L'IA À LA BIBLIOTHÈQUE NATIONALE DE FRANCE : LA PATRIMONIALISATION 4.0

ARNAUD BEAUFORT (X88)
directeur du numérique au secrétariat général du ministère de la transition écologique et de la cohésion des territoires et du ministère de la transition énergétique, ancien DGA de la BnF (jusqu'au 24 octobre 2022)

Les grandes bibliothèques, avec leurs collections de grandes dimensions, trouvent dans le traitement de la donnée et donc dans l'IA des outils précieux pour l'exploitation de leurs immenses collections. Le cas de la Bibliothèque nationale de France (BnF) est particulièrement éloquent en la matière. Voici la présentation de quatre de ses projets utilisant l'IA.

Non seulement l'IA est susceptible d'outiller la collection d'une bibliothèque comme la BnF, mais elle promet aux chercheurs et à tous les amateurs de données un matériau de travail inédit, de nouvelles clés de découverte et de futures trouvailles. C'est d'autant plus vrai que les données et les contenus susceptibles de faire l'objet d'expérimentations et de projets s'y trouvent en très grande quantité et que la collection numérique va considérablement s'étendre à la faveur du dépôt légal numérique.

L'intérêt de la BnF pour l'IA

En sa qualité de service public, la BnF réfléchit à cette double dimension interne et externe. Cela fait une vingtaine d'années qu'elle explore les nombreux champs d'application de l'IA : la reconnaissance optique de caractères (OCR), le traitement automatisé de la langue, l'analyse de données, l'analyse de documents (périodiques, catalogues de vente, cartes, partitions musicales…), etc. Ce paysage fertile et bigarré a ouvert la voie à des projets d'envergure et à des perspectives inédites en termes d'exploration et de traitement.

Ce changement suscite un vif intérêt chez les professionnels des bibliothèques, comme en témoigne le succès des différentes conférences organisées sur le sujet, et il appelle un cadrage capable de prendre en compte les enjeux à la fois managériaux, juridiques, éthiques et sociétaux soulevés par l'usage de l'IA. Les missions de la BnF – collecter, conserver, enrichir et communiquer le patrimoine documentaire national – sont suffisamment ancrées historiquement et légalement pour garantir une continuité d'action et de positionnement, ainsi qu'une qualité de service, au gré de l'introduction des technologies nouvelles. C'est dans cette dynamique que les documents stratégiques récents de l'institution ont intégré l'IA (feuille de route de la BnF sur l'intelligence artificielle, contrat d'objectifs et de performance 2022-2026…).

