Ar(abes)ques
OCTOBRE - NOVEMBRE - DÉCEMBRE 2022

DOSSIER

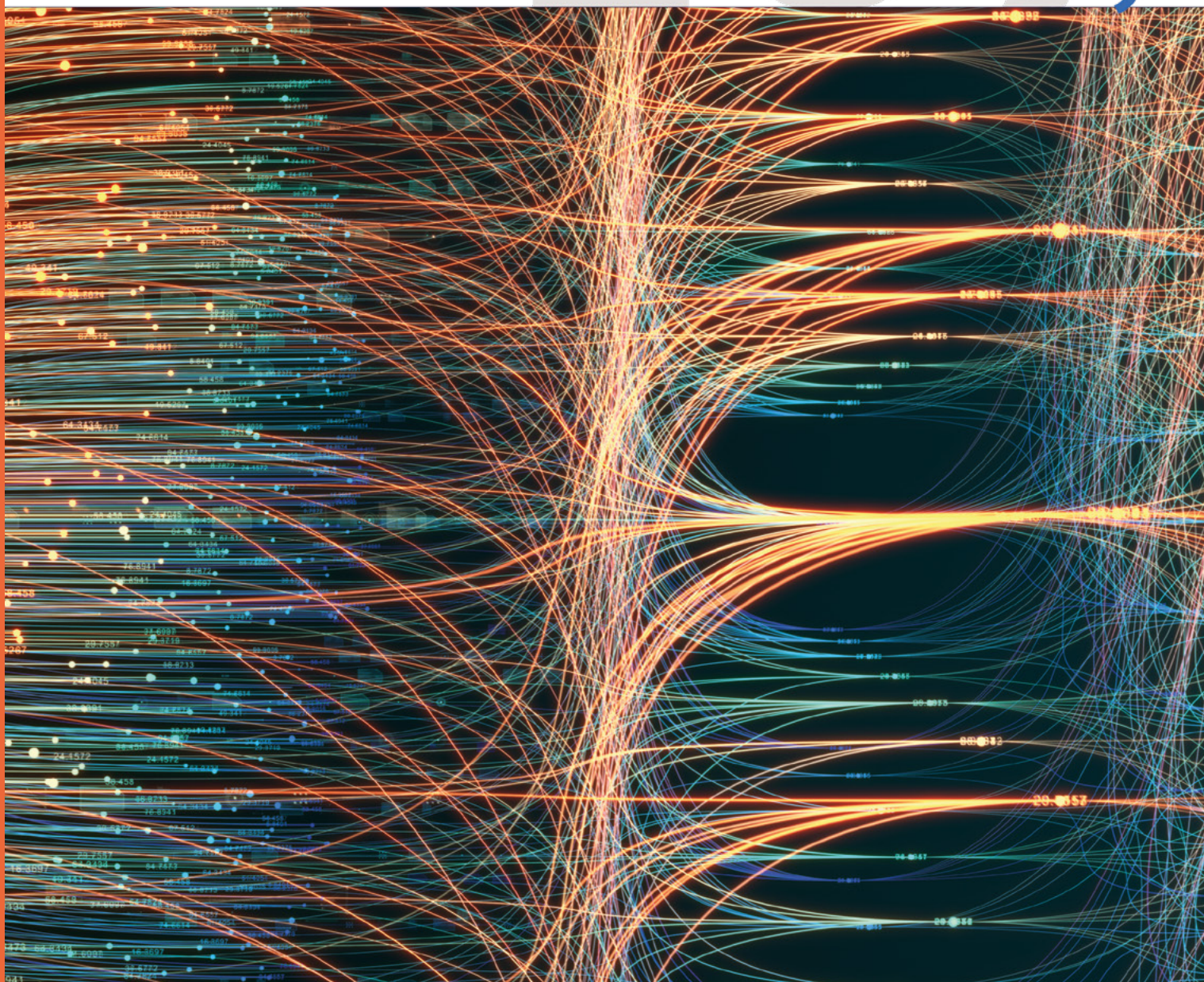
L'intelligence artificielle
Quand la machine se met au service des collections

107

PLEINS FEUX SUR • La Bibliothèque Sorbonne Nouvelle

SYSTÈME D • Transkribus : *l'intelligence artificielle au service du patrimoine documentaire*
eScriptorium : *une application libre pour la transcription automatique des manuscrits*

abes
agence bibliographique de l'enseignement supérieur



(Dossier) L'intelligence artificielle

Présente dans le numérique depuis une cinquantaine d'années déjà, l'intelligence artificielle (IA) conquiert progressivement tous les secteurs d'activité de nos sociétés. Si son application dans le domaine de la documentation en est encore à ses débuts et les possibilités qu'elle ouvre, largement à explorer, le stade de l'expérimentation est cependant dépassé. L'intelligence artificielle intègre peu à peu le quotidien des bibliothécaires, conduisant à repenser les pratiques professionnelles, le numérique au sein d'un établissement, le traitement des données et des collections qui, comme le soulignent Alix Chagué et Laurent Romary dans l'article inaugural, deviennent susceptibles d'être à la fois consultées par les humains et analysées par des machines. Signe de l'intérêt grandissant de la part des professionnels de la documentation pour l'IA, une communauté participative et internationale, baptisée ai4LAM, s'est créée en 2018 pour faire connaître les applications des technologies liées à l'intelligence artificielle dans les domaines des bibliothèques, des archives et des musées. Le groupe francophone créé au sein de cette communauté en juin dernier avec, notamment, l'objectif de mettre à disposition des documents en français et de partager les expériences, devrait contribuer à l'appropriation de ce nouveau champ par les professionnels français.
Bonne lecture !

