IA)	● (sans IA)	● (IA sur la feuille de route)
● (avec IA)	● (sans IA)	✖ (IA sur la feuille de route)	✖ (IA sur la feuille de route - pas de traitement spécifique, mais format «décompressé» en sortie de ReSip et format conteneur type «pst» accepté sans traitement spécifique type «décompression» pour ce dernier)
● (avec IA)	● (sans IA)	● (sans IA)	●
● (avec IA)	● (sans IA)	● (IA sur la feuille de route)	● (IA sur la feuille de route)
● (avec IA)	Possible via developpement personnalisé	●	Sur la feuille de route
●	● (avec IA)	● (sans IA)	● (IA sur la feuille de route - pour identifier des fichiers «non vides» d'un point de vue «binaire» mais vide du point de vue «informationnel». Ex: ne contient que des espaces)
● (avec IA)	● (sans IA)	● (sans IA - à partir du système d'exploitation, pas de visionneuse)	● (IA sur la feuille de route - téléchargement et visualisation par les outils du poste de travail, pas de visionneuse)
● (avec IA)	Possible via developpement personnalisé	✖ (IA envisagée)	● (IA envisagée - identification de format via Siegfried et politique de gestion : liste blanche autorisée, liste noire interdite, acceptation ou rejet des formats inconnus / non identifiés)
● (avec IA)	● (sans IA)	● (avec IA)	● (IA sur la feuille de route)
● (avec IA)	● (sans IA)	✖ (IA envisagée)	● (IA sur la feuille de route - partiellement. À date, contrôle des caractères utilisés dans le nommage des fichers et par extration et réimport après correction d'une série. Par API, possibilité d'appliquer des RegEx et des modifications unitaires «en masse»)
● (avec IA)	● (sans IA)	● (sans IA)	● (IA sur la feuille de route - marquages des erreurs liées à des traitements de contrôles, tels que présence de virus, non respect de la politique de formats)
● (avec IA)	● (sans IA)	✖ (IA envisagée)	Sur la feuille de route avec IA (pour dédoublonner des fichiers au contenu informationnel identique mais dont les empreintes des fichiers sont différentes)
● (avec IA)	● (sans IA)	Sur la feuille de route	● (création par import)
● (avec IA)	● (sans IA)	✖	●
● (avec IA)	● (sans IA)	● (IA envisagée)	● (IA sur la feuille de route - configurable sur la base de script de traitement des MD (JSLT), taggage de métadonnées techniques : format identifié, volume binaire...)
● (avec IA)	● (avec IA - via AI summary)	● (IA prévue - V5)	● (IA sur la feuille de route)
● (avec IA)	● (IA possible via developpement personnalisé)	✖	● (IA sur la feuille de route - par API, par des profils d'utilisateurs disctincts. L'IA sera utilisée pour l'enrichissement des métadonnées)
●	●	● (sans IA - visualisation des docs et métadonnées extraites et de qqes métadonnées de réf.)	● (sans IA - visualisation des transactions identifiant des versements dans des projets de versement, visualisation des docs et métadonnées versées et extraites, SEDA ou non, visualisation et recherche des erreurs identifiées au niveau de chaque pièce quand c'est pertinant / possible)
● (avec IA)	● (sans IA)	● (IA envisagée - édition d'un rapport d'audit)	Partiellement (possibilité d'extraire une description des archives «en erreur» au format CSV)
Via partenaires	● (sans IA)	✖ (IA envisagée pour l'enrichissement des métadonnées)	● (IA sur la feuille de route pour l'enrichissement des métadonnées - en SEDA 2.1, 2.2 et 2.3)
Interne	Utilisation possible de Gemini, Claude et Grok. Autres modèles d'IA en cours d'intégration	Standard (OpenAI) + agentique	Standard (OpenAI) + agentique
L'Oréal, Renault, DILA, Société des Grands Projets	Astro Holding, Sherring Williams, Centre de recherche français	NC	Mission archives du min. de la Culture, Communauté d'agglomération de Cergy-Pontoise, Université de Lille, Cines
90K €	39 € /utilisateur flottant/mois	Pas de coût	Pas de coût (logiciel on-premise, utilisateur VaS, membre du Club utilisateurs)

