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.

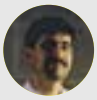
Enfin, mintika a réalisé des développements programmatiques pour constituer les paquets d'archives (SIP) à partir des deux entrants fournis par la CNDA : une arborescence de fichiers rassemblés en dossiers regroupés par année (données) et un tableur d'éléments métiers et archivistiques (métadonnées) 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 diplomatique, soit du code du patrimoine. Pour des petits corpus, 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 Denoeud
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