Ar(abes)ques

REVUE TRIMESTRIELLE DE L'AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR,
227, avenue du Professeur-Jean-Louis-Viala, CS 84308, 34193 Montpellier cedex 5.
Tél. 04 67 54 84 10 / Fax 04 67 54 84 14 / https://abes.fr
Directeur de la publication : David Aymonin.
Coordination éditoriale et secrétariat de rédaction : Véronique Heurtematte.
Comité de rédaction : Christophe Arnaud, Aurélie Faivre, Christine Fleury, Étienne Naddeo, Morgane Parra, Laurent Piquemal, Marie-Pierre Roux.
Iconographie rassemblée par Christophe Arnaud.
Conception graphique : Anne Ladevie (anneladevie.com). Impression : Pure Impression

Couverture : AdobeStock
Les opinions exprimées dans* Arabesques *n'engagent que la responsabilité de leurs auteurs.

ISSN (papier) 1269-0589 / ISSN (web) 2108-7016

Bonjour l'IA

Conformément à l'un des axes forts de son projet d'établissement 2018-2023 (prolongé d'un an à cause de vous savez quoi…), l'Abes s'emploie à «rendre le paysage de l'IST plus lisible». C'est à ce titre qu'elle vous livre aujourd'hui un nouveau numéro de la revue *Arabesques,* consacré à l'intelligence artificielle et dans lequel plus de 20 auteurs ont accepté de rédiger autant de contributions instructives et constructives. Qu'ils et elles en soient ici chaleureusement remercié-e-s.

Chaque article vous révèlera une facette de cette technologie relative aux bibliothèques et à l'information scientifique.

L'IA a plus de 50 ans mais ses concepts sont encore mal connus alors que ses applications concrètes sont aujourd'hui nombreuses, notamment dans le domaine de l'information. C'est donc faire œuvre utile que d'expliquer tout ceci de manière simple et illustrée d'exemples réels, pour partager avec les spécialistes des bibliothèques le savoir de base qui leur permettra d'apprivoiser le sujet.

Petit guide de lecture : n'ayez pas de complexe ! Pour appréhender le sujet, vous pouvez entrer dans ce numéro d'*Arabesques* par n'importe quel article, selon votre connaissance préalable de l'IA ou votre manière d'apprendre : en allant du concret au conceptuel, ou l'inverse. Je vous invite ensuite à lire toutes les contributions pour vraiment saisir la nature et les promesses de l'IA au service des archives, des bibliothèques et des musées.

Un autre point fort de ce numéro vient de ce qu'il nous montre à quel point l'IA s'appuie sur un travail continu engagé dans les bibliothèques il y a des années, et sur l'intégration de multiples sources de données et d'information, patiemment construites, alimentées et enrichies par des professionnels (comme Istex, Gallica, les catalogues, Wikidata, Pascal, Francis, etc. et bien sûr tous les textes produits, collectés, préparés et annotés, y compris les manuscrits).

L'IA est arrivée à un point où ses applications documentaires induisent déjà une évolution positive de nos métiers. *Arabesques* s'en fait l'écho en cet automne 2022, tout comme la revue ID2[1] de l'ADBS en juillet dernier. La création d'un chapitre francophone de ai4lam (voir l'interview page 21) en est également un indice. L'IA est entrée dans les bibliothèques, qu'on se le dise !

À la fin de cet éditorial, le dernier sous ma signature, je souhaite attirer votre attention sur un paragraphe du portrait de Gwenaëlle Marchais qui nous donne son image de l'Abes : une accolade « pour illustrer le contrat que l'Abes remplit chaque jour, qui est de mettre une équipe d'experts au service d'une communauté ». Merci chère Gwenaëlle, c'est la plus belle reconnaissance de nos efforts.

David Aymonin
Directeur de l'Abes

[1] www.adbs.fr/boutique/i2d-la-revue-de-ladbs/i2d-1-2022-lintelligence-artificielle/1140.html

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

Le projet ISSA : *l'intelligence artificielle au service de la recherche bibliographique*

Porté par trois institutions, ISSA, projet d'indexation automatique des publications d'une archive scientifique ouverte, a été conçu comme un outil d'aide aux recherches bibliographiques complexes.

Lauréat de l'appel à projet CollEx-Persée[1] en 2020, le projet ISSA[2] – Indexation Sémantique d'une archive scientifique et Services Associés pour la science ouverte – est porté par trois institutions : le Cirad[3], Inria Sophia Antipolis Méditerranée[4] et IMT Mines Alès[5]. La motivation d'origine, portée par un besoin d'indexation automatique des publications d'une archive scientifique ouverte, s'est rapidement enrichie avec des objectifs plus ambitieux de services de recherche et de visualisation innovants. Les administrateurs d'archives ouvertes gèrent une grande quantité de métadonnées parmi lesquelles les mots-clés qui viennent décrire les publications. Cette indexation est réalisée manuellement la plupart du temps, soit par les déposants eux-mêmes (mots-clés libres en général), soit par des documentalistes spécialistes qui utilisent des descripteurs thématiques ou géographiques issus d'un vocabulaire contrôlé ou d'un thésaurus. Cette activité est exigeante et chronophage, et l'automatisation de l'indexation constitue un besoin clairement identifié par les services d'information scientifique et technique (IST) ou les bibliothèques.

Par ailleurs, ces dernières années, plusieurs évolutions ont radicalement transformé la façon dont les chercheurs et les professionnels en IST interagissent avec la littérature scientifique. En effet, la quantité de publications augmente en flèche, que ce soit dans les revues, les conférences ou par le biais de dépôts de prépublications (par exemple arxiv.org), de sorte qu'il est de plus en plus difficile de trouver des articles correspondants à des critères de recherche parfois très spécifiques.

LA PLACE DES ARCHIVES OUVERTES DANS L'ÉCOSYSTÈME DE LA LITTÉRATURE SCIENTIFIQUE

Dans ce contexte, les archives scientifiques ouvertes jouent un rôle central pour appuyer les recherches bibliographiques. Cependant, les services de recherche classiques à base de mots-clés proposés nativement par les plateformes ne parviennent souvent pas à saisir la richesse des associations sémantiques entre les articles, de sorte que certaines recherches complexes trouvent difficilement des réponses. Il est donc nécessaire de développer de nouveaux outils qui permettent aux utilisateurs de s'orienter dans cette masse de connaissances. Pour relever ces défis, le projet ISSA, guidé par les objectifs de la science ouverte et s'adossant aux principes FAIR, vise à :

- Fournir un pipeline intégré, générique et réutilisable pour l'analyse et le traitement des articles d'une archive scientifique ouverte
- Traduire le résultat en un index sémantique représenté sous la forme d'un graphe de connaissance RDF[6]
- Développer des services de recherche et de visualisation innovants qui exploitent cet index sémantique pour permettre aux utilisateurs d'explorer les règles d'association thématique, les réseaux de copublications, les articles avec des sujets cooccurrents, etc.

