En automatisant certaines tâches et en suscitant de nouvelles applications, l'IA offre aux bibliothèques un potentiel immense pour donner un nouveau souffle à leurs contenus, métadonnées ou documents numérisés.

L'intelligence artificielle, *une ouverture du champ des possibles*

L'intelligence artificielle (IA) fait l'objet d'un intérêt tout particulier depuis quelques années. Pourtant, quand on l'envisage comme un ensemble de processus logiciels permettant d'effectuer des opérations d'analyse ou de décision que des humains seraient normalement susceptibles de réaliser, on se rend compte qu'elle est présente dans le paysage numérique depuis maintenant un bon demi-siècle. C'est une plus grande accessibilité de librairies logicielles, couplée à un accroissement des moyens de calcul, qui caractérise la période la plus récente. Ces composants logiciels permettent à plus de disciplines de s'en approprier les mécanismes et de les appliquer à de nouveaux contextes. Les bibliothèques n'échappent pas à ce mouvement et de nombreux projets ont montré le potentiel de l'IA pour donner un nouveau souffle aux contenus numériques, métadonnées ou documents numérisés disponibles dans les établissements.

Dans les années 1980, les modèles qui avaient le plus de succès reposaient sur des méthodes logico-symboliques qui manipulaient des données, alors vues comme des concepts liés entre eux par des relations (ou prédicats logiques). Les modèles les plus récents s'articulent, eux, autour de méthodes statistiques par apprentissage. Ces méthodes reposent sur des architectures logicielles auxquelles on soumet de grandes quantités d'exemples et qui vont par itérations successives en abstraire les distributions statistiques, dans le cas d'apprentissages dit non supervisés, ou en généraliser l'analyse sur la base d'annotations préalablement fournies, dans le cas d'apprentissages supervisés.

Crédit Adobe stock

DE NOUVEAUX CHAMPS DES POSSIBLES EN MATIÈRE D'USAGE DES CONTENUS

Comme on peut le constater dans les différentes contributions à ce numéro d'*Arabesque,* l'IA est susceptible d'être présente dans une large gamme d'applications touchant aux domaines des bibliothèques ou des institutions patrimoniales. Si l'IA permet dans certains cas d'automatiser des tâches plus ou moins complexes déjà effectuées manuellement ou semi-automatiquement, elle fait également survenir de nouvelles applications qui redéfinissent le champ des possibles en matière d'usage des contenus. Nous voyons par exemple apparaître différents types d'applications qui viennent soutenir l'organisation des fonds existants ou accompagner les processus de numérisation.

Les premières applications intégrant des techniques d'apprentissage machine (*machine learning*) ont été utilisées pour accompagner les activités de catalogage, notamment pour l'indexation ou la classification automatiques de contenus. Cependant, les plus importantes avancées offertes par l'IA dans le domaine patrimonial sont liées à la création et l'enrichissement de contenus sur la base des opérations de numérisation conduites dans ces institutions depuis plusieurs années. Ainsi, les progrès extrêmement rapides de la reconnaissance automatique d'écriture manuscrite, avec la mise à disposition d'environnements libres tels que eScriptorium/Kraken[1], offrent la perspective d'accéder à l'intégralité des textes contenus dans de larges collections manuscrites. Des projets récents en collaboration avec les Archives nationales (LectAuRep[2]) ou la Bibliothèque nationale de France (Gallicorpora[3]) ont ainsi démontré tout le potentiel de telles techniques. Plus récemment, les

travaux menés autour de la suite GROBID permettent d'envisager de reconstituer la structure logique de documents numérisés, qu'il s'agisse d'entrées de dictionnaires ou encore de catalogues de ventes avec le projet DataCatalogue[4] en lien avec la BnF.
Enfin, les méthodes d'apprentissage profond (*deep learning*) ont permis de créer des modèles génériques de codage des informations présentes dans des images ou des textes par simple apprentissage non supervisé. Il s'agit souvent de techniques dites de masquage qui forcent le modèle à prédire un élément graphique ou linguistique en fonction d'un contexte qui lui est fourni. Ces modèles (on parle par exemple de BERT ou de GPT3), même s'ils sont parfois invisibles dans les applications concrètes, jouent un rôle essentiel en termes de performance. Ils font aussi l'objet de critique ou d'analyse quand on constate les biais qu'ils peuvent porter en eux, en lien avec la nature des données d'apprentissage utilisées.

