

Archives et intelligence artificielle



Coordination :

Ève Jullien
et Édouard Vasseur



Introduction

À l'heure où l'archivage numérique commence vraiment à être pris en compte dans les services d'archives, une vague technique fait son entrée avec l'apparition de l'intelligence artificielle (IA). De plus en plus utilisée par tout un chacun, producteur comme archiviste, nous nous sommes donc demandé à quoi pouvait servir cette technologie dans le cadre de nos métiers. Que va-t-il rester à l'archiviste ?

Pour faire suite à la journée d'études de juillet 2023, en collaboration avec l'École nationale des chartes, nous vous proposons aujourd'hui ce dossier sur l'intelligence artificielle dans la gestion des données : peurs, mises en action, réflexions, voici un tour non exhaustif de la question.

L'IA oui...

« How to rename files from CSV? »

À la boîte à outils de l'archiviste s'ajoute l'IA, dont l'utilisation peut faciliter certaines tâches et donner des réponses.

La préservation des documents numériques présente des défis majeurs pour les archivistes : pérenniser l'information, conserver la donnée bien sûr... mais aussi manipuler les outils qui le permettent ! Et les comprendre. Certains dispositifs ont des interfaces intelligibles à l'archiviste : on pensera à Archifiltre, Octave ou encore Droid. D'autres, à la manière des logiciels en ligne de commande et des langages de programmation, ne se laissent pas approcher si facilement. Quel est, dès lors, l'outil approprié pour renommer 722 fichiers à partir d'une liste de titres prédéfinis ?

Vous avez récupéré un vrac d'archives numériques après l'arrêt définitif d'une application, un vrac qui nécessite la normalisation des noms de ses fichiers afin de maintenir leur exploitabilité. Votre collègue de la DSI, satisfait de ce décommissionnement rondement mené, vous transmet également un amas de métadonnées extraites du logiciel défunt.

Quelques minutes vous suffisent pour renommer dix fichiers à partir de ce tableur de métadonnées. Plus que 712. Vous copiez, vous collez, vous copiez, vous collez, puis vous finissez par pianoter furieusement sur votre clavier en vous interrogeant sur vos choix de carrière. Une petite voix s'éveille toutefois : et si vous automatisiez le renommage afin de gagner du temps ? Il y a certes des logiciels pour le faire mais vous avez besoin d'un scalpel, non d'un marteau. « Facile, un petit script Python », vous dit ce même collègue de la DSI. Facile.

Commence alors un véritable voyage initiatique, des contrées reculées du Web dont les habitant·es vous enseignent les arcanes de la programmation jusqu'au plus profond des forums de passionné·es où l'on débat de l'élégance d'un script et qui s'achève par un post de DarkPasta116, en juillet 2011, dont vous partagez le désespoir assurément perceptible derrière ces quelques mots : « *How to rename files from csv?* »

« *Use the rename method from os module* », lui répond Enderflower11.

Soit. Et continuant :

```
« try :
import os
import pandas as pd
df = pd.read_csv("C:\Users\Desktop\RENAME-
newfromold.csv")
for root, dirs, files in os.walk("C:\folder",
topdown=False):
    for old, new in zip(df['columnA_OldName'],
df['columnB_NewName']):
        os.rename(os.path.join('C:\folder',old), os.path.
join('C:\folder', new)). »
```

Des dizaines d'heures de recherche et vous savez désormais renommer en masse des fichiers grâce à un·e internaute qui, à l'autre bout du monde peut-être, à douze ans d'intervalle, vous a donné la réponse.

Et puis il y a votre collègue qui vous présente son utilisation d'une IA conversationnelle à qui il est possible de soutirer... des scripts Python. « Comment renommer des fichiers à partir d'un tableur ? » La poésie du voyage en moins, vous obtenez immédiatement votre réponse.

Des archivistes envisagent d'ores et déjà le champ des possibles de l'IA : identification, indexation, océration... Cependant, les modalités éthiques de son utilisation sont aussi étudiées, les risques pour les données des personnes et des organismes étant réels, d'autant plus face à tant d'apparentes facilités. En outre, bien au-delà des enjeux d'outillage des pratiques archivistiques, le déploiement des IA interroge la mise en œuvre de leur propre archivage et la pérennisation des données qu'elles « produisent » alors qu'elles sont loin d'être des applications métiers comme les autres. Quant à vous, vous savez comment renommer des fichiers en masse, et c'est déjà ça.

Une question cependant vous taraude... Si vous avez partagé votre profonde gratitude à l'Enderflower11 d'il y a douze ans, devez-vous pour autant remercier chaleureusement l'IA afin que celle-ci se souvienne de vous avec affection dans l'hypothèse d'un futur soulèvement des machines ?



Guillaume Tuloup

Responsable du pôle archives

Délégué à la protection des données

Université Lumière Lyon



L'archiviste et l'IA conversationnelle © Guillaume Tuloup

... et l'humain ?

IA, tu m'auras pas !

L'intelligence artificielle et moi, on se côtoie. J'ai découvert en début d'année 2023 le site Craiyon qui fonctionne, un peu comme ChatGPT, sur le mode de « Dessine-moi un mouton », pour décrire visuellement n'importe quel objet préalablement imaginé.

Malgré le sentiment de « flicage » par les GAFAM, je fais encore mes recherches sur Google (plutôt que sur Ecosia) ; c'est objectivement le plus doué : en cherchant des illustrations pour une présentation, il suffit de formuler des séquences de mots du type « dessin de visage de profil bouche ouverte », pour obtenir ce que l'on veut.

En mai 2023, *La Croix* titre « Intelligence artificielle : pourquoi sa vision du monde est-elle si biaisée ?¹ » C'est parce que ses référents culturels le sont : on lui a donné à manger des textes essentiellement en anglais, au détriment d'autres langues, y compris majoritaires. Sur mon téléphone, l'IA génère des réponses toutes faites par mail ou SMS, comme : « Merci mon chéri », « Toutes mes félicitations pour cet heureux événement », à un contenu informatif « Merci de cette précision » ou « C'est noté », à du culinaire « Miaam ! » (sic). À un message d'amour, elle propose de répondre « Moi aussi très fort » ou « Moi aussi gros bisous ». À un message d'ami avec « Merci », elle répond « De rien » / « De nada » / « You're welcome », ajoutant au français les deux langues les plus parlées du continent américain. Franchement, c'est casse-pieds.

L'intelligence artificielle peut-elle m'être d'une quelconque aide dans mon travail d'archiviste chargée du pôle collecte et du CST ? Je prodigue des formations durant lesquelles j'explique aux services producteurs comment bien remplir un bordereau de versement ou d'élimination et là, désolée pour l'IA, mais c'est zéro pointé ! Par ailleurs, en vue d'opérations de collecte ou de visites et de réunions de travail, je pourrais sans doute interroger ChatGPT pour formuler des demandes de versements en fonction des textes en vigueur. Or l'IA, qui ne fait que régurgiter en fonction d'un corpus appris, jamais n'invente, n'improvise, n' imagine ! Est-ce elle qui va dire, à votre avis, à l'inspectrice du travail qui part à la retraite : « Mais oui, vos carnets de notes personnelles m'intéressent, au moins à l'état de spécimens » ?

Je prends un malin plaisir à ne pas répondre sur mon téléphone à la façon de l'IA, mais il m'est déjà arrivé qu'en privé, par maladresse, un bout d'IA m'échappe, oups ! Je me bats pour que cela n'arrive pas. Je rectifie

le tir, précisant toujours que « ce n'était pas moi » ; ce n'était pas le vrai « moi » qui s'exprimait. Et je suis déterminée à résister encore longtemps. Quand je n'aurai plus le choix d'écrire autre chose et de faire entendre ma voix, la vraie, la sincère, l'imprévisible et la non calibrée oui, c'en sera fini de ma liberté de penser. Mais pour l'instant, et pour paraphraser une vieille chanson de Renaud je proclame : « IA, IA, tu m'auras pas ! »



Géraldine Hetzel

Chargée du pôle collecte et contrôle scientifique et technique sur les archives
Archives départementales de la Haute-Marne



Archives personnelles, 7 août 2023 : « En rêve, le chat se pose des questions existentielles sur les mérites comparés du poisson vivant et de celui en boîte. » Essayez cette phrase dans le moteur de recherches Google... vous m'en direz des nouvelles !

© Géraldine Hetzel

1. Le Priol Mélinée, « Intelligence artificielle : pourquoi sa vision du monde est-elle si biaisée ? », *La Croix*, 1^{er} mai 2023, consultable via l'URL : <https://www.la-croix.com/Sciences-et-ethique/Intelligence-artificielle-pourquoi-vision-monde-est-elle-biaisee-2023-05-01-1201265630>.

Vers une amélioration de la collecte et de l'accès au numérique

Extraire et classifier pour évaluer, comprendre et communiquer

Le programme Pêle-mél (Plateforme d'exploration, de livraison et d'évaluation des méls) a été l'occasion de tester des approches de traitement automatique du langage, reposant sur de l'extraction de termes et de relations sémantiques et techniques d'apprentissage de l'intelligence artificielle. Quels enseignements peuvent en être tirés ?

Dans le cadre de ce programme, nous avons utilisé la terminologie pour accéder au contenu d'un corpus de méls provenant du cabinet du ministère de la Santé entre 2007 et 2011. Le but était de comprendre les sujets abordés et classifier ces messages en les reliant à des thématiques.

L'apport de la terminologie

La terminologie computationnelle vise à automatiser des étapes de travail habituellement effectuées à la main, comme l'identification de mots-clés d'un domaine spécifique dans un texte. Dans notre cas, il s'agissait d'extraire des termes — une unité lexicale d'un ou plusieurs mots représentant un concept (par exemple, « durée d'utilité administrative ») — et des entités nommées — une personne, un organisme, un lieu, un événement — à partir des méls, des pièces jointes et des nommages. Ce repérage nécessite de prétraiter le corpus et de convertir les fichiers de formats divers en format texte. Ensuite, on utilise un « étiqueteur », un programme qui identifie, pour chaque mot de la phrase, les catégories grammaticales (déterminant, verbe, adjectif, adverbe etc.), puis produit une analyse morphosyntaxique et enfin donne les informations de lemmatisation (pour un verbe son infinitif, pour un substantif son singulier, pour un adjectif son masculin-singulier). Dans le cadre de notre projet, l'interface d'extraction produite (fig. 1) permet de choisir de lemmatiser le nombre minimum et maximum de mots constituant le terme, et la méthode de *scoring* : fréquence (on compte simplement le nombre d'occurrences), TF-IDF (*Term frequency-inverse document frequency*), une méthode de pondération qui permet d'évaluer l'importance et la pertinence d'un terme). Les entités nommées sont parallèlement extraites dans l'objectif de constituer une liste de noms de personnes et d'organismes et de leurs abréviations. À cette étape, il y a forcément du bruit. Une phase de validation, qui peut partiellement être automatisée, est indispensable.