AGRITROP, CAS D'USAGE DU PROJET ISSA

Pour démontrer la pertinence et l'efficacité de la solution, le projet ISSA s'appuie sur un cas d'usage qui sert de preuve de concept : Agritrop[7], l'archive ouverte institutionnelle du Cirad, contenant plus de 110 000 ressources dont 12 000 articles en libre accès, spécialisée dans les domaines de l'agronomie, de la biodiversité et du développement durable. Le thésaurus multilingue Agrovoc[8], géré par l'Organisation des nations unies pour l'alimentation et l'agriculture, est utilisé pour l'indexation comme vocabulaire de référence spécifique au domaine.

Le processus de construction du graphe de connaissance (ou index sémantique) fait appel à plusieurs techniques d'intelligence artificielle : traitement du langage naturel, ingénierie des connaissances, Web sémantique et données liées. La première étape consiste à récupérer les informations contenues dans l'archive ouverte grâce au protocole OAI-PMH[9]. Dans un premier temps, toutes les métadonnées récupérées sont transformées au format RDF et viennent peupler l'index sémantique : titre, auteurs, résumé, licence, date, langue, identifiants de la publication, lien les PDF en accès libre, etc. Par la suite, les données textuelles des articles telles que le titre, le résumé ou le corps du texte sont traitées afin d'en extraire

[1] www.collexpersee.eu
[2] https://issa.cirad.fr
[3] www.cirad.fr
[4] https://inria.cl/en/centre-inria-sophia-antipolis-mediterranee
[5] https://www.imt-mines-ales.fr
[6] Resource Description Framework : langage de base du Web sémantique développé par le W3C
[7] https://agritrop.cirad.fr
[8] www.fao.org/agrovoc
[9] www.openarchives.org/pmh

automatiquement des descripteurs thématiques et géographiques et des entités nommées, c'est à dire des mentions d'entités reconnaissables dans le texte. Descripteurs et entités nommées sont liés à des bases de connaissance généralistes comme Wikidata[10] et DBpedia[11], géographiques comme GeoNames[12] ou encore à des ressources terminologiques plus spécifiques adaptées à un domaine scientifique donné, par exemple le thésaurus Agrovoc dans le cas d'Agritrop. Ces informations sont transformées en RDF et viennent enrichir à leur tour le graphe de connaissance qui contient alors toutes les informations utiles à la description des publications de l'archive – métadonnées classiques et pour les articles en accès libre, descripteurs thématiques et entités nommées liées. L'ensemble, décrit selon les formats du web sémantique, est naturellement relié au Web des données et interrogeable via un point d'accès SPARQL (langage de requête de données RDF). Les connaissances de milliers de publications produites par des milliers de chercheurs se retrouvent ainsi connectées, publiées sur le Web et interrogeables !

PROPOSITION DE SERVICES À VALEUR AJOUTÉE

Le graphe de connaissance sert de clé de voûte au développement d'outils de recherche et de visualisation.

Un premier résultat quasi immédiat est la possibilité de consulter les notices de l'archive ouverte par le biais d'une visualisation enrichie : métadonnées classiques, résumé avec entités nommées surlignées et liens vers les bases de connaissance, affichage des descripteurs obtenus automatiquement, visualisation cartographique des entités nommées géographiques du texte.

Deux autres outils de visualisation permettent d'aider à la résolution de requêtes complexes :

- ARViz extrait et visualise des règles d'association reliant les descripteurs thématiques des articles.

La *Figure 1* illustre comment les concepts mentionnés dans les articles de l'archive peuvent être utilisés pour découvrir et visualiser les règles d'association. Dans l'exemple, les articles mentionnant les concepts Covid-19 et sécurité alimentaire (a) mentionnent fréquemment le concept de pandémie (b).

- LDViz permet quant à lui d'explorer les réseaux sémantiques formés par des entités aussi variées que des descripteurs thématiques, des auteurs, des institutions, etc. En visualisant ces réseaux, LDViz permet aux utilisateurs de résoudre des questions de compétence complexes.

Avec différentes techniques de visualisation, la *Figure 2* montre comment un utilisateur peut rechercher des articles mentionnant le concept de santé ou l'un de ses sous-concepts (a) et (b), découvrir qu'il est souvent mentionné avec le changement climatique (c), et obtenir la liste des publications

