PRÉDIRE L'ÉTAT MATÉRIEL DES DOCUMENTS : DALGOCOL, UN PROGRAMME DE RECHERCHE EN INTELLIGENCE ARTIFICIELLE À LA BNF

Entretien avec Philippe Vallas

Philippe Vallas

Adjoint du directeur du département de la Conservation à la BnF,
chargé notamment de la coordination des activités de conservation physique

DALGOCOL (Fouille de Données et ALGOrithmes de prédiction de l'état des COLlections) a été lancé en 2018 par Alaa Zreik dans le cadre de sa thèse de doctorat. Il prend la forme d'un projet collaboratif en science des données entre la Bibliothèque nationale de France (BnF) et l'université Versailles-Saint-Quentin-en-Yvelines. Ce travail de recherche est financé par le laboratoire d'excellence (Labex) Patrima[1], devenu Fondation des sciences du patrimoine (FSP). L'objectif n'est pas d'aboutir à un outil opérationnel, mais de tester la possibilité d'utiliser des méthodes d'intelligence artificielle (IA) sur les données informatiques produites par la BnF dans le cadre de ses activités de conservation.

Pourquoi recourir à l'IA ? L'un des enjeux est notamment de tirer parti de très grandes masses de données : ainsi, environ deux millions de documents ont été testés au début du travail. L'automatisation permet de croiser des données de différentes natures pour aider des spécialistes à prendre la décision de retirer ou non une ressource de la consultation, le temps de la restaurer. Si le travail est piloté par Alaa Zreik, il mobilise les compétences des professionnels de la BnF.

Pour le *Bulletin des bibliothèques de France*, Philippe Vallas revient sur les opportunités ouvertes par Dalgocol. Archiviste-paléographe, conservateur général des bibliothèques, Philippe Vallas a consacré sa carrière à la conservation des collections de la Bibliothèque nationale de France ; il est actuellement adjoint du directeur du département de la Conservation, chargé notamment de la coordination des activités de conservation physique. Principal rédacteur de la Charte de la conservation de la BnF (dernière version validée en 2015), il a représenté son établissement dans plusieurs instances nationales (Comité technique de restauration du MCC/SLL) ou internationales comme l'IFLA (2007-2014) ou EROMM (depuis 2006). Il supervise les activités de la coordination informatique et du laboratoire scientifique et technique de son département, qui conduisent ou participent à de nombreux programmes de recherche en conservation du patrimoine écrit.

*

1 https://www.chcsc.uvsq.fr/labex-patrima

***BBF*: Depuis de nombreuses années, les professionnels produisent et archivent des informations concernant la communication des ouvrages, l'état des documents, ou encore l'historique des traitements de conservation. Quels sont les apports de Dalgocol ?**

Philippe Vallas : Comme vous le savez, la Bibliothèque nationale de France compte parmi ses missions prioritaires la conservation pérenne de ses collections patrimoniales, en premier lieu celles issues du Dépôt légal. Bien que les moyens dont elle dispose soient considérables, ils ne permettent évidemment pas le traitement systématique de tous les documents qui le nécessiteraient. L'élaboration d'une stratégie de conservation, et notamment le choix des documents à traiter, de façon préventive comme curative, est compliquée et chronophage. Dalgocol avait pour but de tester l'utilisation de l'intelligence artificielle pour faciliter les activités de conservation, à partir des données informatisées produites en masse par la BnF depuis maintenant une vingtaine d'années. Plus précisément, l'objectif était de parvenir à prédire l'état physique d'un document à partir des données produites au cours de son « histoire » : descriptions informatisées existantes de ses caractéristiques physiques et bibliographiques, des historiques des traitements dont il a pu bénéficier, de ses communications éventuelles au public, des dégradations diverses qui ont pu l'affecter, etc.

Comment s'est nouée la collaboration avec le chercheur Alaa Zreik ?

Dalgocol est un projet de recherche élaboré en 2018 par la BnF en collaboration avec le laboratoire DAVID de l'université de Versailles Saint-Quentin-en-Yvelines, spécialisé en intelligence artificielle, sous la forme d'une thèse de doctorat financée par la Fondation des sciences du patrimoine (ex-Labex Patrima). Alaa Zreik, le doctorant, a donc été accueilli à temps partiel pendant trois ans à la BnF, à partir d'octobre 2018, pour un accès le plus aisé possible aux données nécessaires, comme aux professionnels de l'établissement.