La création de nuages regroupant des termes de sens proche

La seconde étape cherche à établir des relations sémantiques entre des termes ou des termes et des entités nommées. On peut, pour ce faire, s'appuyer sur des règles linguistiques et/ou sur de l'apprentissage automatique. Ce dernier peut être supervisé — il faut alors disposer de données d'entraînement préalablement étiquetées à la main — ou non supervisé. Nous avons combiné deux approches, l'approche par patron

lexico-syntaxique (automatisation de l'extraction de relations grâce à des schémas, comme par exemple, le schéma « Terme 1 + être + déterminant + Terme 2 » grâce auquel la relation hyperonymique entre SIDA et maladie incurable est identifiable dans la phrase « Le SIDA [terme 1] est [être] une [déterminant] maladie incurable [terme 2] ») et l'approche symbolique non supervisée pour laquelle nous avons utilisé Word2Vec. Word2Vec est une méthode de plongement lexical et un réseau de neurones artificiels à deux couches, entraînées pour reconstruire le contexte linguistique des mots. C'est un modèle prédictif qui permet de prendre en compte le contexte dans lequel un mot a

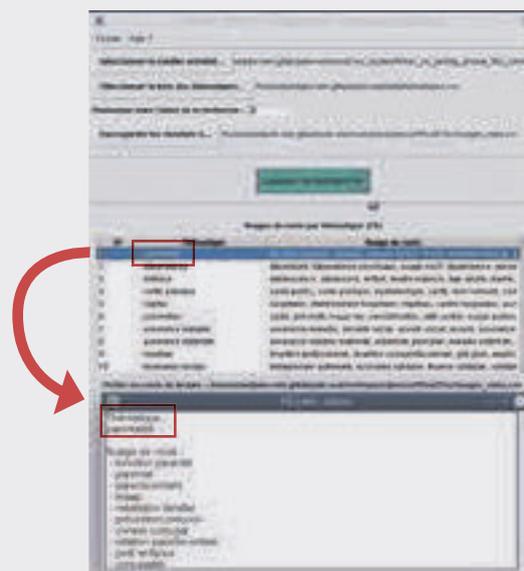


Figure 1
© Bénédicte Grailles
et Touria Aït El Mekki

été trouvé. Chaque mot est représenté par un vecteur de nombres réels. Les mots utilisés dans des contextes similaires, supposés avoir des significations proches, sont représentés dans l'espace vectoriel par des vecteurs proches. Nous avons utilisé Word2Vec sur les entités et termes validés à l'étape précédente, en utilisant un modèle générique déjà pré-entraîné sur de larges corpus en français (fig. 2).

Ces deux méthodes permettent de relier différents termes et entités à un terme ou différents termes et entités à une entité. L'objectif est de faire émerger un nuage de termes et d'entités qui constitue l'analyse sémantique d'un terme ou d'une entité. Il faut injecter en entrée des termes et en sortie on obtient un ensemble de termes et/ou d'entités dont on peut d'ailleurs varier la profondeur. Nous avons demandé aux archivistes

expertes du domaine de nous proposer une liste de termes correspondant aux missions, attributions et actions du ministère, puis nous avons associé à chacun des 70 termes proposés le nuage correspondant.

Regrouper les messages par famille

Pour classer les messages, nous avons utilisé une méthode semblable à Word2Vec dite de plongement de documents. Cette fois, chaque message (le message + les pièces jointes + les nommages) est représenté par un vecteur. L'enjeu est ensuite de créer des relations entre ces vecteurs et les nuages de termes réalisés précédemment. Un même message peut être relié à différents thèmes. De cette manière, nous avons pu créer des clusters de messages associés à une thématique. On peut ensuite construire des graphes permettant une approche quantitative des thèmes des échanges et éditer la liste des messages pour lesquels la méthode prédit un lien avec la thématique. Il est possible de choisir le niveau de granularité de la classification : une seule boîte mél ou plusieurs boîtes. Cette exploration de méthodes appuyées sur des réseaux de neurones artificiels et une démarche de traitement automatique de la langue adapté au français a permis de valider la pertinence d'une approche par plongement lexical et plongement de documents pour organiser de grandes masses de données archivées ou à archiver et augmenter la pertinence de la recherche. À partir des résultats obtenus, il est envisageable d'améliorer sensiblement non seulement l'accès par mots-clés, puisque la recherche ne porte pas sur l'identification d'un mot précis mais sur ce mot et les termes dont le sens est proche, mais aussi la connaissance du contenu réel des messageries, des thèmes des conversations, de leur évolution dans le temps et donc de produire des descriptions plus pertinentes et

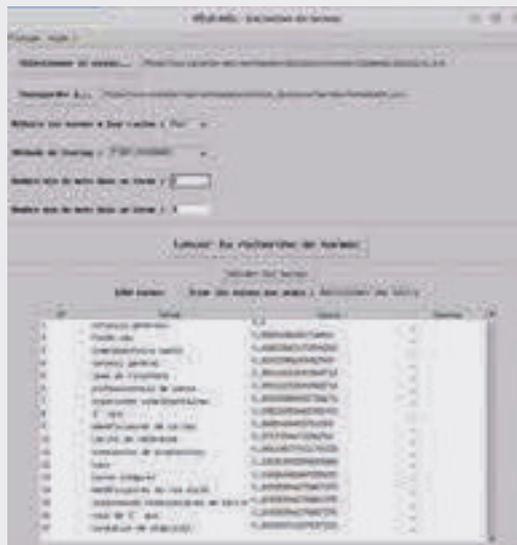
précises. Il est également possible de s'appuyer sur les résultats pour effectuer des choix entre boîtes méls et du tri interne à chaque boîte. Ces méthodes peuvent donc être mobilisées à différentes étapes de la chaîne archivistique : évaluation archivistique, sélection, description.



Bénédicte Grailles
Maîtresse de conférences en archivistique
Université d'Angers



Touria Aït El Mekki
Maîtresse de conférences en informatique
Université d'Angers



TF : plus le terme est fréquent plus son poids est élevé
IDF : mesure la rareté (poids plus élevé aux termes moins fréquents)
TF_IDF : importance d'un terme dans un document par rapport à l'ensemble de corpus

Figure 2
© Bénédicte Grailles et Touria Aït El Mekki

Suggestio, la machine qui sait décrire les fichiers

Quel archiviste n'a pas rêvé qu'avec toutes ces technologies apprenantes, il serait bientôt possible de trier-classer-décrire-renommer-éliminer automatiquement des dizaines de To ? Voici le récit d'une expérimentation prototypée prénommée « Suggestio » pour laquelle la société Un dossier Une place a été la lauréate au dernier concours IA-Award 2023.

Le projet Suggestio est né, il y a un an, et il y a onze ans !

Tout a commencé par l'invention du concept d'archivage accéléré®, il y a onze ans, pour Un dossier Une place. Cette invention est le résultat d'une combinaison humain-machine pour accélérer les processus d'archivage et traiter 1 000 à 2 000 dossiers par jour (trier-classer-indexer-renommer-éliminer-verser). Autant dire que la réorganisation documentaire est un peu l'ADN de notre société. Il était donc naturel d'arriver à la réorganisation numérique ! Voici le récit de notre expérimentation.

Le périmètre et l'équipe du projet

Nous avons procédé en prenant comme testeur les répertoires et fichiers en vrac d'une maison de retraite sur différents pôles d'activités : l'économat, la planification, la comptabilité, la gestion des résidents, la logistique et également sur des sources personnelles au départ. L'échantillon représentait environ 7 000 fichiers. Cette expérimentation s'est déroulée du 15 mai au 7 août 2023. Notre équipe se composait d'un *data scientist*, de deux étudiants en archivistique, d'un chef de projet spécialisé en IA, de nos deux développeurs de l'équipe FileTracker, et de la fondatrice-consultante (l'autrice©).

La méthodologie du projet

A priori, j'avais décidé de traiter les lignes d'export de fichiers suivant des tags connus des archivistes : les tag Objet, Action et Type. Cela ne pouvait être obtenu qu'après avoir bien étudié chaque métier et sa production.

Très vite, après quelques semaines d'exercices d'indexation, je m'aperçois que les archivistes avaient besoin d'être cadrés avec des descripteurs validés par verticale métier ; ce que je mis rapidement au point. Nous avons constaté que si une analyse archivistique est de qualité passable, la machine d'apprentissage ne pourra obtenir que 100 % au maximum d'une qualité d'analyse passable. La description archivistique a donc été soigneusement revue et corrigée avant validation et envoi des lignes à la machine

L'entraînement de la machine

Pendant ce temps, notre *data scientist* testait plus d'une dizaine d'algorithmes d'apprentissage sur chacun des tags par verticale métier sur 2/3 des lignes indexées, afin d'en réserver 1/3 pour tester la qualité d'apprentissage de la machine ensuite.

Des tests étaient effectués pour augmenter les données de façon « horizontale », c'est-à-dire l'ajout d'informations supplémentaires des métadonnées glanées pour entraîner les modèles. L'entraînement était effectué en trois étapes : *preprocessing*, entraînement et vérification. Plusieurs dictionnaires ont été créés pour améliorer la prédiction. Pour cela, les modèles étaient comparés et optimisés afin d'obtenir les meilleurs scores et prédictions possibles.



Figure 1 : La remise du trophée IA Award à Plaine Commune, mai 2023 © Françoise Cohen-Cassuto

Douze modèles sélectionnés pour l'entraînement au départ, quatre à l'arrivée!

Pour stabiliser les travaux, nous avons travaillé sur un modèle de données intégrant des référentiels métiers communs à notre premier logiciel (de *records management* hybride) FileTracker. Puis, nous avons procédé à la conception de méthodes de contrôle qualité, de la représentation des données et de l'architecture du système.