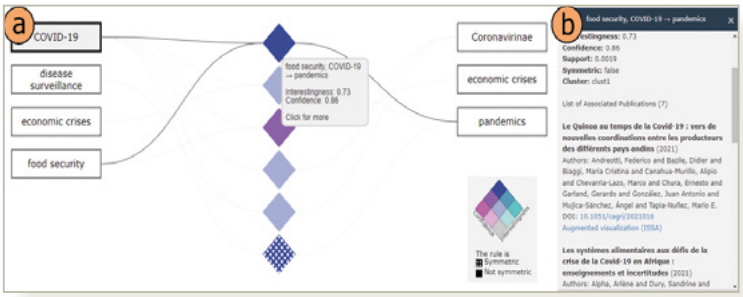


Figure 1 - Recherche par règles d'association

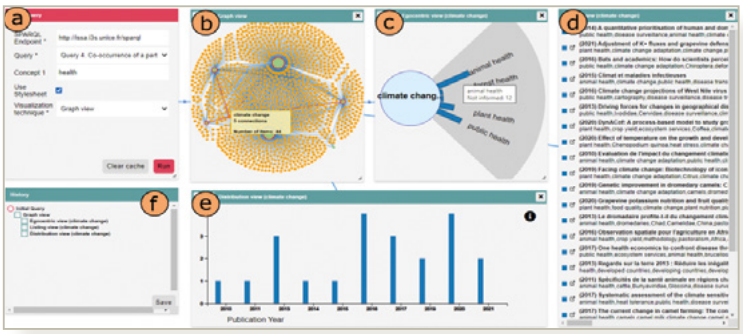


Figure 2 - Recherche par cooccurrence de concepts

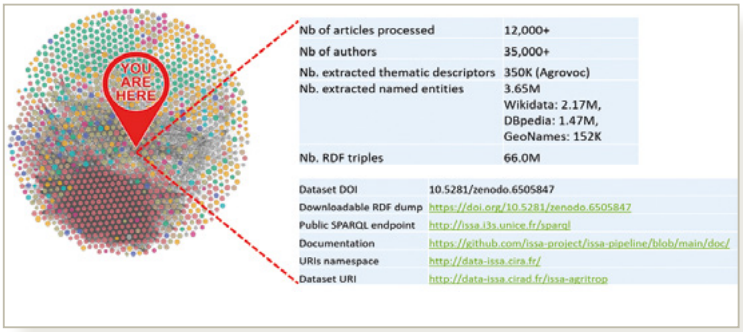


Figure 3 - Le jeu de données ISSA Agritrop dans le LOD

associées (d) et leur répartition dans le temps (e).

S'inscrivant pleinement dans la dynamique de la science ouverte et des données *FAIR*, le travail présenté est rendu disponible sous licence ouverte avec tous les documents d'accompagnement nécessaires pour faciliter sa réutilisation. Le jeu de données ISSA généré dans le cas Agritrop est publié sur le *Linked Open Data Cloud* également sous licence ouverte (*Figure 3*).

Anne Toulet
Coordinatrice scientifique du projet ISSA pour le Cirad
anne.toulet@cirad.fr

Franck Michel
Coordinateur scientifique du projet ISSA pour Inria
fmichel@i3s.unice.fr

Andon Tchechmedjiev
Coordinateur scientifique du projet ISSA pour IMT Mines Alès
andon.tchechmedjiev@mines-ales.fr

[10] www.wikidata.org
[11] www.dbpedia.org
[12] www.geonames.org

Vers l'intelligence artificielle et au-delà !
Une feuille de route pour la BnF

En 2020, la Bibliothèque nationale de France s'est dotée d'une feuille de route sur l'intelligence artificielle afin de mieux répondre aux nombreux défis, technologiques, professionnels, culturels, éthiques posés par l'IA.

Depuis plusieurs années déjà, la Bibliothèque nationale de France (BnF) travaille avec des partenaires du secteur académique sur des expérimentations à base d'intelligence artificielle (IA), notamment dans le domaine de l'OCR et de l'analyse d'images *(computer vision)*. Elle a mis en place avec l'IR Huma-Num le BnF DataLab, qui vise à développer de nouveaux usages de recherche sur les collections numériques massives, mobilisant entre autres la fouille de données et l'apprentissage machine. Cependant, si ces projets expérimentaux portant sur des corpus réduits montrent des résultats satisfaisants, les industrialiser à l'échelle des 9 millions de documents de Gallica est une autre affaire.

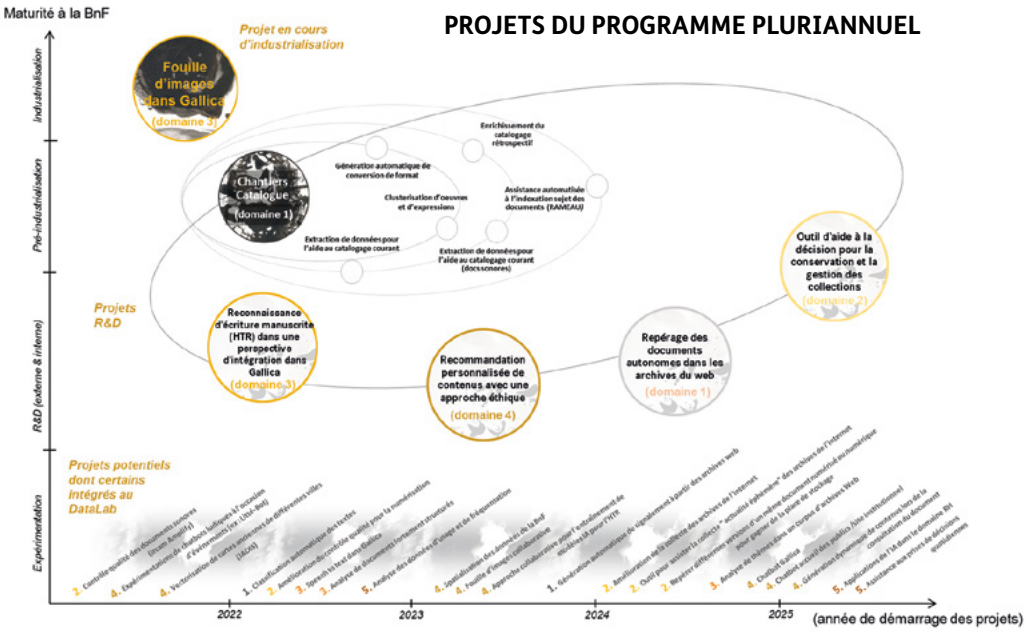
Plus généralement, l'intelligence artificielle représente un défi à bien des égards pour une institution comme la BnF : parce que le terme est devenu omniprésent sans pour autant être clairement défini, parce que la technologie montre une grande maturité dans le privé mais que les cas d'usage propres aux bibliothèques sont encore en devenir, parce qu'elle représente des investissements considérables et remet en cause l'infrastructure, la gouvernance des données, l'organisation du travail ; enfin, parce qu'elle pose de nombreuses questions éthiques quant à son impact sur l'environnement et plus encore, sur l'humain.