LES CORPUS DE QUALITÉ, INDISPENSABLES À L'IA

La performance des différentes applications mentionnées ci-dessus reposent évidemment sur des modèles informatiques appropriés, associés à des capacités de calcul suffisantes, mais avant tout, elle découle directement de la production en amont de corpus de données de qualité. Ces données servent à la fois à entraîner les modèles d'apprentissage mais aussi à les tester pour en évaluer les résultats. Elles sont en général coûteuses à réunir, à nettoyer et à documenter correctement (origine, contenu, nature des annotations). C'est pourquoi on ne peut s'engager dans des activités intégrant de l'intelligence artificielle sans identifier très tôt une stratégie de gestion et si possible d'ouverture des données, qu'il s'agisse de données génériques issues du Web – par exemple le corpus OSCAR[5] – ou des données spécialisées telles que celle produites dans le cadre du projet LectAuRep avec les Archives nationales. La mise en commun de telles données passe souvent par l'établissement d'infrastructures de partage comme c'est le cas pour la reconnaissance d'écriture manuscrite avec l'initiative *HTR-United*. Enfin, dans une perspective plus large d'ouverture des données et de reproductibilité, il faut pouvoir associer à tout résultat d'entraînement non seulement les données source mais aussi les paramètres d'apprentissage (qui pilotent le comportement des modèles informatiques) et bien sûr les modèles obtenus. De cette façon, ils pourront être réutilisés ou comparés avec les résultats d'autres équipes.
Alors que du point de vue de la recherche en informatique le domaine semble encore en pleine ébullition, il est difficile d'effectuer des prédictions précises sur les enjeux de recherche à venir. Si nous nous restreignons au lien entre IA et gestion des données patrimoniales, il y a clairement des progrès importants à faire pour faciliter son appropriation et son utilisation dans des environnements disposant de moindres ressources informatiques. Cela passe probablement par un investissement plus important dans les normes de représentation des données et d'interfaçage des processus d'IA dans des logiciels métiers. Il semble aussi essentiel d'aller vers des modèles plus sobres pour faciliter leur usage en dehors de grosses plateformes de calcul avec comme effet supplémentaire, mais non négligeable, d'en réduire l'empreinte carbone.

PENSEZ AUTREMENT LE NUMÉRIQUE

Pour les institutions patrimoniales, l'arrivée massive de l'intelligence artificielle dans leur processus de numérisation crée une réelle révolution intellectuelle et organisationnelle qu'il est indispensable d'anticiper et de bien intégrer à leurs missions plus classiques. Comme on l'a vu rapidement dans cette introduction, il ne s'agit plus de concevoir ces processus comme des logiciels à l'ancienne, dont on peut confier la réalisation à son département informatique ou à une sous-traitance sélectionnée à l'occasion. La mise en œuvre d'une application reposant sur l'apprentissage automatique implique de gérer sur le moyen terme non seulement des algorithmes, mais aussi des données de référence (la vérité de terrain) dont la sélection, la description ou l'enrichissement par le biais de campagnes d'annotation doivent intégrer en continu les spécialistes métier. Par ailleurs, il faut identifier des moyens de calculs proportionnés qui permettront de bien gérer les processus d'apprentissage machine en relation avec les volumes de données à traiter. Elles devront enfin définir des stratégies de R&D qui puissent intégrer l'évolution rapide de l'état de l'art en la matière, probablement sur la base de collaborations stratégiques avec des laboratoires de recherche publics.
Avant tout, les institutions concernées devront se donner la capacité de penser autrement le numérique en leur sein, pour ne pas simplement (bêtement, dirait-on…) le voir comme un appendice aux logiciels existants, notamment de gestion des informations ou de consultation par les usagers, mais bien de repenser l'ensemble du dispositif autour des données dans un continuum où catalogues et contenus sont susceptibles d'être à la fois consultés par les humains et analysés par des machines.

Alix Chagué
Doctorante en humanités numériques au sein de l'équipe ALMnaCH (Inria – Paris) et du GREN (université de Montréal)
alix.chague@inria.fr

Laurent Romary
Directeur de la culture et de l'information scientifiques, Inria
laurent.romary@inria.fr

[1] Voir l'article p.25 : « *eScriptorium : une application libre pour la transcription automatique des manuscrits* ».

[2] https://lectaurep.hypotheses.org

[3] www.bnf.fr/fr/les-projets-de-recherche#bnf-gallic-orpor-a

[4] https://hal.inria.fr/hal-03618381

[5] https://oscar-corpus.com