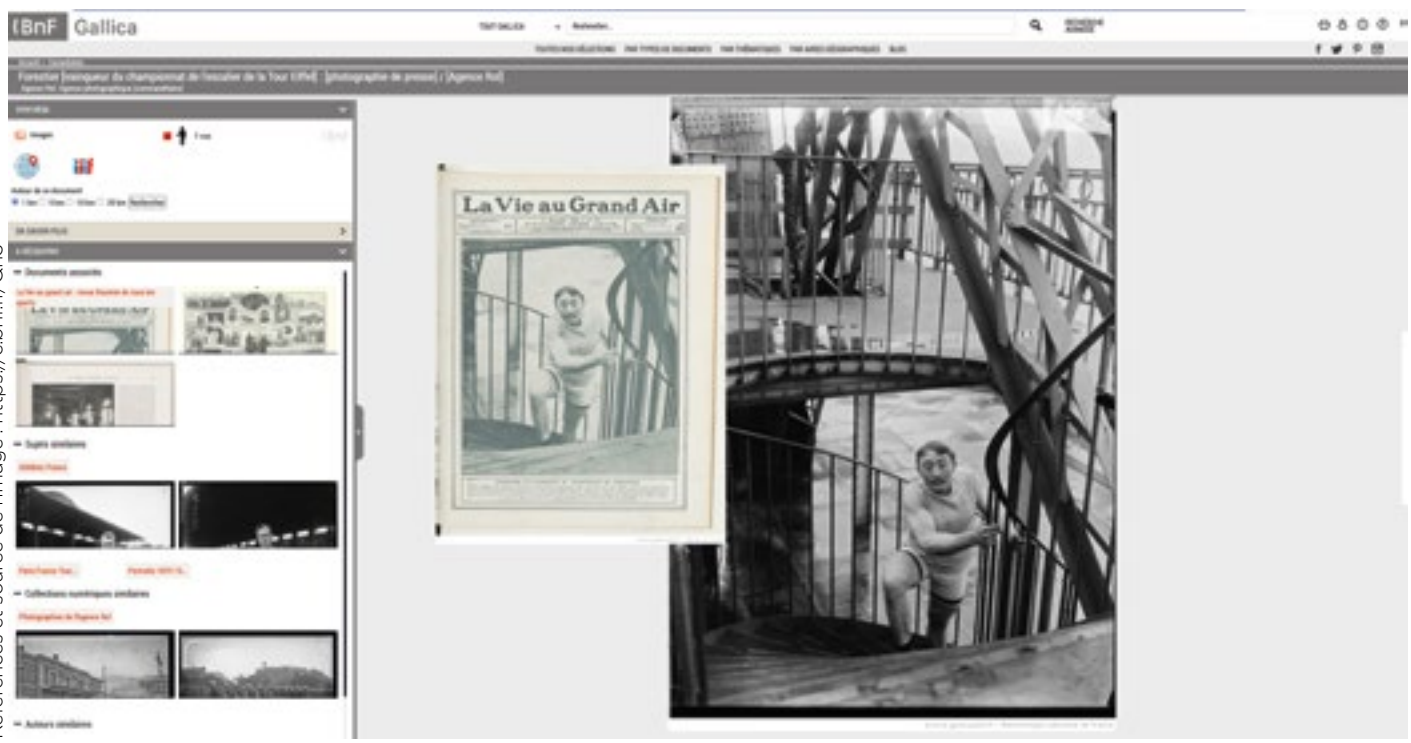
Penser l'IA selon une logique d'ouverture

Les chercheurs qui souhaitent entraîner leurs algorithmes peuvent trouver d'importants jeux de données à la BnF. Au-delà des ressources actuellement disponibles (le site api.bnf.fr, qui donne accès aux jeux de données de la BnF et à ses API, et le DataLab, son équivalent physique), il s'agit de faire en sorte que les données francophones puissent servir d'entraînement à des projets dans le sillage de SQuAD (The Stanford Question Answering Dataset) : lancé par une équipe de recherche de l'université Stanford à la fin des années 2010, ce projet consistait à apprendre à une machine à répondre à des questions à partir d'un algorithme et d'un jeu de données constitué d'articles de Wikipédia, de questions sur ces articles et de réponses. Un enjeu global de *découvrabilité* des contenus culturels numériques en ligne se précise, partagé avec de nombreuses institutions, en particulier francophones. Pour relever ces défis, la Bibliothèque s'inscrit dans un positionnement résolument coopératif, fondé sur l'intelligence collective et la mutualisation, pour mieux prendre en compte les enjeux environnementaux, tout en faisant face aux dépenses importantes suscitées par l'IA, et en préservant les valeurs du service public. Quatre projets illustrent les bénéfices attendus de l'IA en termes de service.

Gallica Images

Gallica (https://gallica.bnf.fr) est la bibliothèque numérique de la BnF et de ses partenaires. Les images y sont omniprésentes, que ce soit dans la presse, dans les livres et bien sûr dans les fonds iconographiques. Ce projet de fouille d'images répond à des cas d'usage très pratiques, par exemple à la volonté de trouver plus facilement la source des images publiées dans les journaux numérisés : les collections numériques de la BnF comprennent à la fois de nombreux titres de presse et des fonds d'agence photographique, qui pourraient être rapprochés de manière automatisée.