En 2018, la première conférence « Fantastic Futures » organisée par la Bibliothèque nationale de Norvège et la bibliothèque de Stanford nous a convaincus de l'importance de nous doter d'une feuille de route pour nous mettre en ordre de marche et accueillir le potentiel transformateur de l'IA. Le besoin s'est donc fait sentir de formaliser la stratégie de l'établissement dans ce domaine afin d'éclairer la route pour les années à venir : c'était l'objectif de la lettre de mission confiée en septembre 2020 par le directeur général de l'établissement à Emmanuelle Bermès, premier signe d'une volonté institutionnelle d'affirmer cette orientation nouvelle.

CINQ ACTIONS POUR SE LANCER

Élaborée au moyen d'une série d'ateliers réunissant les collègues intéressés, d'une enquête interne, d'un parangonnage international et d'une veille extensive, la feuille de route synthétise le chemin à parcourir en cinq actions.

1. La première action à entreprendre consiste à inscrire l'IA dans la stratégie de l'établissement. Alors que se posait justement la question du renouvellement du contrat d'objectifs et de performance avec la tutelle pour la période 2022-2027, cette opportunité a été saisie pour faire de l'IA une dynamique transverse du nouveau contrat, susceptible d'apporter des pistes d'innovation dans tous les domaines métier de la bibliothèque : des entrées numériques à la médiation, du signalement à la conservation, du traitement des documents numérisés à leur accès dans Gallica, de l'analyse des usages des publics aux fonctions de gestion administrative. Cette stratégie globale met l'accent, dès cette étape, sur les enjeux éthiques, en particulier la question des données personnelles, les risques liés aux biais dans les données, l'impact environnemental de cette technologie et l'accompagnement du changement. Celui-ci touche particulièrement la fonction informatique, qui doit se mobiliser pour devenir une force motrice des évolutions à venir.



Ces projets relèvent de 5 domaines

1. Aide au catalogage et signalement
2. Gestion des collections, des entrées à la conservation
3. Exploration, analyse des collections et amélioration de l'accès
4. Médiation, valorisation et éditorialisation des collections
5. Aide à la décision et au pilotage

Information et contact
www.bnf.fr/fr/feuille-de-route-ia

2. La deuxième action se focalise sur les projets de recherche et développement, ainsi que sur les mesures à prendre pour faciliter l'industrialisation des résultats lorsque ceux-ci sont convaincants.

3. La troisième action porte sur le développement des compétences, non seulement en matière d'expertise pour les agents (informaticiens ou bibliothécaires) qui auront à participer aux projets mobilisant l'intelligence artificielle, mais aussi pour l'ensemble du personnel. En effet, pour les professionnels des bibliothèques, comprendre ce qu'est l'intelligence artificielle, comment elle fonctionne et les risques qu'elle présente relève de la culture générale. Au-delà de la stratégie de la BnF sur ce sujet, il s'agit d'une question de société qui rejoint notre mission de sensibilisation des publics. Les algorithmes sont présents dans notre quotidien, ils utilisent nos données et influencent nos décisions : avoir conscience de ces mécanismes est un enjeu citoyen.

4. La quatrième action est la plus technique: industrialiser l'IA nécessite d'adapter l'infrastructure informatique, mais aussi d'agir sur la gestion des données et leur qualité, un point fort de la BnF de par sa mission mais qui se heurte encore trop, dans le système d'information actuel, au silotage des collections de nature différente (par exemple, les archives Web d'une part et la bibliothèque numérique d'autre part).

5. Enfin, la cinquième action vise à doter la BnF d'un programme pluriannuel, avec des partenaires du secteur académique mais aussi du privé, et d'autres bibliothèques, afin de créer du lien entre les projets et de mutualiser les briques qui peuvent l'être.

UNE GALAXIE DE PROJETS EN DEVENIR

Dans la feuille de route, le programme pluriannuel prend la forme d'une galaxie de projets, représentés ici en fonction de leur niveau de priorité, de la maturité de la BnF sur le cas d'usage concerné et d'une ébauche de planification. Chaque planète représente un projet considéré, durant l'étude, comme majeur.

En haut à gauche, le projet le plus immédiat mais aussi le plus mature est l'industrialisation de la fouille d'images dans Gallica, qui fait suite à plus de dix ans d'expérimentations et de R&D sur ce sujet. Financé par France Relance dans le cadre de l'appel à projet « Numérisation de l'architecture et du patrimoine », en partenariat avec l'INHA (Institut national d'histoire de l'art) et la bibliothèque nationale et universitaire de Strasbourg, le projet Gallica Images sera lancé en 2023 ; il utilisera l'IA pour segmenter et caractériser plusieurs millions de contenus iconographiques aujourd'hui peu accessibles dans Gallica.

Ensuite viennent d'autres projets jugés essentiels pour le développement de la BnF dans les années à venir : une planète dotée de nombreux satellites symbolise les expérimentations diverses à mener dans le champ des catalogues ; la reconnaissance d'écritures manuscrites (HTR) et la recommandation personnalisée avec une approche éthique sont des sujets sur lesquels la coopération avec d'autres institutions s'annonce prometteuse. Enfin, dans le champ de la conservation, outiller le futur site d'Amiens d'une intelligence de la donnée est l'une de nos perspectives de long terme, une ambition qui nécessite de poursuivre les expérimentations, telles que le projet Dalgocol récemment achevé.

Au bas du schéma, dans une zone métaphoriquement encore plongée dans des brumes d'incertitudes, sont recensés tous les autres cas d'usage qui ont émergé lors de l'élaboration de la feuille de route, mais qui n'ont pas été, à ce jour, priorisés.

Le BnF DataLab sera sans nul doute un dispositif essentiel pour leur développement.

ET APRÈS 2026 ?

Si la BnF s'est dotée, avec cette feuille de route, de jalons à atteindre et livrables à réaliser d'ici à 2026, il est certain qu'on ne pourra pas parler alors d'achèvement. Il s'agit davantage d'enclencher une dynamique : l'intelligence artificielle trouve dans les bibliothèques un terrain de développement naturel, à la croisée des humanités et des technologies, mais n'est pas pour autant une fin en soi. Élément d'une culture numérique plus vaste, elle nous amène à nous interroger sur les principes de transparence, d'explicabilité, de justice (équité/égalité) et de sobriété qui devraient régir toute innovation numérique, et qui pourront sans nul doute nous guider dans les nouveaux défis qui émergeront à l'avenir, au-delà de l'IA.