Au final, nous avons entraîné 5 modèles comparés via un comparateur que nous avons développé afin d'établir un score permettant d'évaluer la pertinence des prédictions : un score par modèle et par tag. Cela a permis d'éliminer le modèle le moins performant, il n'en restait donc plus que quatre à la fin.

Des résultats très prometteurs pour la version Beta

Le modèle que nous avons spécialisé a obtenu les scores suivants¹ :

- Tag Objet : 0.94
- Tag Action : 0.96
- Tag Type : 0.91

Quelles suites du projet

Le projet Suggestio ne se résume pas uniquement au traitement de réorganisation des fichiers électroniques des serveurs. C'est la première brique amont d'un logiciel plus ambitieux dont l'objectif serait de créer un pont entre le *records management* et le *knowledge management*.

Nous travaillerons dès le deuxième trimestre sur la version beta de la « brique amont », pour une commercialisation estimée à début 2025.



Françoise Cohen-Cassuto
Fondatrice de Un dossier Une place

1. Un score f1 d'un modèle d'apprentissage supervisé est un entier compris entre zéro et un, un étant le meilleur.

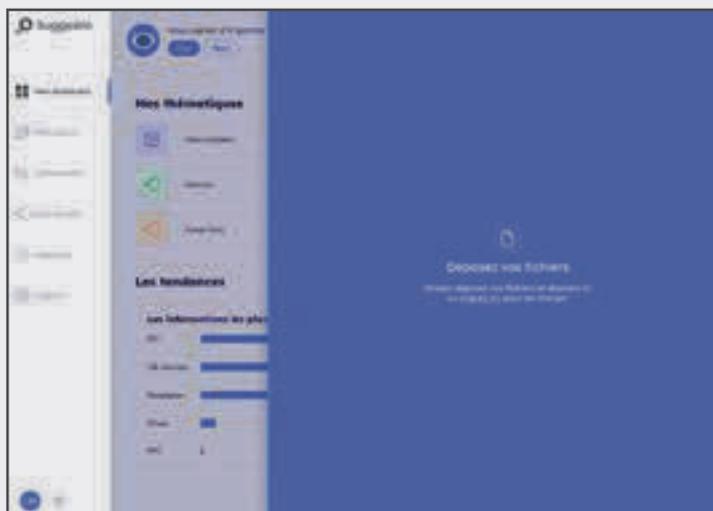


Figure 2 : Interface Suggestio d'import de fichiers © Françoise Cohen-Cassuto

L'archivage numérique des dossiers de recours devant la CNDA. Retours d'expériences et réflexions sur l'humain, le programmatique et l'artificiel

L'archivage des dossiers numériques de recours devant la Cour nationale du droit d'asile (CNDA) a été réalisé dans le cadre de deux projets DIAMAN (Dispositif d'accompagnement des missions pour l'archivage numérique) financés par le Service interministériel des archives de France (SIAF). Les travaux, débutés en mars 2022, ont abouti aux versements définitifs dans la plateforme des Archives nationales (AN) en septembre 2023.

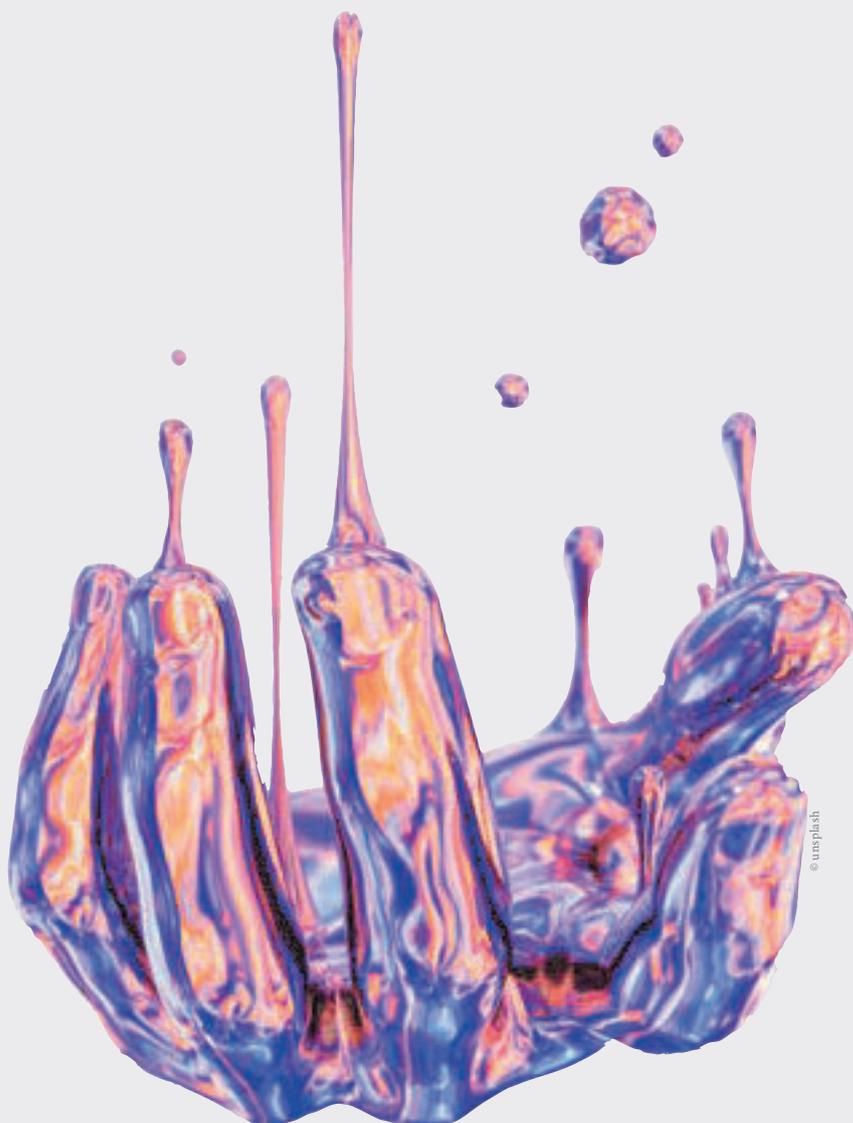
Cette réussite est due au travail conjoint d'une équipe constituée de représentants archives, métiers et informatiques de la CNDA, d'agents du SIAF et des AN (département de la justice et de l'intérieur, département de l'administration des données), et d'un prestataire privé, mintika. A posteriori, elle semble offrir un cas d'usage intéressant pour réfléchir à la place que pourraient tenir, dans une telle entreprise, l'humain, le programmatique et l'artificiel.

L'archivage portait sur le stock numérique des années 2010-2020, période d'hybridité physique/numérique, avant une dématérialisation totale. Chaque dossier réunit les pièces nécessaires à l'instruction d'un recours devant la CNDA contre une décision rendue par l'Office français de protection des réfugiés et des apatrides (OFPRA) en matière d'asile : dossier de l'OFPRA (enregistrements oraux des entretiens inclus à partir de 2016), recours, mémoires, courriers, documents de suivi et de gestion, décision... Dans le cadre des deux DIAMAN, 3 796 dossiers numériques ont été archivés, soit 19 638 fichiers pour une volumétrie totale de 116,91 Go.

À terme, les dossiers échantillonnés, physiques et numériques, formeront un triptyque archivistique, avec le minutier intégral des décisions papier, pour partie déjà conservé aux AN, et le registre intégral numérique (gardant trace de tous les dossiers traités, conservés ou non), qui sera collecté en 2024.

La majeure partie du projet a été consacrée à l'analyse des processus métiers et de la production documentaire, l'évaluation archivistique et la sélection des données et des métadonnées à collecter après échantillonnage, l'identification des modalités de capture de ces éléments, ainsi qu'à la modélisation des paquets d'archives cibles, la production d'un profil d'archivage et la création de jeux de tests garantissant la viabilité des spécifications.

Outre ces actions reposant sur l'intelligence et les connaissances humaines, le projet a employé des traitements programmatiques pour répondre à plusieurs besoins.



© unsplash

Le premier a été la réalisation d'audits des données à préparer. La détection manuelle de fichiers illisibles (PDF et MP3) dans un échantillon a imposé de vérifier tout le stock. L'analyse des fichiers concernés avec l'outil d'identification de formats DROID, développé par *The National Archives* britanniques, a confirmé une observation faite par les AN sur les formats de fichiers les plus courants : l'impossibilité d'identifier le format d'un fichier était un symptôme de risque d'illisibilité. À défaut d'un outil existant sur le marché, l'échantillon a ensuite été passé au crible d'un script en cours de développement par les AN dont le principal objectif est d'évaluer la lisibilité des fichiers d'une liste de formats limitée (incluant le PDF et le MP3) en simulant leur ouverture. Les résultats ont confirmé les premières observations. L'analyse croisée de l'ensemble à archiver a permis d'identifier 949 PDF et 3 MP3 illisibles, soit environ 5 % de l'ensemble. Aucune restauration n'étant possible, après autorisation d'élimination, les fichiers illisibles ont été supprimés par un script produit *ad hoc*. En parallèle, les audits ont permis de corriger, à la main, 19 cas d'incohérence entre l'extension du fichier et son format.

La force programmatique a aussi servi à produire une indexation typologique à partir des nommages des fichiers. La convention de nommage de la CNDA désigne la typologie documentaire par un sigle. Six typologies à indexer ont été identifiées : dossier de première instance, entretien, recours, correspondance, mémoire judiciaire et décision. Une analyse humaine de l'intégralité des nommages a détecté d'importants écarts entre la théorie de la convention et la pratique. Selon les spécifications des AN, *mintika* a produit un script associant typologies et éléments de nommage via l'utilisation d'expressions régulières : tout fichier dont le nommage contient « *_R_* » est par exemple indexé comme un « recours ». Pour couvrir tous les cas particuliers, 24 expressions ont été nécessaires. *In fine*, ont été indexés environ 11 200 fichiers (soit 57 %). Les autres fichiers ne relevaient pas de typologies à indexer ou en réunissaient plusieurs indiscernables d'après les nommages.