Gallica Images s'inscrit dans la continuité d'expérimentations engagées dès le début des années 2010 à la Bibliothèque : les premiers projets de recherche menés dans ce domaine avec plusieurs laboratoires ont été l'occasion d'approfondir les ressources de la numérisation et d'évaluer l'apport des algorithmes en matière d'indexation. En 2016, GallicaPix, prototype de moteur de recherche sémantique réalisé à partir des API de récupération des contenus de Gallica, des données et d'outils d'intelligence artificielle (dont IBM Watson Visual Recognition, Google Cloud Vision, OpenCV), a pu satisfaire des situations classiques de recherche par mot clé, par type ou par thème dans des corpus d'images. D'autres expérimentations ont développé l'usage de moteurs de recherche visuelle favorisant une →



Ici, une photo de l'agence Rol et un journal dans lequel elle a été publiée. Les appariements aujourd'hui réalisés à la main dans le cadre de projets ponctuels pourraient être automatisés grâce à l'IA.

Références et source de l'image : https://c.bnf.fr/Qn3

→ recherche de similarités entre deux images, tel GallicaSnoop, développé avec l'Inria et l'Ina à partir du moteur Snoop, utilisé par l'application PlantNet.
Gallica Images sera lancé en 2023 avec le soutien du Programme d'investissements d'avenir (France 2030). Il a pour objectif d'étendre ces travaux à l'ensemble de Gallica. Il s'agit de rendre toutes les images largement accessibles en industrialisant une technologie de segmentation (repérage des images à l'intérieur des livres, presse et revues numérisées à l'aide du protocole IIIF, International Image Interoperability Framework) et de caractérisation (format, couleurs, typologie…) par intelligence artificielle. Piloté par la BnF, la Bibliothèque nationale et universitaire (BNU) de Strasbourg et l'Institut national de l'histoire de l'art (INHA), ce projet soulève quelques questions majeures : tout d'abord le traitement de volumes aussi importants (le nombre total d'images qui seront ainsi distinguées dans Gallica est estimé à plus de 100 millions) suppose une puissante machine *ad hoc*, et donc une approche raisonnée des entraînements nécessaires et du volume de nouvelles données générées. De plus, pour garantir la juste compréhension des résultats des recherches futures, la BnF mettra l'accent non seulement sur les tests préalables, mais aussi sur l'interface utilisateur et sur l'environnement documentaire des résultats. Elle y veille déjà en ce qui concerne l'OCR : Gallica indique le taux de reconnaissance atteint pour tel ou tel document et un lien est présent pour ceux qui souhaitent en savoir plus. Ce souci relève de la littératie ou « habileté numérique » : l'objectif est d'inviter les utilisateurs à prendre conscience des biais inhérents aux ressources et à compléter leurs approches. Enfin, quelle que soit la solution technique retenue, le respect des données personnelles et des contenus protégés par la propriété intellectuelle sera essentiel (comme il l'est actuellement), *a fortiori* dans le cas où ces technologies seront appliquées à la collection du dépôt légal numérique dans Gallica intra muros.