Emmanuelle Bermès

Adjointe pour les questions scientifiques et techniques auprès du directeur des services et des réseaux, Bibliothèque nationale de France
emmanuelle.bermes@bnf.fr

Céline Leclaire

Chargée de production de supports stratégiques, Bibliothèque nationale de France
celine.leclaire@bnf.fr

REPÈRES BIBLIOGRAPHIQUES

La BnF et l'intelligence artificielle, feuille de route : **www.bnf.fr/fr/feuille-de-route-ia**

« L'intelligence artificielle à la BnF », Dossier Grand angle, *Chroniques* n° 93, janvier-mars 2022. **http://chroniques.bnf.fr/pdf/Chroniques_93.pdf**

Jean-Philippe Moreux, « Recherche d'images dans les bibliothèques numériques patrimoniales et expérimentation de techniques d'apprentissage profond », *Documentation et bibliothèques,* volume 65, numéro 2, avril–juin 2019, p. 5-27. **https://id.erudit.org/iderudit/1063786ar**

Emmanuelle Bermès, Eleonora Moiraghi, « Le patrimoine numérique national à l'heure de l'intelligence artificielle. Le programme de recherche Corpus comme espace d'expérimentation pour les humanités numériques », *Revue Ouverte d'Intelligence Artificielle,* Volume 1 (2020) no. 1, pp. 89-109. **https://roia.centre-mersenne.org/articles/10.5802/roia.5**

Céline Leclaire, Lucie Termignon, « Pour une éthique de la recommandation personnalisée à la Bibliothèque nationale de France ». Présenté lors du satellite IFLA *New Horizons in Artificial Intelligence in Libraries,* 21-22 juillet 2022, Galway, Irlande. À paraître.

Philippe Vallas, « Prédire l'état matériel des documents : Dalgocol, un programme de recherche en intelligence artificielle à la BnF : entretien avec Philippe Vallas », *Bulletin des bibliothèques de France (BBF),* 2022-1. **https://bbf.enssib.fr/consulter/bbf-2022-00-0000-008**

Pensé comme un outil de pilotage et de suivi, le Baromètre de la science ouverte utilise l'intelligence artificielle pour optimiser ses missions.

L'utilisation de l'apprentissage automatique dans le Baromètre de la science ouverte : *une façon de réconcilier bibliométrie et science ouverte ?*

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté Égalité Fraternité

Dès le lancement du Plan national pour la science ouverte (PNSO) en 2018, le Baromètre de la science ouverte (BSO) a été pensé comme un outil de suivi et de pilotage de politiques publiques. D'abord centré sur l'accès ouvert aux publications, le BSO a permis en quelques mois d'objectiver un « point de départ » du taux d'ouverture des publications françaises. Le BSO a vocation à élargir son périmètre, en s'intéressant à d'autres productions que les seules publications, et à approfondir ses analyses pour fournir des éléments d'aide à la compréhension et à la décision pour ses différents utilisateurs (décideurs au niveau national ou établissement, négociateurs, financeurs, chercheurs).

UNE ALLIANCE OBJECTIVE AVEC LES OUTILS D'APPRENTISSAGE AUTOMATIQUE

D'autres outils (internationaux notamment) de suivi de la science ouverte préexistaient au BSO, alors pourquoi en créer un nouveau ? D'abord, parce que les outils existants utilisent des données issues de bases propriétaires, ni partageables, ni réutilisables et introduisent des biais de couverture. De plus, le Ministère voulait disposer d'un outil souverain, adapté à ses propres besoins de suivi. Néanmoins, le pragmatisme conduit la plupart des institutions à se tourner vers les bases de données propriétaires. Notamment, les champs disciplinaires et les affiliations ne sont pas ouvertes en général (les affiliations sont absentes de Crossref dans plus de 75 % des cas en 2021). Impossible de suivre l'ouverture des publications sans analyser les tendances d'une discipline à l'autre qui sont le reflet de pratiques différentes entre les communautés. De même, il est nécessaire de connaître les pays d'affiliation : seules les publications avec une affiliation française sont analysées dans le BSO, encore faut-il savoir si une publication a une affiliation française !

Le manque de métadonnées ouvertes est parfois vécu comme un obstacle infranchissable, plaçant certaines institutions dans la situation de devoir avoir recours à des données propriétaires, dans l'attente de la mise en place, sur un temps plus long et incertain, d'infrastructures ouvertes et centralisées de métadonnées riches. Une troisième voie mérite d'être pensée, et l'apprentissage automatique s'avère être un outil déterminant dans sa mise en œuvre.

TRANSFORMER DES BASES EXISTANTES EN DONNÉES D'APPRENTISSAGE

L'apprentissage automatique recouvre de nombreuses méthodes. Mais un invariant demeure : l'apprentissage automatique utilise des données d'apprentissage (d'entraînement) pour construire un modèle, permettant ensuite d'enrichir de nouvelles données non rencontrées dans les données d'apprentissage. Il faut donc d'une part des données d'apprentissage suffisamment riches pour construire un modèle pertinent et d'autre part des données à enrichir grâce au modèle. Ces données à traiter doivent porter un minimum d'informations, sans quoi le modèle sera incapable de calculer quoi que ce soit d'utile.

Dans le cadre du BSO, la détection d'accès ouvert repose sur l'outil (ouvert) *Unpaywall*[1]. Les problématiques principales restantes portent sur l'inférence des champs disciplinaires et des pays d'affiliation. Nous faisons l'hypothèse qu'il est possible de déterminer une discipline à partir du titre du document et de la revue. Le module *scientific tagger*[2] utilise les bases PASCAL et FRANCIS[3] comme bases d'apprentissage. Le modèle est construit avec l'algorithme *fastText*[4] qui a le mérite d'être très léger et rapide. Une approche similaire est mise en place pour inférer la langue ainsi qu'une classification spécifique au domaine biomédical[5].