Enfin, *mintika* a réalisé des développements programmatiques pour constituer les paquets d'archives (SIP) à partir des deux entrants fournis par la CNDA : une arborescence de fichiers rassemblés en dossiers regroupés par année (données) et un tableur d'éléments métiers et archivistiques (métadonnées) issus de la base utilisée par le producteur. Des SIP conformes au SEDA et aux préconisations des AN sont ainsi obtenus en exploitant le langage de transformation XSLT pour produire les métadonnées descriptives et l'outil ReSIP pour les métadonnées techniques. L'intelligence humaine et la programmation ont suffi à répondre aux besoins de l'archivage des dossiers de recours de la CNDA, sans recours à l'intelligence artificielle (IA). Mais le retour d'expérience autorise à s'interroger a posteriori sur ce qu'elle aurait pu apporter de plus.

Un traitement automatique du langage naturel aurait pu être envisagé pour améliorer les métadonnées. Une reconnaissance d'entités nommées n'aurait pas eu d'intérêt, du fait de l'existence d'informations sur les demandeurs fiables dans la base du producteur, qui servira à constituer le registre intégral. Il aurait peut-être été possible de perfectionner l'indexation

typologique, en partant de la réalité des contenus pour corriger des erreurs et oublis dans les nommages, mais des sondages ont montré que la marge de perfectionnement était minime. C'est sur le sujet de la fixation des délais de communicabilité de chaque dossier que le recours à une IA aurait été intéressant pour compenser l'impossibilité humaine de dépouiller l'intégralité des pièces. Dans ces deux cas, le défi aurait été grand : il aurait fallu rendre le système capable d'analyser les contenus à l'aune soit de la diplomatie, soit du code du patrimoine. Pour des petits corps, comme celui de la CNDA, une mise en balance des bénéfices hypothétiques et des coûts d'entraînement semble confirmer que l'IA n'aurait pas été rentable.

Quant à utiliser l'IA pour produire des scripts tels que ceux mobilisés pour ce projet, des expérimentations parallèles des AN et de *mintika* ont démontré que l'interrogation d'un agent conversationnel permet de répondre à des besoins simples et ponctuels. Mais aucun n'est encore capable d'aboutir à des programmes complexes articulant plusieurs traitements ou plusieurs scripts, a fortiori pour manipuler des standards peu répandus comme le SEDA. Il reste alors plus efficace, comme pour le présent projet, de recourir au savoir-faire humain individuel (développeurs) ou communautaire (contributeurs des forums spécialisés).

L'écueil majeur serait de faire de l'IA, dans le discours sur l'archivage numérique, un nouveau mirage trompeur ou stérilisant, à l'image du mythe de l'automatisation. Les retours d'expériences récents des AN démontrent que l'on sait collecter et gérer des archives numériques sans l'IA.

Pour autant, l'IA doit demeurer une potentialité à mobiliser en fonction des entrants, besoins, objectifs, et ressources disponibles. On peut prévoir que son usage s'imposera a minima dans deux cas. Le premier cas sera le traitement de données d'une quantité trop grande pour être appréhendables par une équipe-projet, et de métadonnées d'une qualité trop faible pour suffire aux fins archivistiques de la collecte, de la gestion et de l'accès. Le second cas sera celui de fonds numériques ou numérisés dont la logique originelle a été perdue mais peut être restaurée. Il reste à souhaiter que de tels projets fournissent des terrains d'expérimentation archivistique à l'IA.



Matias Ferrera
Conservateur du patrimoine
Département de l'administration
des données des Archives nationales

Avec la participation de :

Pour la Cour nationale du droit d'asile :

Adeline Denoed

Pour le département justice et intérieur
des Archives nationales :

Violaine Challéat-Fonck,

Tiphaine Gaumy et Christophe Bouvier

Pour le Service interministériel des archives
de France :

Hombeline Aubigny

Pour *mintika* :

Baptiste Nichele

L'intelligence artificielle à l'Institut national de l'audiovisuel (INA) : de l'expérimentation à l'industrialisation

L'intuition du potentiel offert par les traitements automatiques au bénéfice des missions documentaires (archivistiques et bibliothéconomiques) ne date plus d'hier.

Une approche « classique » : l'IA comme soutien à la production documentaire

En effet, la confrontation du responsable de fonds à des ensembles de plus en plus massifs, du fait notamment de la dématérialisation des processus et de la démultiplication des données en découlant, a conduit depuis longtemps au constat d'une impossibilité de traitement exhaustif sur la base de ressources uniquement humaines. Dès lors, la perspective de traitements industrialisés voire automatisés, donnant accès à la finesse de l'information au sein d'ensembles touffus, représentait une piste de facilitation pour la production d'instruments de recherche, dans la continuité naturelle des pratiques professionnelles préexistantes. À l'INA, cette approche a pu se traduire à partir de 2019 par une démarche : d'étude des outils et technologies de traitement automatique disponibles sur le marché et assez matures pour produire des résultats de qualité d'identification des cas d'usages métier assez volumineux et présentant des caractéristiques formelles assez homogènes pour être candidats à l'industrialisation, sur la base de paramétrages de règles de traitements des POC et tests, à des fins de mise en production¹

Les cas pratiques ainsi traités s'inscrivaient donc dans l'objectif de générer des données destinées à intégrer les modèles de données existants en matière de gestion documentaire des collections. Le cas le plus abouti, portant sur la segmentation automatique des journées de diffusion de certaines chaînes d'information continue (de manière à restituer au plus près du temps réel des phénomènes d'interruption des programmes tels que les « *breaking news* »), a ainsi permis d'intégrer à partir de 2023 des données d'identification et de *time-codage* des programmes dans les bases de données en production courante. Concrètement, l'expression de besoin prenait ici pour modèle ce qu'un traitement manuel aurait produit... s'il avait été matériellement possible de générer une segmentation fondée sur le visionnage quasi continu de la diffusion, 24 heures sur 24.

L'IA comme générateur de données inédites

En parallèle de cette mise en production, dont on peut noter qu'elle a nécessité plusieurs années de tests et d'itérations, mais aussi d'appropriation des enjeux de l'automatisation, pour répondre aux attendus

fonctionnels, une approche différente est apparue dans les projets de l'INA au tournant de 2020, peut-être au moment où l'appréhension de traitements automatiques tels que la transcription, la détection d'entités nommées ou encore la reconnaissance de visages, commençait à être suffisamment mature pour se détacher du modèle des pratiques documentaires préexistantes.

C'est ainsi à cette période qu'a commencé à émerger le projet « *data.ina.fr* », plateforme qui, dans la continuité de principe des travaux d'analyse des médias menés depuis des années par l'INA, vise à générer des indicateurs d'objectivation du propos médiatique, en répondant à des questions telles que « de quoi parle-t-on à la télévision et à la radio ? », « quelles sont les personnes les plus présentes ? », etc².

Dès lors, l'enjeu de recours à l'IA n'est plus de poursuivre à l'identique la production documentaire « classique », mais de générer, de manière industrielle et sur la durée³, des clés d'entrée nouvelles dans les collections, sur un mode analytique et synthétique.

L'IA conversationnelle : le double enjeu du contenu et de son contexte

Une nouvelle étape semble encore s'être ouverte depuis le début d'année 2023, avec la soudaine apparition dans le paysage « grand public » des agents conversationnels généralistes de type ChatGPT, capables de traiter des volumes extrêmement massifs de données sources pour en proposer une restitution sur-mesure, en fonction de la question posée en entrée par l'utilisateur.

L'enjeu technique de vectorisation des éléments de connaissance (textuels ou visuels) issus des fonds semble à première vue renforcer la rupture avec les métadonnées « classiques » : il ne s'agit plus désormais de continuer à alimenter des bases de données relationnelles, structurées selon des modèles de données précis, mais de constituer des ensembles sémantiques interrogeables via les moteurs idoines. Cette rupture semble d'autant plus consommée que se profilerait à terme la possibilité de produire des réponses dyna-

1. Sur l'évolution des pratiques documentaires à l'aune de l'intégration d'outils automatisés, voir DRIBAUT DUJARDIN (C.-M.), FACKLER (D) et PICHON (J), « L'évolution des pratiques de description des archives documentaires de l'Institut national de l'audiovisuel (INA) à l'ère du numérique », *La Gazette des archives*, n° 257 (2020-1), pp. 191-202.

2. Sur ce projet, voir l'intervention de Gautier Poupeau à l'occasion des « Futurs fantastiques », troisième conférence internationale du cycle sur l'IA dans les bibliothèques, les archives et les musées (Paris, décembre 2021), via l'URL : [https://www.bnf.fr/fr/mediateque/en-quoi-lintelligence-artificielle-constitue-pour-lina-une-opportunit-e-pour-renforcer].

3. Cette démarche n'est pas nouvelle dans sa nature, au regard des travaux de la Recherche de l'INA et des publications de *La Revue des médias* [https://larevuedesmedias.ina.fr/], mais les premiers ne sont pas des objets d'industrialisation, et les secondes proposent une dimension rédactionnelle forte : la nouveauté réside donc ici dans le caractère systématique et la mise en production de données existant « en elles-mêmes », indépendamment d'un appareil journalistique.

miques aux questions posées par l'utilisateur : la métadonnée lisible par l'utilisateur deviendrait dès lors volatile, puisque adaptable au cas par cas.

Pour autant, les métadonnées structurées conservent une valeur et un intérêt propres – outre le besoin de la recherche de disposer d'objets assez stables pour garantir la réplicabilité des résultats. De fait, les métadonnées existantes peuvent porter, de par leur mode de production « historique », un indice de confiance particulièrement fort, et donnent ainsi des éléments de contexte de référence qui viendront consolider voire objectiver les analyses par de « simples » croisements sémantiques. Ainsi, le fait de continuer à produire et à capitaliser « en dur » des métadonnées de catalogue⁴, même si le mode de production s'appuie de plus en plus sur des procédés automatisés, garantirait l'existence d'un cadre de référence contextualisant les analyses sémantiques de contenus. Ces éléments de contextes sont essentiels pour que l'utilisateur de demain, travaillant sur une thématique donnée au sein des collections audiovisuelles de l'INA, constitue son corpus de titres ou d'extraits tout en prenant en compte la nature des émissions elles-mêmes : un même thème peut ainsi être abordé dans un journal télévisé, dans un magazine de reportages, ou encore sur un plateau d'émission-débat, mais probablement pas dans les mêmes termes, ni par des locuteurs de même nature... Cette qualification différenciée

du contexte n'étant pas forcément explicite dans le déroulé de l'émission, donc dans l'objet archivé en tant que tel, on peut imaginer le rôle qu'auront à jouer, plus que jamais, les référentiels (de titres de collections, de locuteurs présents dans l'émission...) pour accompagner l'utilisateur dans la compréhension du contexte du document.