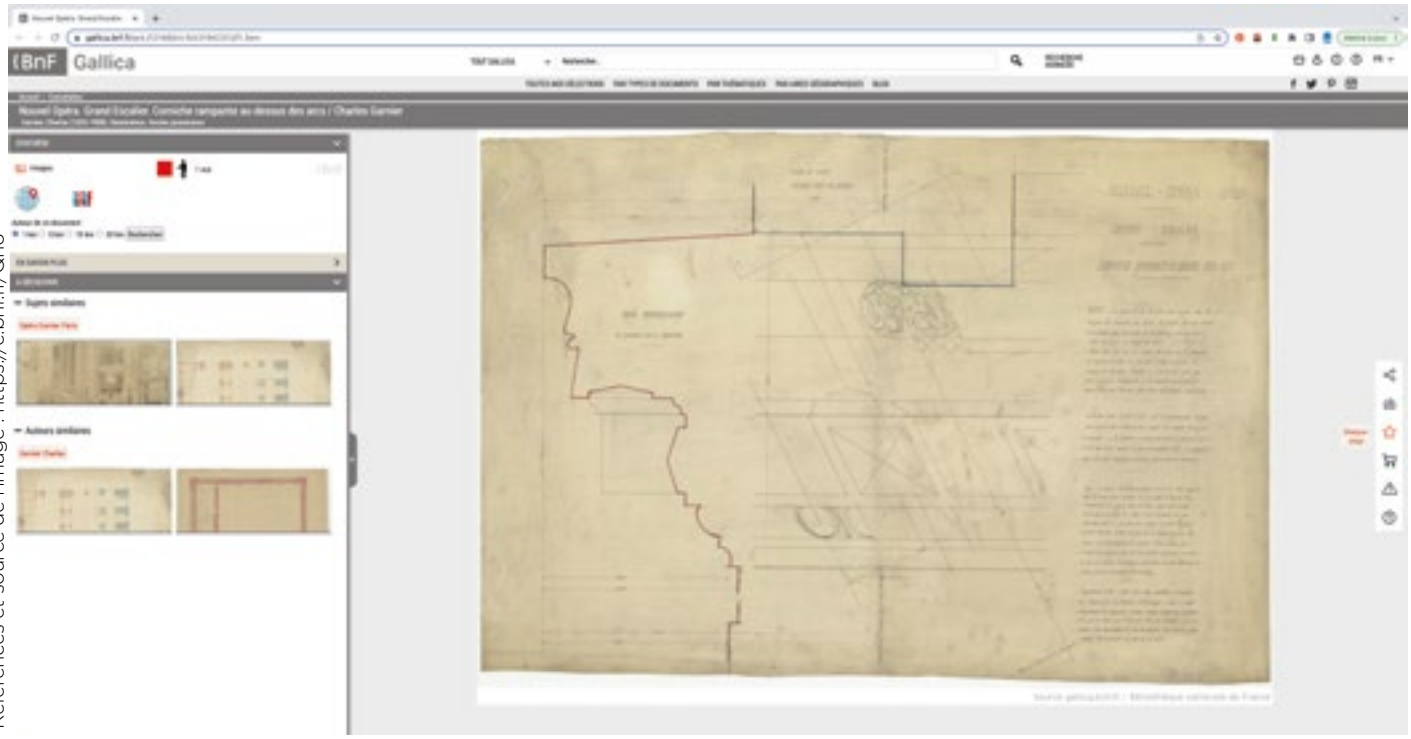
La reconnaissance de l'écriture manuscrite (HTR)

Si les caractères imprimés font à présent l'objet d'une reconnaissance industrialisée grâce à des technologies matures (OCR), il n'en va pas de même des écritures moins standardisées ou plus rares (écritures manuscrites de différentes époques, mais aussi imprimés anciens, tapuscrits, textes en langues rares…) : le repérage d'un lieu, d'un nom de personne ou d'un simple mot courant dans les manuscrits de Gallica passe surtout, aujourd'hui, par une lecture cursive des textes et non par des outils de recherche plein texte. Prenons l'un des plans que fit Charles Garnier du grand escalier de son opéra : avec ses différents titres, avec ses mesures et annotations verticales, il pose quelques défis à l'HTR.
Comme le projet de fouille d'images, le projet d'HTR s'appuie sur les expérimentations des années 2010. Il s'agit d'entraîner un système à partir d'un échantillon représentatif d'un corpus homogène en lui fournissant une transcription manuelle, puis d'étendre la transcription à l'ensemble du corpus de manière automatisée en s'appuyant sur l'IA. Plusieurs plateformes – en particulier eScriptorium et Transkribus – peuvent aujourd'hui être utilisées à ces fins. Chaque type d'écriture (voire chaque main) ayant ses spécificités, la fourniture de la première transcription peut nécessiter des compétences pointues en paléographie ou en liaison avec le contenu. En plus des questions éthiques soulevées par le projet de fouille d'images, qu'il partage, le projet d'HTR nous invite donc à considérer avec attention la phase d'entraînement des algorithmes, qui nécessite un important travail humain.