Pour les pays d'affiliation, l'obstacle à franchir est plus haut : les métadonnées ouvertes ne contiennent en général pas d'information sur les affiliations. Le problème ne porte pas sur les données d'entraînement mais bien sur les données à traiter. Sans données à traiter, l'algorithme se retrouve dans une impasse. Impasse qui semble pourtant paradoxale : les affiliations sont à la fois sous nos yeux à la première page des publications et invisibles dans les métadonnées ouvertes. Un outil de collecte et

[1] https://unpaywall.org
[2] https://github.com/dataesr/scientific_tagger
[3] https://pascal-francis.inist.fr
[4] https://fasttext.cc
[5] E. Jeangirard; Content-based subject classification at article level in biomedical context; 2021; hal-03212544
[6] https://github.com/dataesr/affiliation-matcher

Crédit Adobe stock

d'analyse des pages Web publiques des publications a été développé. Il en extrait les affiliations plein texte. Le module *affiliation-matcher*[6] permet à partir d'une affiliation plein texte (Université de Paris Dauphine, France) de deviner le pays associé (France). Cela peut être plus subtil qu'il n'y paraît dans cet exemple. Ainsi, la présence du mot « France » dans l'affiliation n'est ni nécessaire (CERMICS Université Paris Est), ni suffisante (Hôtel Dieu de France, Beirut, Lebanon).
Le module *affiliation-matcher* s'appuie sur des données référentielles (notamment le RNSR et ROR[7]) qui jouent le rôle de données d'apprentissage. L'algorithme en place ne relève néanmoins pas entièrement de l'apprentissage automatique car les règles d'appariement ne sont pas décidées par la machine, mais contrôlées par l'utilisateur du module.

L'EXTENSION DU BSO AUX DONNÉES DE LA RECHERCHE ET CODES LOGICIELS AMÈNE À L'UTILISATION DE NOUVEAUX OUTILS

Le deuxième Plan national pour la science ouverte (PNSO2)[8] fixe comme objectif au BSO de proposer de nouveaux indicateurs de suivi au-delà des publications. Dans sa déclinaison santé, le BSO analyse notamment les essais cliniques. À présent, nous travaillons à l'analyse des données de la recherche et des codes logiciels. Une piste suivie consiste à tenter de repérer dans le texte des publications, les références aux logiciels et aux données de la recherche. Il faut donc avoir accès au *full-text* des publications, et disposer d'un moyen pour y repérer une mention de logiciel ou de jeu de données. C'est un niveau supplémentaire de complexité. L'accès aux *full-text* est encore très difficile (hors accès ouvert) malgré l'existence d'accords TDM (*Text and Data mining*) dans les contrats avec certains éditeurs et les dispositions liées à la fouille de texte dans un décret[9] récent. De plus, ce type d'outil de détection fait appel à des techniques dites « d'apprentissage profond » (*deep learning*). Nous travaillons avec Patrice Lopez (*science-miner*[10]), un des experts internationaux de l'utilisation des techniques d'apprentissage profond sur les textes scientifiques.

NE PAS LÂCHER LA « PROIE » DU RÉEL POUR « L'OMBRE DES VÉRITÉS ALGORITHMIQUES »[11]

Quelles que soient les données et les techniques utilisées, les algorithmes d'apprentissage automatique produisent des erreurs, que nous tentons de contrôler et de mesurer. Une étude récente de Lauranne Chaignon et Daniel Egret[12] a validé l'efficacité de la méthode de détection des affiliations françaises du BSO en menant une comparaison avec les principales bases bibliométriques.
L'utilisation de l'apprentissage automatique reste un moyen fiable de pallier le manque de métadonnées ouvertes et riches. Des réseaux d'échanges d'informations peuvent aussi se structurer, comme nous le proposons avec les déclinaisons locales du BSO[13].
Le développement des outils avancés d'enrichissement automatique ne doit donc pas nous faire oublier un objectif essentiel, la poursuite de la mise en place d'un réseau d'infrastructures ouvertes, proposant des métadonnées riches pour rendre à la communauté scientifique le contrôle de l'information qu'elle produit elle-même.

Éric Jeangirard
Data Scientist, Département des outils d'aide à la décision, SIES – MESR
eric.jeangirard@recherche.gouv.fr

[7] https://ror.org

[8] www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte

[9] www.legifrance.gouv.fr/jorf/id/JORFTEXT000045960058

[10] https://science-miner.com

[11] R. Gori, 2022, La Fabrique de nos servitudes

[12] L Chaignon, D Egret; Identifying scientific publications countrywide and measuring their open access: The case of the French Open Science Barometer (BSO). Quantitative Science Studies 2022; doi:10.1162/qss_a_00179

[13] https://barometredelascienceouverte.esr.gouv.fr/about/declinaisons

Connectôme : *vers des services de données scientifiques ouverts grâce à l'IA*

L'objectif du projet suisse Connectôme est d'organiser les métadonnées ouvertes nationales et internationales pertinentes de manière durable afin de les rendre facilement accessibles.

BIBLIOTHÈQUE CANTONALE ET UNIVERSITAIRE BCU LAUSANNE

Personne ne doute plus du fait que les données de la recherche sont un atout essentiel pour les universités. L'accélération technologique de la recherche entraînant par ailleurs une augmentation de la masse de données ouvertes qui se retrouvent stockées dans diverses archives (inter)nationales décentralisées, et dépôts institutionnels, c'est donc un grand défi pour les universités et leurs bibliothèques de savoir quelles données sont stockées et où, quelle est leur qualité, comment elles peuvent être obtenues, récoltées collectivement, réutilisées, et comment elles peuvent être liées entre les disciplines afin de pouvoir engendrer de nouvelles recherches révolutionnaires.
Afin de compléter l'infrastructure nationale, SWITCH[1], le prestataire du réseau national suisse de recherche et d'éducation, a lancé le projet Connectôme[2]. La vision de Connectôme est d'interconnecter et d'organiser les métadonnées ouvertes nationales et internationales pertinentes de manière durable dans toutes les disciplines, afin de les rendre facilement trouvables, largement accessibles, interopérables et à valeur ajoutée.