Panorama des solutions de traitement de vracs numériques

● oui ✖ non

PRÉSENTATION	Programme Vitam (ministère de la Culture)	Service interministériel des Archives de France (SIAF) - ministère de la culture	Spark Archives
Année de création	2015	2019	2012
Nom de la solution	Vitam, ReSip	Octave, basé sur Docuteam Packer de la société Docuteam	Zéro Vrac
Mode d'accès	Outil bureautique	Outil bureautique	SaaS
Solution open source	●	●	✖
FONCTIONNALITÉS			
Import d'un ensemble de fichiers	● (sans IA)	● (sans IA)	● (sans IA)
Import des contenus de messagerie	● (sans IA - format pst (et autres) décompressés par ReSip en une organisation de répertoires et d'archives représentant chaque message et PJ)	✖	● (sans IA)
Ajout manuel d'éléments aux lots importés	● (sans IA)	● (sans IA)	● (sans IA)
Navigation dans l'arborescence	● (sans IA)	● (sans IA)	✖
Identification de répertoires semblables	● (sans IA - fichier semblable uniquement, pas répertoire)	● (sans IA - fichier semblable uniquement, pas répertoire)	✖
Identification de fichiers vides	● (sans IA)	✖	✖
Ouverture des documents	● (sans IA - téléchargement et visualisation par les outils du poste de travail, pas de visionneuse)	● (sans IA - visualisation dans l'outil)	● (sans IA)
Alerte sur les formats non identifiés	● (sans IA - identification de format via DROID)	● (sans IA - identification de format via DROID)	✖
Renommage d'un intitulé	● (sans IA)	● (sans IA)	● (sans IA)
Normalisation en masse des noms de fichier	● (sans IA - possibilité de renommer les intitulés - en tant que métadonnées - via l'import et le réimport d'un fichier CSV)	● (sans IA)	● (sans IA)
Marquage des répertoires ou des fichiers à supprimer	✖ (suppression directe des fichiers importés dans l'outil)	● (sans IA)	● (sans IA)
Dédoublonnage automatique des objets identiques	● (sans IA - dédoublonnage après action manuelle de l'utilisateur)	● (sans IA)	● (sans IA)
Création de nouveaux répertoires	● (sans IA)	● (sans IA)	✖
Transfert du contenu d'un répertoire dans un autre	● (sans IA)	● (sans IA)	● (sans IA)
Taggage des éléments	● (sans IA)	✖	● (sans IA)
Enrichissement de la description	● (sans IA)	● (sans IA)	● (sans IA)
Enrichissement par un tiers	✖	✖	● (sans IA)
Visualisation des informations relatives à l'import	● (sans IA - visualisation des documents et des métadonnées versées et extraites, SEDA ou non)	● (sans IA - information sur les anomalies relatives aux formats des fichiers via pop-up)	● (sans IA)
Édition de rapport d'analyse	NC	● (sans IA - rapport d'éxécution sur les anomalies de formats des fichiers)	● (sans IA)
Production du SIP au format SEDA 2.1	● (sans IA - en SEDA 2.1, 2.2 et 2.3)	● (sans IA - en SEDA 2.1, 2.2 et 2.3)	✖
Modèle IA utilisé	✖	Sans objet	Mistral
DIVERS			
Références	Archives départementales de la Gironde, mission archives des ministères sociaux, ARS, Commissariat à l'énergie atomique	42 services d'archives départementales, 46 services d'archives communales et intercommunales	NC
Coût d'un projet (à partir de...)	Pas de coût	Pas de coût	NC

vrac numérique : les **intégrateurs**, **intermédiaires** entre éditeurs et clients

Les intégrateurs ont aussi leur rôle à jouer dans le déploiement d'une solution dédiée au traitement du vrac numérique. Objectif : adapter des logiciels ou des briques de solutions à la réalité technique et humaine du client.