Quels types de données Dalgocol permet-il d'exploiter ?

Nous avons cherché à exploiter tous les types de données informatisées utiles pour la prédiction de l'état matériel des documents, qui sont plus variées qu'on ne l'imagine : bien sûr, celles produites par les activités de conservation proprement dites, qui décrivent l'état physique des documents (par exemple « plats détachés » ou « papier acide »), les traitements, conditionnements dont il a pu bénéficier (« reliure mécanisée », « numérisation », « désacidification » par exemple, avec souvent des détails plus précis), les sinistres éventuellement subis (« dégât des eaux » de telle date), mais aussi des données bibliographiques (les dates d'édition notamment sont importantes), les descriptions matérielles fournies par les éditeurs et qui sont en partie conservées dans les notices des catalogues (« couv. ill. en coul. », « dépl. » par exemple) ; et les statuts (« communicable », « hors d'usage ») et données de communication aux lecteurs, critère souvent décisif pour la priorisation des traitements. À l'exception de cette dernière catégorie, chiffrée, toutes ces données sont essentiellement textuelles.

Au total, nous avions comptabilisé au moins 28 bases de données « maison », dont nos catalogues, susceptibles de fournir des informations utiles, utilisant des formats et logiciels très divers (Excel, Lotus Notes…), et évidemment rarement interopérables.

En quelques mots, pourriez-vous nous expliquer les étapes de traitement des données de conservation-restauration de Dalgocol ?

Un préalable important pour Alaa Zreik était de comprendre l'organisation de la conservation à la BnF, notamment le mode de sélection et de priorisation des traitements, ainsi que le fonctionnement et le rôle des applications et bases de données utilisées. Ensuite, il a recensé et classé les types de dégradation (« déchirures », « pliures », « feuillets détachés », « couture rompue »…) et de traitements, en notant les équivalences afin d'aboutir à une terminologie stable et homogène et permettre les comparaisons entre documents. Une trentaine de types d'« évènements » susceptibles d'influencer la vie, ou « trajectoire », d'un document ont ainsi été recensés, à partir de plus de 7 millions d'évènements concernant près de 2 millions de documents.

À partir de ce travail, Alaa a créé une ontologie, c'est-à-dire un modèle conceptuel qui définit des relations possibles entre les différents évènements qui peuvent affecter l'état d'un document, et il a représenté l'histoire de la vie et donc de la conservation des documents par des trajectoires sémantiques.

Une fois ce modèle établi, il a pu tester l'intelligence artificielle – en l'occurrence des algorithmes de type K-Mean[2] – pour comparer, en fonction des données disponibles sur les documents, la similarité des trajectoires et prédire leur état (en l'occurrence, sont-ils communicables aux lecteurs ou hors d'usage).

À l'heure actuelle, quel est le degré de fiabilité de Dalgocol, en termes de prédiction de l'état physique des documents ?

Les derniers tests de tri entre les documents communicables et hors d'usage ont été très concluants (50 documents mal classés sur un panel de 700, soit 85 % d'orientations justes). Ces résultats auraient certainement été encore améliorés si nous avions disposé de plus de temps pour affiner les trajectoires et rajouter des données.

Quelles sont les conditions pour que Dalgocol fournisse des performances optimales ?

La condition essentielle est la quantité et la qualité (précision, datation…) des données disponibles pour chaque document, et leur proportion au sein d'une collection (possibilité d'opérer des extrapolations).

Pourriez-vous nous présenter en quelques mots les différentes compétences représentées au sein de l'équipe ? Quel est le rôle de chacun ? Quelles sont vos méthodes de travail ?

Nous avions organisé une équipe très complémentaire par les compétences de ses membres : Alaa Zreik est évidemment un informaticien expert en fouilles de données, mais il ne connaissait pas les documents et leurs dégradations ni les problématiques et stratégies de conservation liées. Pour ma part, j'étais parfaitement ignorant dans sa spécialité mais bon connaisseur des collections de l'établissement, des pratiques et stratégies

2 [NDLR] Pour fournir une définition très simplifiée, un algorithme de type K-Mean mobilise une technique d'apprentissage permettant de définir des ensembles (clusters).

de conservation ainsi que des données nécessaires pour les mettre en œuvre. Nous étions appuyés par deux collègues bibliothécaires appartenant à la coordination informatique du département de la Conservation, dont les compétences étaient intermédiaires, ainsi que plus ponctuellement par des ingénieurs du département des Systèmes d'information de la BnF, qui ont « ouvert » à Alaa Zreik les bases nécessaires.