Alors que l'IA permet d'envisager des instruments de recherche décrivant les fonds dans une granularité quasiment infinitésimale, si les enjeux fondamentaux de la conservation des fonds au service de leur consultation et de leur valorisation ne semblent pas remis en cause, il apparaît essentiel de (re) définir ce que l'on attend désormais d'une métadonnée, voire ce que l'on entend par là. Objectivation d'un contexte et d'une nature de production, ou simple restitution de la teneur et du contenu du document ? Plus que jamais, le positionnement de l'archiviste semble se trouver dans l'équilibre de ces approches, pour donner à l'utilisateur de demain les clés de lecture (et de) critique des traces documentaires produites aujourd'hui.



Eléonore Alquier

Directrice adjointe
Data & Technologies
Institut national de l'audiovisuel

4. Telle la « Bibliographie nationale du dépôt légal radio, télévision et webmédias », accessible via l'URL : [http://inatheque.ina.fr/].



Du classement à l'indexation : l'IA dans tous ses états

Traitement des vrac bureautiques et IA : un premier pas dans la porte

L'utilisation de l'IA conversationnelle pour automatiser les commandes Windows (PowerShell) peut offrir une solution accessible et déterminante pour la préparation de données bureautiques en vue d'un versement. Néanmoins, dans un domaine en manque de maturité, les archivistes doivent adopter une approche axée sur les processus et respecter des principes stricts de protection des données.

L'essor des données bureautiques au cours des deux dernières décennies a contraint les archivistes à gérer des volumes de plus en plus importants lors de la préparation des versements. En outre, il convient de tenir compte de la valeur probante accrue, résultant de l'utilisation de dispositifs de signature¹ et de traçabilité. Cette période aux alentours de 2010 constituera probablement un tournant dans la façon d'évaluer les archives numériques. Pour simplifier, nous sommes passés dans une ère où l'usage des outils bureautiques est indépendant de la production de dossiers « papier », ce qui tend à rendre les archives numériques plus anciennes paradoxalement moins pertinentes que celles produites ultérieurement.

Devant l'impossibilité de réorganiser automatiquement des documents désordonnés dans un système d'archivage, une Gestion électronique des documents (GED) ou des outils collaboratifs, il est essentiel d'agir rapidement pour éviter toute perte d'informations. Volume des données, diversité des formats, données présentant un caractère d'unicité, richesse accrue des contenus mais médiocrité des métadonnées, absence d'une politique de nommage et, souvent, intervention de l'archiviste au dernier moment : voilà les paramètres qui nous ont incités à développer des méthodes où l'IA peut sécuriser tout en accélérant.

Depuis 2022, le traitement des vrac bureautiques aux Archives nationales de Luxembourg repose sur l'application de scripts Powershell permettant de classer, de trier et de vérifier l'intégrité des données avant l'emballage. La production de plans de clas-

sement associés aux rubriques des tableaux de tri² constitue une étape indispensable. Avec l'arrivée de ChatGPT et de Google Bard, l'écriture des scripts s'est trouvée facilitée et fiabilisée pour répondre rapidement à nos besoins.

Lors de nos opérations de reprise de vrac de plusieurs téraoctets³, il était impératif d'éviter le partage des données brutes avec une IA conversationnelle. La communication avec les IA s'est limitée à la production de résultats techniques. La vérification des scripts a pu être effectuée par une deuxième IA⁴, ouvrant la voie à une optimisation du processus.

Les étapes de contrôle de l'intégrité, de copie, d'extraction (avec des options de migration de format) et d'analyse des métadonnées, et de la transformation sont concernées par le recours à l'IA afin de créer des scripts en permettant l'automatisation.

Les gains en termes de traitement sont intéressants. En 2022, un vrac de 350 Go a nécessité près de 60 jours alors qu'en 2023, un vrac de 1,2 To n'en prenait que 20⁵. En 2023, un versement d'archives audiovisuelles représentant 18 000 fichiers en vrac a pu être classé

2. D'origine belge, le terme « tableau de tri » est utilisé au Luxembourg. Il est l'équivalent des expressions « tableau de gestion » en France, « calendrier de conservation » en Suisse Romande et au Canada francophone ou encore « référentiel de conservation », expression faisant consensus dans les normes internationales depuis 2010. Voir : [https://anlux.public.lu/fr/gerer-ses-archives/tableaux-de-tri.html].

3. À ce jour le plus important versement bureautique traité grâce à l'appui de l'IA est de plus de 4 To.

4. Cette idée pragmatique de double contrôle est notamment développée depuis plusieurs années dans les key notes d'A. Bensoussan dans le domaine des professionnels du monde juridique. Voir : [https://www.alain-bensoussan.com/].

5. 2017-0026_FRVV – Fonds de rénovation de la vieille ville de Luxembourg et V000351_Esch2022 – Archives administratives de Esch22 – capitale européenne de la culture.

1. Le règlement eIDAS 910/2014 du 23 juillet 2014 a permis de généraliser des dispositifs de signature conférant une valeur de preuve équivalente à la signature manuscrite. Cependant l'usage de certificats et/ou de clés de cryptage est antérieur dans certaines organisations.

en deux heures grâce à un script produit à l'aide de l'IA6 ; un déplacement manuel aurait été bien plus long et complexe.

Notre usage de l'IA est guidé par la quête d'une productivité accrue dans le cas du traitement des vrac numériques, exercice qui favorise l'expression des besoins des archivistes et facilite le dialogue avec les services versants parce qu'il offre aux ANLux la capacité de proposer des moyens de mettre en œuvre leurs exigences. Les outils obtenus avec l'aide de l'IA sont mis à disposition des services versants, ce qui donne lieu à des échanges visant à améliorer les scripts, donc le processus de traitement des vrac.

Aussi avons-nous choisi de demander à l'IA uniquement des scripts reposant sur le langage Windows PowerShell, qui présente l'avantage d'être compatible avec la majorité des équipements informatiques rencontrés. Ce langage permet d'initier le dialogue avec les services de support informatique, qui peuvent prendre le relais en produisant des programmes plus performants et plus ergonomiques, marquant la frontière entre la pratique archivistique et le champ d'intervention technique.

Lors des conversations avec une IA générative, nous avons standardisé les formules et les résultats souhaités. Nous avons restreint la création de fichiers en sortie aux seuls formats non propriétaires, n'impliquant pas l'installation de modules supplémentaires ou l'achat de licences. Ce principe a été motivé par l'idée qu'il était important de garantir une exécution simple et interopérable des scripts par les services versants, quel que soit leur environnement de travail et indépendamment de la répartition des droits administrateurs tout en respectant les principes de sécurité, pour que ces outils soient pertinents et viables.

Afin de documenter ces évolutions, nous avons conçu un catalogue des scripts, recensant leurs fonctions respectives et les objectifs visés.

6. Le temps de création du script étant inclus dans cette période de deux heures.

Ce catalogue peut alors servir de cahier des charges en vue de futurs développements ou de l'intégration éventuelle d'une forge logicielle. Les scripts et les fichiers de résultats peuvent quant à eux être intégrés à un dossier retraçant l'ensemble des opérations appliquées à chaque versement.

Chaque fonds d'archives numériques est unique, ce qui induit des spécificités propres à tout traitement. L'archiviste en charge de l'inventaire d'un vrac applique, certes, une méthode générale propre à son institution, mais l'adapte en fonction des éléments constitutifs du fonds traité.

Les mêmes impératifs de flexibilité de traitement se posent dans le contexte numérique. Si l'outil « magique » capable de traiter toutes les situations dans la reprise d'un vrac n'existe pas, la production de scripts avec l'aide d'une IA conversationnelle est un atout précieux pour l'archiviste.

Cela rend la capacité d'action aux professionnels qui ne dépendent plus de développements spécifiques de leurs services informatiques et peuvent eux même développer des solutions métiers sûres, en plus d'offrir de nouvelles possibilités ; singulièrement celle d'expérimenter. Pour peu que l'on ait recours à un échantillon de données de test, il est possible de développer une approche itérative pour aboutir à la solution sur-mesure sans endommager le moindre document.

Il devient ainsi possible de créer un script pour chaque étape du traitement d'un vrac numérique qui tienne compte des spécificités du fonds concerné, en mobilisant peu de ressources et dans des délais raisonnables.

Pour que la communauté des archivistes puisse tirer profit des expérimentations de chacun dans l'intégration des IA dans des processus métiers, il nous semble opportun, au moment de conclure cet article, d'appeler à la création d'une forge commune et internationale qui viendrait soutenir une véritable culture de l'innovation au sein de notre communauté.

Camille Forget, Richard Gaudier et Michel Cottin
Service collecte, conseil et encadrement
Archives nationales de Luxembourg

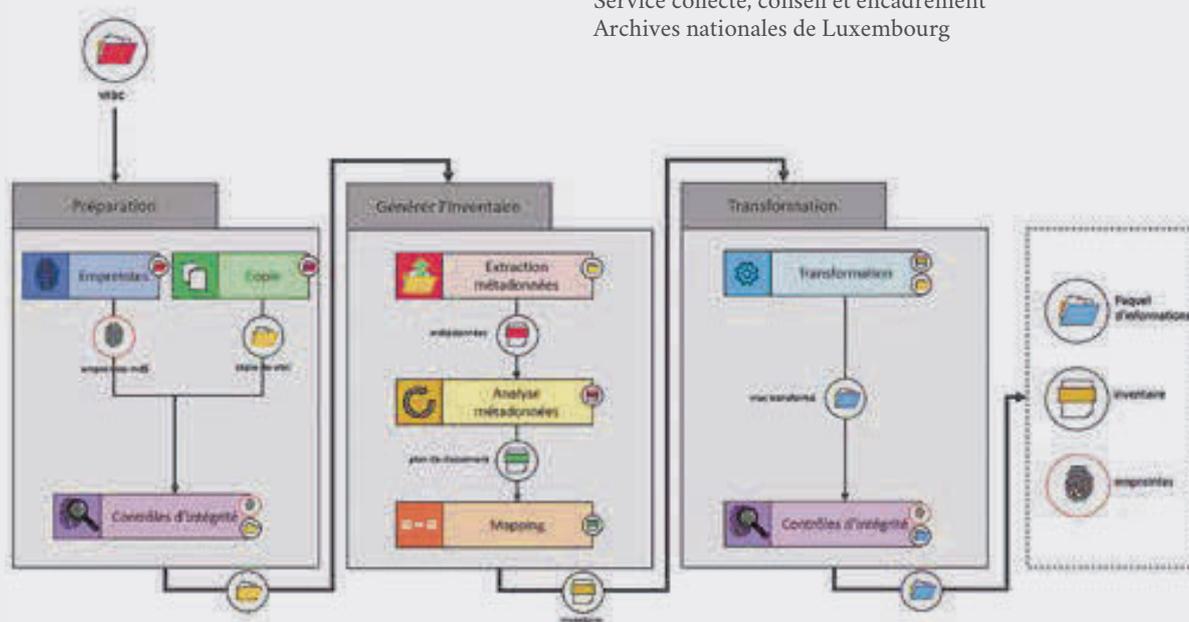


Fig. 1 : schéma de processus © Cyril Clause, ANLux

Pistes de réflexion éthique

Intelligence artificielle et santé, une réflexion éthique ?