> *"Le numérique apparaît comme une véritable culture."*

L'assistance au catalogage

La BnF gère quotidiennement l'arrivée de centaines de documents de toute nature, dont la description est essentielle à la visibilité des ressources disponibles et à la satisfaction des besoins documentaires des utilisateurs, à commencer par les chercheurs. Ce travail bibliographique des catalogueurs alimente un écosystème de données de qualité dont profitent les moteurs de recherche, ainsi qu'un dialogue fécond avec de nombreuses structures, en particulier avec les industries culturelles et créatives. L'intelligence artificielle alimente l'espoir d'un gain de productivité important dans ce domaine. Il est par exemple possible d'imaginer que, en analysant le fichier d'un document numérique, on aide le catalogage tant du document numérique que du document imprimé arrivés tous deux par la voie du dépôt légal. Cependant, l'introduction de l'IA dans des processus de catalogage complexes n'est pas simple.
Les deux principales questions éthiques qui se posent ici sont celle de l'ouverture (comment, dès le début, envisager la perspective de mettre à la disposition d'autres utilisateurs des algorithmes spécifiquement développés pour ou par la BnF) et celle de l'implication de l'humain dans le processus, afin de garantir les responsabilités en cas de défaillance de l'algorithme (ce qui suppose par



Références et source de l'image : https://c.bnf.fr/Qn6

Charles Garnier, Grand escalier : corniche rampante au-dessus des arcs.

exemple des protocoles de validation), afin de favoriser le travail collaboratif et surtout afin de limiter la « fracture numérique », qui peut être considérée selon deux angles : celui des compétences, de l'aisance face aux outils, et celui de l'identité professionnelle dès lors qu'une partie des activités se voit assistée par la machine et que les tâches habituelles se déplacent.

La recommandation personnalisée dans Gallica

Enfin, un projet de recommandation personnalisée pourrait venir pallier les insuffisances du moteur de Gallica, occasionnées notamment par le choix fait de ne pas utiliser les données des utilisateurs (historiques de recherche, etc.). L'intelligence artificielle pourrait compléter la puissance du moteur par un dispositif de recherche inédit dans un cadre qui respecte la déontologie actuelle. Ainsi, le travail de délégation à l'IA de certaines tâches ou fonctionnalités se ferait avec toutes les garanties, par exemple en proposant aux utilisateurs de choisir s'ils veulent recourir ou non à la fonctionnalité de recommandation personnalisée.

Une question d'éthique

De même que, au-delà des seules questions techniques, le numérique doit être considéré dans toutes ses composantes et apparaît comme une véritable culture, source d'une patrimonialisation d'un genre nouveau à la BnF, de même l'intelligence artificielle trouve dans les bibliothèques un espace de développement naturel, au croisement des humanités et des technologies. Les principes éthiques liés à l'introduction de l'IA – transparence, explicabilité, justice (équité, égalité) et sobriété – ne sont pas éloignés des valeurs fondamentales de la BnF, qui depuis des décennies alimentent la confiance des usagers dans l'institution.

Références

- Conférence « Futurs fantastiques 2021 », coorganisée par la BnF et l'Université Paris-Saclay avec la communauté AI4LAM et congrès 2022 de l'IFLA (International Federation of Library Associations and Institutions)
- Stanford Question Answering Dataset par The Natural Language Processing Group
- Ministère de la Culture, « La France et le Québec dévoilent une stratégie commune pour améliorer la découvrabilité des contenus culturels francophones en ligne », communiqué de presse, 30 novembre 2020
- Céline Leclaire et Lucie Termignon, « Pour une éthique de la recommandation personnalisée à la Bibliothèque nationale de France », congrès de l'IFLA, 2022
- Milad Doueihi, *Pour un humanisme numérique*, Seuil 2011
- Emmanuelle Bermès, *Le numérique en bibliothèque : naissance d'un patrimoine – l'exemple de la Bibliothèque nationale de France (1997-2019)*, thèse de doctorat, 2020