Nous avons très longuement dialogué avec Alaa, surtout au début, pour lui expliquer comment était organisée la conservation à la BnF, quelles étaient les données importantes, où les trouver, la signification et les équivalences entre les termes des différents traitements ou dégradations, les liens de cause à effet, les correspondances entre dégradations et traitements, etc. Nous avons ainsi passé en revue tous le vocabulaire des bases principales, celles qui décrivent les caractéristiques et l'état physique des documents et celles qui décrivent les traitements de conservation réalisés.

En tant que professionnel des bibliothèques, quelles conclusions tirez-vous de ce projet scientifique ?

Je retiens bien sûr l'intérêt de travailler avec une personne d'une spécialité très éloignée de la mienne, et la richesse des échanges une fois défini le vocabulaire commun indispensable pour se comprendre mutuellement. Plus concrètement, je suis désormais convaincu de l'apport potentiel de l'intelligence artificielle, une notion maintenant un peu moins vague pour moi ; surtout, l'enseignement principal que nous en tirons, concernant notre établissement, est que les données informatisées disponibles sont encore loin d'être suffisantes pour permettre à l'IA de produire sa pleine efficacité : si les collections sont certes précisément décrites du point de vue intellectuel, bibliographique, les descriptions physiques sont très incomplètes, voire sommaires, et les dégradations ne sont indiquées, dans la grande majorité des cas, que pour les documents dirigés vers une filière de traitement de conservation selon un processus informatisé (ce qui n'est le cas, à la BnF, que pour les monographies et périodiques imprimés, à l'exception de tous les documents dits spécialisés, et seulement depuis une vingtaine d'années à peine). La coordination informatique du département de la Conservation a donc lancé, en coordination avec la direction des Collections, un travail de sensibilisation et d'homogénéisation des règles de description physique des différents types de documents, afin d'intensifier, voire de systématiser la production de ces « données utiles à la conservation » selon l'appellation donnée à ce projet, qui est bien sûr de long terme mais qui devrait considérablement faciliter le travail de nos successeurs pour la gestion de la conservation pérenne de nos collections. •

Deux articles d'Alaa Zreik pour aller plus loin

- ZREIK, Alaa et Zoubida KEDAD. « Matching and analysing conservation–restoration trajectories », *Data & Knowledge Engineering*. 1er mai 2022, vol. 139. p. 102015. En ligne : https://doi.org/10.1016/j.datak.2022.102015 [consulté le 31 mai 2022].
- ZREIK, Alaa et Zoubida KEDAD. « Matching Conservation-Restoration Trajectories : An Ontology-Based Approach » in Samira CHERFI, Anna PERINI et Selmin NURCAN (dir.). *Research Challenges in Information Science*. Cham : Springer International Publishing. 2021 (Lecture Notes in Business Information Processing).

Résumé

DALGOCOL (« Fouille de Données et ALGOrithmes de prédiction de l'état des COLlections ») a été lancé en 2018 par Alaa Zreik dans le cadre de sa thèse de doctorat. L'objectif n'est pas d'aboutir à un outil opérationnel, mais de tester la possibilité d'utiliser des méthodes d'intelligence artificielle sur les données informatiques produites par la Bibliothèque nationale de France (BnF) dans le cadre de ses activités de conservation. Pourquoi recourir à l'IA ? L'un des enjeux est notamment de tirer parti de très grandes masses de données : ainsi, environ deux millions de documents ont été testés au début du travail. L'automatisation permet de croiser des données de différentes natures pour aider des spécialistes à prendre la décision de retirer ou non une ressource de la consultation, le temps de la restaurer. Si le travail est piloté par Alaa Zreik, il mobilise les compétences des professionnels de la BnF.

Tous les contenus publiés sur le site du Bulletin des bibliothèques de France sont placés sous licence libre CC BY-NC-ND 2.0 : Attribution – Pas d'utilisation commerciale – Pas de modification 2.0 France.