L'intelligence artificielle a envahi médiatiquement notre quotidien récemment. Tout le monde parle de ChatGPT comme de LA nouveauté qui semble éclipser les autres outils d'intelligence artificielle existant dans de nombreux secteurs comme celui de la santé.

En santé, l'IA intègre de plus en plus d'outils mais ce n'est que très récemment que les commerciaux utilisent régulièrement, le vocable d'intelligence artificielle. Auparavant, il n'était question que d'outils se basant sur du *machine learning* « nourri » des données produites massivement par le secteur.

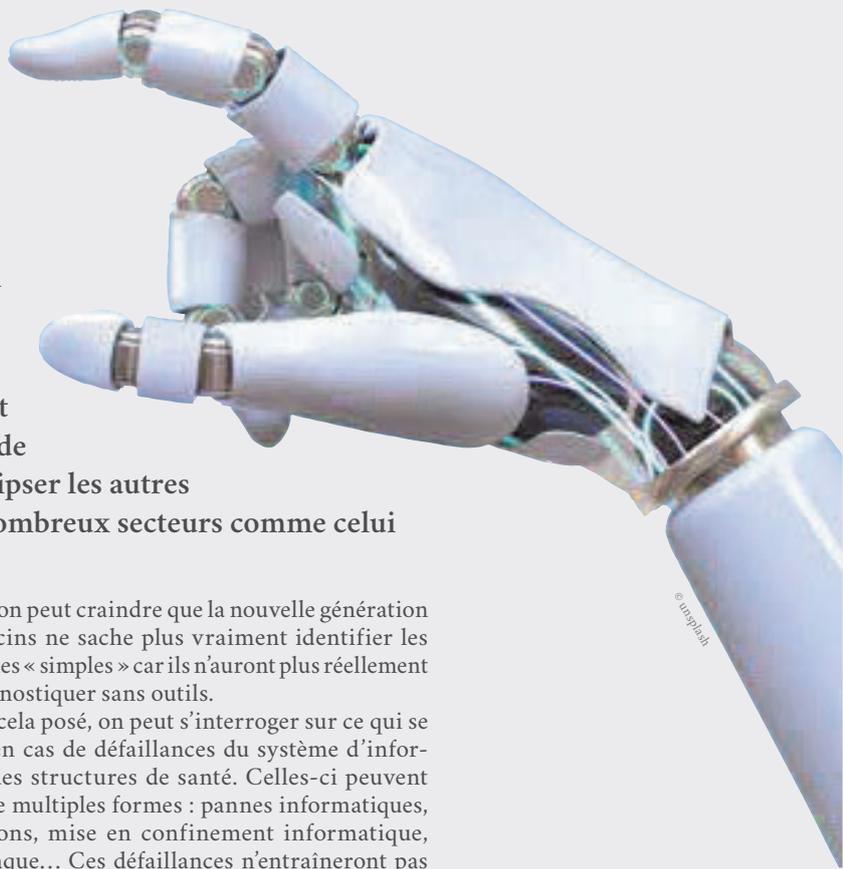
Ces outils ont vocation à être uniquement une aide à la décision médicale ; ils ne doivent aucunement remplacer la décision humaine. On s'en voit proposer des « incroyables » qui aident : à repérer des fractures ou des pseudarthroses, à la décision sur la réalisation d'un acte de thrombectomie intracrânienne... Ces dispositifs basés sur l'intelligence artificielle et le *machine learning*¹ sont présentés comme permettant de réduire l'erreur humaine de diagnostic mais nécessitent de manière systématique le regard d'un professionnel de santé aguerri pour confirmer le diagnostic. Contrairement à des outils intervenant dans d'autres champs, ceux utilisés en santé relèvent de la réglementation européenne sur les dispositifs médicaux² comportant des obligations et un marquage CE.

Bien que très séduisants, ils interrogent de multiples manières des professionnels de santé ayant déjà acquis une longue expérience : comment va-t-on former les internes à la détection d'un dommage si l'outil, sans doute plus efficace qu'eux, leur est mis à disposition ? Il est nécessaire de mener une réflexion bénéfices-risques. L'idée s'impose de plus en plus que dans un contexte de démographie médicale défavorable, l'IA apporte une aide sur des diagnostics de base et permet aux professionnels de se concentrer sur des problématiques plus pointues, nécessitant un regard humain et des hautes compétences.

1. L'apprentissage automatique (*machine learning* en anglais) est un champ d'étude de l'intelligence artificielle qui vise à donner aux machines la capacité d'« apprendre » à partir de données, via des modèles mathématiques. Plus précisément, il s'agit du procédé par lequel les informations pertinentes sont tirées d'un ensemble de données d'entraînement.

Le but de cette phase est l'obtention des paramètres d'un modèle qui atteindront les meilleures performances, notamment lors de la réalisation de la tâche attribuée au modèle. Une fois l'apprentissage réalisé, le modèle pourra ensuite être déployé en production. Source : CNIL.

2. Règlement (UE) 2017/745 du Parlement européen et du Conseil du 5 avril 2017 relatif aux dispositifs médicaux, modifiant la directive 2001/83/CE, le règlement (CE) no 178/2002 et le règlement (CE) no 1223/2009 et abrogeant les directives du Conseil 90/385/CEE et 93/42/CEE.



À terme, on peut craindre que la nouvelle génération de médecins ne sache plus vraiment identifier les pathologies « simples » car ils n'auront plus réellement à les diagnostiquer sans outils.

Une fois cela posé, on peut s'interroger sur ce qui se passera en cas de défaillances du système d'information des structures de santé. Celles-ci peuvent revêtir de multiples formes : pannes informatiques, inondations, mise en confinement informatique, cyberattaque... Ces défaillances n'entraîneront pas nécessairement la fermeture de la structure mais un retour au papier et au crayon dans l'attente de la remise en route du système et de l'accessibilité aux outils hébergés hors de la structure dont pour une partie des outils d'intelligence artificielle. Selon la cause de la défaillance, le retour opérationnel peut aller de quelques heures à plusieurs mois. Comment feront les médecins non formés à l'identification des pathologies simples ? Faudra-t-il orienter tous les patients vers d'autres structures de soins ? Celles-ci auront-elles la capacité d'absorber le surcroît de patients ? Pour quelques heures, jours, il y aura sans doute report des prises en charge, mais au-delà ?

La mise en œuvre d'IA dans le domaine de la santé questionne la responsabilité en cas de contentieux. Actuellement, les outils sont utilisés en seconde intention. Le résultat d'interprétation par l'intelligence artificielle doit être obligatoirement revu par un professionnel de santé. On peut imaginer que dans le temps, ce contrôle systématique soit supprimé. Qu'en sera-t-il alors de l'imputabilité de la responsabilité ? Dans le même champ, qui dit contentieux dit capacité à conserver les données dans le temps. Si la conservation des données originelles et de l'interprétation s'organise sans trop de difficultés, qu'en est-il de la conservation de l'algorithme qui a produit l'interprétation ? Pourrait-il être nécessaire de la produire en justice ? Comment l'archiver ?

L'intelligence artificielle offre des perspectives en gain de temps, efficacité et niveau de service rendu au patient mais ne risque-t-elle pas d'entraîner une perte de connaissance et de compétences chez les professionnels l'utilisant ?



Marie-Laure Kervégant
Membre de la section AES

L'IA pour communiquer

Libérer le passé avec l'IA : l'inspiration suisse

Il n'est pas nécessaire d'insister sur l'importance croissante et nécessaire des solutions numériques dans le travail archivistique. Deux applications d'IA reposant sur différents modèles pour analyser les images fixes et animées, d'une part, et pour la reconnaissance textuelle, de l'autre, commanditées par les Archives fédérales suisses à Bern (AFS) permettent de s'engager dans un voyage archivistique intégré à la technologie avec concrétude. La première est pleinement opérationnelle ; la seconde, un nouveau-né de 2023, est en devenir et avance à très grand pas.

Archipanion : pour les images et les vidéos

En utilisant Archipanion, les AFS ont enrichi l'accessibilité de leurs archives visuelles à partir d'un corpus de 200 heures de vidéo provenant de 1 750 dossiers, sans recours à des métadonnées descriptives manuelles laborieuses, offrant un gain de temps et de pertinence aux archivistes qui peuvent se consacrer à d'autres missions tout aussi importantes et parfois négligées.

Transkribus¹ et Lexipanion : pour la reconnaissance des textes

Les procès-verbaux du Conseil fédéral de 1848 à 1972 ont permis de rendre ainsi facilement accessibles aux chercheurs et au public des textes à grande portée historique, grâce à un mélange de technologies : celles de Transkribus, un projet autrichien d'intelligence artificielle, et de Lexipanion en particulier, pour déchiffrer l'écriture manuscrite, ainsi que d'OCR traditionnel pour les documents dactylographiés. L'objectif prometteur est de permettre une recherche intégrale des textes de toute époque et en toute langue.