Au-delà des éditeurs présents dans notre tableau comparatif *(pages précédentes)*, les intégrateurs apparaissent également comme des acteurs du traitement des vracs numériques. Car si les éditeurs conçoivent la solution technologique, les intégrateurs interviennent pour adapter le logiciel de l'éditeur à la réalité technique et humaine de leurs clients.

« En matière de vrac numérique, la valeur ajoutée d'un intégrateur est avant tout méthodologique », confirme Thierry Georges, directeur de l'entité ECM au sein de Coexya. *« L'intégrateur joue le rôle d'intermédiaire entre l'éditeur et l'utilisateur : il met en œuvre et assemble les meilleures briques du marché afin de trier, classer et restructurer durablement l'information. »*

Dans un premier temps, l'intégrateur procède à un audit documentaire approfondi (volumes, typologies, droits d'accès…) qui va conditionner le choix des briques logicielles préconisées. Ce diagnostic permet également d'identifier les doublons et documents obsolètes afin d'orienter chaque document vers le sort final approprié : Ged, SAE, mais aussi potentiellement serveurs de fichier, applications métiers ou bien destruction. En résumé, l'intégrateur permet de définir une stratégie de traitement adaptée au contexte (volumétrie, criticité des données, usages), ainsi que des règles de gestion réutilisables pour industrialiser les opérations.

gouvernance documentaire

L'intégrateur configure ensuite la connexion automatisée à ces flux documentaires vers les cibles de stockage, puis accompagne la transformation des pratiques des utilisateurs afin d'éviter que ce vrac ne soit reconstitué sur le plus long terme.

« Dans cette perspective, l'intégrateur met en œuvre des traitements en masse et peut déployer des mécanismes de versement automatisés vers les systèmes cibles afin de fiabiliser et de pérenniser les flux », indique Thierry Georges. *« Il contribue également à structurer une gouvernance documentaire en formalisant des règles de classement, de nommage et de conservation, permettant d'inscrire les actions dans la durée. »*

traiter les vracs avec l'IA

En matière de traitement de vracs numériques, les apports de l'IA sont nombreux. Selon Marine Césaire, consultante fonctionnelle chez Coexya, *« l'IA augmente la pertinence de l'indexation et de la classification automatique des documents, qui facilitent ensuite leur traitement et leur classement appropriés »*. Elle permet notamment d'améliorer la compréhension du contenu des documents non structurés (texte libre, PDF, scans) en identifiant automatiquement des typologies documentaires, des informations clés ou des thématiques métier, ce qui limite les traitements manuels.

L'IA est aussi mise à contribution pour étendre le périmètre des documents susceptibles d'être traités automatiquement et efficacement au sein d'un vrac numérique. *« Cette technologie est particulièrement utile pour traiter des volumes importants de données hétérogènes, pour lesquels les approches classiques montrent leurs limites, notamment en matière de classification par le sens ou d'enrichissement automatique de métadonnées »*, indique Marine Césaire.

L'intégrateur attire cependant l'attention sur le coût énergétique du recours à l'intelligence artificielle. Celle-ci doit être réservée aux documents les plus complexes, tandis que les traitements plus classiques et moins énergivores (basés sur des métadonnées, sur des traitements de Lad/Rad ou bien sur des expressions régulières) doivent être appliqués en priorité aux documents les plus simples.

À noter que certains éditeurs de solutions de traitement de vracs numériques disposent d'une branche *« services »* et intègrent eux-mêmes leur propre solution. ■

Bruno Texier

repères

Intégrateurs Ged/ECM/SAE/ records management présents sur le segment du traitement de vracs numériques :

- Access Group
- Coexya
- Eliadis
- Exakis Nelite
- Infotel Software
- Smile
- Xelians