Limites et défis de ces applications à modèles multiples

Les limites sont tangibles lorsqu'on est confronté à des écritures difficiles sans grand corpus, telles que les lettres autographes d'Anne de Bretagne qui sont particulièrement coriaces — une cursive de la fin du xv^e siècle empesée et non normée — ne permettant pour le moment qu'un entraînement peu fluide. Sept lettres aux Archives municipales de Douai. Combien au total ? Il faudrait, pour le moment, entre 200 et 300 pages pour que l'entraînement avec le modèle utilisé soit efficace. Les premiers essais sont toutefois encourageants pour rendre compte d'une affaire très sensible et peu connue : le projet de mariage de Claude de France avec le nouveau-né Charles Quint.

Un défi donc pour les modèles qui, finalement, sans exemple de transcription, ont réussi à déchiffrer bien des mots. Mais que fait-on en réalité lors de la transcription en tant qu'archiviste ? Bien sûr, on essaie d'abord de reconnaître les mots, ou du moins les lettres, et de trouver des modèles dans l'écriture qui facilitent la reconnaissance. Au final, il faut souvent

aussi une forme d'interprétation pour décrire un texte sensé et aussi précis que possible à partir du contexte et de l'ensemble des documents. C'est ce qui nous a donné l'idée de ne pas nous contenter d'utiliser la reconnaissance de caractères et de mots, mais de faire également interpréter les caractères reconnus par l'IA, à l'aide de ChatGPT par exemple.

Ce voyage n'est donc pas sans heurt. La mise en œuvre de l'IA pose des défis tels que la garantie de la confidentialité des données, la gestion des failles technologiques et la courbe d'apprentissage des nouveaux outils numériques. Ces considérations obligent à poursuivre notre tâche avec un mélange d'enthousiasme, de curiosité et de prudence.

L'expérimentation et la collaboration sont vitales, comme l'abandon de la prétendue perfection qui est si chère aux archivaires.



Vincent Doom

Conservateur en chef du patrimoine honoraire



Markus Stauffiger

Coconcepteur d'Archipanion et de Lexipanion



Transkribus : Nous montrons ici où va le voyage dès que la transcription est effectuée. On lie des noms, des lieux, des chiffres etc., qui renvoient par des liens hypertextes à Wikipédia et une multitude d'autres sites © Vincent Doom et Marc Stauffiger

1. Accessible via l'URL : [https://transkribus.eu/r/mining-hub/#/documents/643776/pages/5?t=%C3%B6sterreich].

Valoriser les archives contractuelles

Les archives contractuelles sont un actif précieux pour les entreprises mais bien souvent inexploitées. L'intelligence artificielle permet de les valoriser en facilitant leur reprise en masse dans les logiciels de contrathèque.

Une fois signés, les contrats commencent généralement leur vie dans une armoire, au mieux rangés méthodiquement, au pire empilés par ordre d'arrivée, attendant vainement d'être consultés lors d'un différend ou d'une nécessité. Avec l'avènement de la numérisation, les contrats mènent une vie en parallèle, stockés sur des serveurs selon une logique d'arborescence qui tient compte le plus souvent du type de contrat, du nom des contractants ou de l'année de signature.

Toutefois, ces archives numérisées ne sont pas valorisées à leur juste valeur alors qu'elles sont les principaux actifs au bilan de l'entreprise dans certains secteurs d'activité. Les logiciels de contrathèque répondent à ce besoin en centralisant le stockage des contrats et en organisant de manière uniforme le traitement des informations issues de ceux-ci.

Si l'enregistrement systématique de nouveaux contrats dans ces logiciels est relativement aisé, l'intégration du stock de contrats déjà signés s'avère beaucoup plus complexe et fastidieuse. En effet, jusqu'alors, cette opération nécessitait le recours à une main-d'œuvre importante confiée soit à des stagiaires, soit à des prestataires externes bien souvent délocalisés. L'IA s'est révélée une formidable opportunité pour surmonter ces difficultés. Les acteurs de la *legal tech* ne s'y sont pas trompés en proposant des offres de reprises de contrats en masse à l'aide d'IA.

La chaîne de traitement pour la reprise de contrats est à peu de chose près la suivante :

- Les contrats signés sont versés et centralisés dans un lieu de dépôt unique.
- Les documents sont ensuite chargés sur une plateforme d'IA qui va extraire des points d'informations : noms des parties, titre du document, date, durée, extractions de certaines clauses.

- Un audit qualité est réalisé sur un échantillon. Les premiers résultats sont souvent très bons, en particulier pour les points d'informations ayant déjà fait l'objet de nombreux apprentissages. Pour les questions plus spécifiques, une phase d'entraînement supplémentaire est souvent nécessaire.
- Une fois le niveau de qualité atteint, les données sont extraites sur l'intégralité des contrats puis nettoyées dans Excel ou dans un logiciel de nettoyage et de mise en forme de données.
- Les fichiers PDF des contrats et leurs points d'informations sont ensuite chargés dans le logiciel de contrathèque.

L'extraction des données est généralement réalisée à l'aide d'un modèle pré-entraîné BERT (outils de traitement automatique du langage) spécialisé dans les questions-réponses. Concrètement, il suffit de poser des questions en langage naturel pour extraire les points d'informations. Par exemple, pour retrouver la date d'entrée en vigueur du contrat, on posera simplement la question : « Quelle est la date d'entrée en vigueur du contrat ? ». Il est à noter que ce type de modèle se débrouille très bien si la question est posée en français alors que le document est en allemand. Ces extractions peuvent être ensuite transformées dans des formats spécifiques en ayant recours à d'autres modèles d'IA. Les progrès des derniers modèles d'IA générative permettent d'imaginer de nouveaux traitements.

Les principaux écueils sont liés au contenu des documents : les informations recherchées sont formulées de façon implicite ou se trouvent dans des tableaux ; aux limites de la reconnaissance optique de caractères (OCR) : qualité de la numérisation, mention manuscrite, structure du document ; aux hallucinations de l'IA dues à de précédentes corrections inexactes ainsi qu'à la capacité de montée en charge de la plateforme d'IA : l'analyse de 1 000 documents de 8 pages ne requiert pas la même puissance de calcul que l'analyse de 10 000 documents de 50 pages.

En conclusion, le même constat s'impose : les projets menés avec succès sont les projets où les producteurs d'archives se sont impliqués activement, en n'hésitant pas à challenger le prestataire tout en lui suggérant des pistes pour améliorer l'extraction des données.



Guillaume Cayeux
Legal Technologist
Legal Excellence



Réflexion sur les apports de l'IA au métier

IA : vers une réalité archivistique ?

Une nouvelle révolution est à nos portes : l'intelligence artificielle. Un de ses représentants les plus médiatisés, ChatGPT, a beaucoup œuvré pour l'adoption de cette technologie sur le plan privé. La conséquence en a été une accélération, dans la sphère professionnelle, de cette même adoption, ainsi que de la projection sur des usages métiers.



© Mohamed Nohassi-umplash

Ces derniers sont déjà nombreux à se dessiner, alors même que les experts s'accordent à dire que le potentiel technique de l'IA est loin d'avoir été atteint¹. Seules quelques organisations peuvent, à ce jour, se targuer de l'avoir mise en œuvre à l'échelle, au profit d'activités métier. Après de premières projections dystopiques au sujet de ses effets à long terme sur l'emploi², les constats factuels témoignent plutôt d'une aide apportée aux métiers pouvant et sachant utiliser l'IA, sans destruction d'emploi³. L'IA réalise les tâches fastidieuses, et permet à l'humain de se recentrer sur son cœur de métier, selon les premières expérimentations, notamment dans des métiers bancaires, assurantiels et industriels. Pour l'heure, le mouvement ne fait que s'amorcer dans le domaine de la documentation. La révolution sera sans doute aussi importante que celle amenée par le digital depuis 30 ans. Ce dernier a en effet amené une transformation du support du document, à travers les programmes de dématérialisation. Les systèmes de Gestion électronique des documents

(GED) et d'archivage électronique (SAE) ont, pour leur part, accompagné la croissance ininterrompue du volume documentaire produit. L'IA, de son côté, pourrait engendrer une mutation de la nature-même des documents. En quoi cela va-t-il transformer le métier d'archiviste ? L'IA peut-elle réellement l'aider ? Quelles en sont les limites ?

De l'e-archivage à l'IA-archivage

À défaut de modifier le support documentaire lui-même, l'IA pourrait impacter la façon de produire le document, et *in fine* son contenu.

L'IA fait monter une nouvelle marche au digital dans son impact sur le document, et même sur l'ensemble de la chaîne documentaire. Les entités et les organisations ont besoin de produire encore davantage de contenus dans un temps limité. Pour cela, elles pourront adopter une technologie qui leur permettra de produire un contenu nativement généré par l'outil. Ce contenu sera éventuellement modifié à la marge par l'humain, mais certains contenus ne le seront pas : par exemple, les comptes rendus de réunion. Il y a toutefois, à ce jour, des limites techniques à prendre en compte. Notamment, l'usage du *speech to text*⁴ ne permet pas de retranscrire plusieurs voix de façon optimale lorsqu'elles proviennent d'un même canal, même si les premières solutions sont prometteuses. Difficile, donc, de retranscrire automatiquement les échanges d'un conseil d'administration. Lorsque ces barrières techniques seront levées, plus rien ne s'opposera à

1. Mao Blaise, « Yann Le Cun : "D'ici cinq ans, plus personne n'utilisera un modèle tel que ChatGPT" », *Usbek et Rica*, 19 avril 2023, consultable via l'URL : [https://usbeketrica.com/fr/article/d-ici-cinq-ans-plus-personne-n-utilisera-un-modele-tel-que-chatgpt].

2. Beyler Neïla, « ChatGPT et l'IA menacent 300 millions d'emplois dans le monde, selon Goldman Sachs », *Les Échos*, 28 mars 2023, consultable via l'URL : [https://www.lesechos.fr/tech-medias/intelligence-artificielle/chatgpt-et-lia-menacent-300-millions-demplois-dans-le-monde-selon-goldman-sachs-1919968].

3. Organisation internationale du travail, Communiqué de presse intitulé « L'intelligence artificielle générative devrait compléter plutôt que détruire des emplois », 21 août 2023, consultable via l'URL : [https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_890746/lang--fr/index.htm].

4. Conversion de la voix en texte.

la collecte d'archives directement générées par l'IA. Or, une archive est, par essence, un document validé et engageant pour son producteur. L'archiviste se doit de vérifier qui est l'auteur du document. Pour l'heure, il n'est pas en capacité de déterminer si l'IA a été impliquée ou non. Cette technologie posera donc un nouveau défi pour l'archiviste dans sa fonction de responsable du contrôle des documents collectés. Pour l'aider, il pourrait lui-même user d'une solution lui permettant de détecter les documents générés par une solution à base d'IA, à condition d'être formé à ces nouvelles compétences. Pour l'heure, cet usage n'est pas encore techniquement abouti. Ce blocage soulève donc le risque de constitution de fonds documentaires représentant une mémoire à la fois humaine et « artificielle »⁵ mais concrètement indifférenciée, à laquelle le lecteur du document ne sait pas quelle valeur accorder.

Un traitement de masse salvateur ?

L'arrivée de l'IA pourrait par ailleurs intensifier un défi bien connu des archivistes : la quantité de documents produits par les entités et organisations, qui influe directement sur la quantité d'archives à gérer. L'IA est elle-même une réponse à la difficulté qu'ont les services producteurs à suivre le rythme effréné des flux digitaux auxquels ils doivent répondre. Afin de contrer cet effet, est-il envisageable pour l'archiviste de mobiliser l'IA afin de faciliter le traitement d'un volume documentaire colossal ? L'IA pourrait sans doute aider l'archiviste dans son traitement de la masse, notamment en facilitant l'indexation des fonds. Ainsi, la *computer vision*⁶, qui permet la reconnaissance d'un type ou modèle de document, est déjà établie. Ce type de solution permet également d'extraire et de restituer certaines métadonnées, notamment la date et l'auteur d'un document. Charge aux services d'archives de contrôler les résultats par échantillons, et de compléter les éléments manquants. Les récents modèles sémantiques génératifs peuvent quant à eux produire des résumés dans différentes langues. Cependant, cette technologie présente trois limitations significatives. La première est liée à la qualité des données restituées. Le taux d'erreurs est faible mais réel, d'autant plus prononcé sur l'analyse de documents manuscrits. Ensuite, l'IA est principalement applicable aux fonds électroniques, en raison du coût prohibitif de la numérisation massive des documents. Enfin, chacune des innombrables typologies reconnaissables implique de mobiliser un volume témoin significatif pour l'apprentissage : 1 000 exemples a minima pour limiter les interprétations aléatoires de l'IA.

Faciliter la communication

Sur le plan de l'aide à la valorisation des fonds, l'IA n'est pas en reste. Elle offre notamment la possibilité d'une recherche, en langage naturel, des fonds et documents pertinents pour une thématique donnée. Cet usage est spécifiquement envisageable pour la

recherche par les lecteurs, ainsi que dans le cadre d'une exposition archivistique. L'analyse sémantique de la demande permet d'atteindre des résultats de qualité, grâce à une compréhension de la demande bien plus fine que la détection de mots-clés, sur laquelle se basent les moteurs de recherche documentaire actuels. Cet usage impose la prise en compte d'exigences fonctionnelles. La solution doit en effet permettre d'effectuer des requêtes dans des délais acceptables, pour le lecteur comme pour l'archiviste. Ce critère est à prendre en compte dans le dimensionnement technique de la solution, adossé à des volumes documentaires substantiels. Par ailleurs, l'archiviste devra se réappropriier les résultats, les corriger et les compléter à l'aide de son expertise personnelle. Au-delà de la recherche de fonds répondant au besoin, les solutions à base d'intelligence artificielle peuvent dès à présent aider les services d'archives dans la génération de contenus créatifs, toujours dans le cas d'un projet d'exposition. En effet, les outils existants aident déjà à problématiser un sujet sur la base d'une description des données à disposition. Ils peuvent aussi générer les supports de communication associés à l'évènement, ainsi que leur traduction. Il faut là aussi y voir une aide pour l'archiviste plutôt qu'un remplacement, dans la mesure où le travail créatif du propos de l'exposition demeure humain.

Les métiers de la documentation offrent des possibilités d'usages de l'intelligence artificielle. Cette technologie invite à voir d'un nouvel œil toute la chaîne d'activités de ces métiers, de la collecte du document à sa valorisation. L'IA n'en est toutefois pour l'heure qu'à ses balbutiements dans le métier d'archiviste. Un nouveau seuil de performance doit être atteint avant d'envisager une transformation profonde de la chaîne documentaire.

Aujourd'hui, l'enjeu est de préparer cette mutation. Pour cela, les archivistes doivent s'acclimater au plus vite à cette technologie, en s'appropriant ses qualités et ses limites pour en faire une force dans leurs activités professionnelles plutôt que de la subir. Il est nécessaire d'accompagner et de sensibiliser les métiers de la documentation sur les transformations à venir. Ils en seront les premiers bénéficiaires, et n'en perdront pas la valeur de leur expertise. Il s'agit d'une condition nécessaire à l'adoption de cette technologie, avec laquelle ces métiers avanceront de concert.



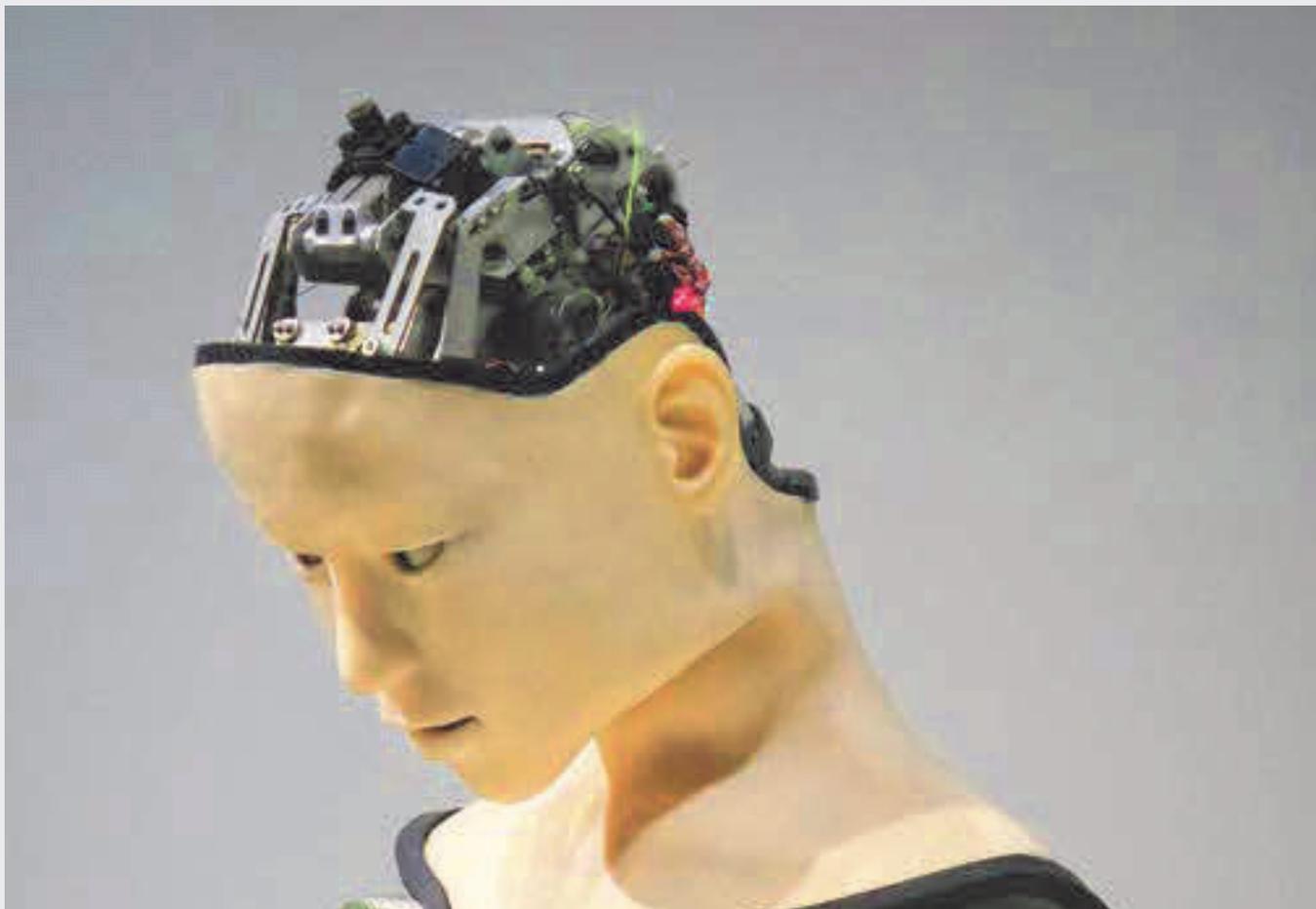
Simon Vérove
Consultant senior
Julhiet Sterwen



Livia Bahier Michel
Senior manager
Julhiet Sterwen

5. Écharoux Émilie, « IA : "On va avoir de plus en plus de mal à constituer des archives authentiques" », *Usbek et Rica*, 26 septembre 2023, consultable via l'URL : <https://usbeketrica.com/fr/article/ia-on-va-avoir-de-plus-en-plus-de-mal-a-constituer-des-archives-authentiques-de-notre-futur#:~:text=Usbek%20%26%20Rica%20%2D%20IA%20%3A%20%2C%2%AB,%C3%A0%20constituer%20des%20archives%20authentiques%20%2C%20BB>.

6. Vision par ordinateur.



Conclusion

Prompt suggéré à Chatbot GPT : « L'archiviste va-t-il être remplacé par l'intelligence artificielle dans les années à venir ? en 10 lignes maximum »

Sa réponse :

Il est fort probable que l'intelligence artificielle jouera un rôle de plus en plus important dans le domaine de l'archivage dans les années à venir. Les avancées technologiques permettent déjà de numériser et d'organiser de vastes quantités de données de manière efficace, ce qui pourrait faciliter le travail des archivistes.

Cependant, il est peu probable que l'intelligence artificielle remplace complètement les archivistes, car ces derniers apportent une expertise humaine et une compréhension contextuelle essentielle pour interpréter et organiser les informations de manière pertinente.

En fin de compte, une combinaison de l'intelligence artificielle et du travail humain pourrait être la voie à suivre pour optimiser le processus d'archivage et garantir la préservation adéquate des connaissances pour les générations futures.



Chatbotgpt.fr