HARMONISER, ENRICHIR, INTERCONNECTER LES DONNÉES

L'objectif de l'équipe à l'origine de cette initiative est de récolter, d'harmoniser, d'enrichir et d'interconnecter les métadonnées des fournisseurs de données décentralisés, et de les faire entrer dans le graphe de connaissances de Connectôme en utilisant des normes ouvertes et les meilleures pratiques internationales. Les données ouvertes liées qui en résultent sont alors utilisées pour activer de nouvelles fonctions de recherche et de découverte susceptibles de soutenir les cycles de vie des données de la recherche et de l'éducation. La mise en place de cette infrastructure représente un nouveau domaine de développement stratégique de SWITCH. Dès son lancement, ce projet a intéressé les bibliothèques, étant donné que le Connectôme permettra d'offrir non seulement des services d'enrichissement automatique de métadonnées décrivant les publications et l'ORD, mais également la classification automatique des publications électroniques selon la systématique utilisée par une bibliothèque. Connectôme sera également en mesure de fournir des résumés des articles scientifiques aux lecteurs de niveau primaire ou secondaire, et d'ouvrir ainsi ces contenus à un nouveau périmètre de lecteurs.

UNE MISE EN ŒUVRE EN 2022

L'infrastructure a été créée en 2020 en collaboration avec 6 partenaires initiaux et étendue à 9 partenaires suisses en 2021. Certains partenaires ont été impliqués dans le développement du concept, la conception et les efforts de préparation, tandis que d'autres ont fourni un soutien par la livraison de logiciels *open source* et/ou ont offert des conseils et un retour d'information continu[3]. Des chercheurs de diverses disciplines ont été inclus dans le codéveloppement continu de l'infrastructure. Ils ont contribué au développement du graphe de connaissances en apportant des compétences sur les ontologiques spécifiques.
Début 2022, le projet est entré dans sa phase de mise en œuvre. L'accent est actuellement mis sur l'extension des partenariats ainsi que sur le développement, l'évaluation stratégique et le déploiement plus large de l'infrastructure de base et des services de données tels que décrits ci-dessous.

L'INFRASTRUCTURE DE BASE DE CONNECTÔME

L'infrastructure de base permet d'assurer des services de données de recherche ouverts, autant du point de vue de la plateforme que des personnels nécessaires pour développer, exploiter et améliorer l'infrastructure de recherche, qui comprend le mappage des données pour convertir les métadonnées en une structure de données commune (RESCS.org), le graphe de données permettant d'importer et stocker les métadonnées liées, ainsi que l'API Connectôme pour rechercher, découvrir et extraire les données ouvertes liées. L'équipe fournit un savoir-faire en matière de données afin de collaborer avec la communauté suisse de l'éducation, de la recherche et de l'innovation (ERI) à la cocréation des solutions. Les éléments-clés de l'infrastructure de base sont alignés sur les engagements internationaux pertinents, par exemple à travers la participation à deux groupes de travail d'EOSC.

Les services de données de l'infrastructure de base permettent de répondre aux besoins d'amélioration de l'utilisation des données de recherche ouverte des parties prenantes du secteur ERI. Les services de données sont créés conjointement avec la communauté et gérés par SWITCH. Quatre services de données sont actuellement en développement et en évaluation continue :

• L'enrichissement des données

Ce service fournira différents outils pour augmenter la qualité et la réutilisation des données ouvertes par l'enrichissement des métadonnées. Les cas d'usage en cours de développement sont les suivants :
– Extraction et/ou génération d'informations de localisation à l'aide d'algorithmes de reconnaissance d'entités nommées afin d'enrichir les métadonnées de la collection Memobase dans le but de réutiliser les emplacements extraits dans les cartes de l'interface utilisateur et les chronologies des services.

– Désambiguïsation des noms d'auteurs pour les métadonnées de publications et de projets de recherche à l'aide d'un réseau neuronal profond entraîné sur les noms et prénoms suisses en utilisant des données ouvertes, manuellement étiquetées et vérifiées, provenant du Fonds national suisse de la recherche scientifique.

– Enrichissement des métadonnées d'organisation en utilisant des sources de données externes telles que les propriétés de Wikidata.

– Extraction de métadonnées contextuelles à partir d'images à l'aide d'algorithmes de vision formés manuellement sur des données d'images ouvertes provenant de Wikidata en collaboration avec la Haute école spécialisée bernoise et Wikimedia Suède.

– Transformation de métadonnées d'archives (Renouvaud Patrinum) en données ouvertes liées à des fins de réutilisation (par exemple, pour afficher des données ouvertes liées connexes).

• L'API Connectôme

L'API Connectôme vise à permettre aux fournisseurs de services tels que les archives, les dépôts ou les plateformes de découverte d'interroger un graphe de données ouvertes et d'extraire des informations à des fins de recherche et d'éducation.

Les cas d'utilisation actuellement en cours de développement sont les suivants :
– Accès aux données ouvertes liées via des fonctions API génériques (par exemple, récupération, recherche, exportation).
– Exportateurs spécifiques au client qui restructurent automatiquement les données ouvertes liées dans des structures de table spécifiques pour faciliter la réutilisation[4].

• *Insights* en tant que service

Les *insights* sont des modèles significatifs dans les données qui permettent une prise de décision plus efficace.
Ce service vise à générer de nouvelles connaissances à partir de données ouvertes liées. Pour ce faire, des techniques d'intelligence artificielle sont utilisées pour analyser les données ouvertes liées dans le graphe de connaissances du Connectôme. Le public cible est constitué de fournisseurs de services et de données pour la recherche et l'éducation.

Les cas d'utilisation en cours de développement sont les suivants :
– Découverte d'informations de recherche similaires et/ou connexes et suggestions (recherche en cours de saisie) à l'aide d'algorithmes de recommandations.

– Analyse des réseaux (d'auteurs) pour identifier et/ou classer la pertinence des auteurs pour un groupe de sujets de recherche thématiques générés automatiquement et visualisation des graphes des réseaux résultants.
– Extraction automatique de mots-clés et leur catégorisation à partir de textes téléchargés et de documents scientifiques à l'aide d'algorithmes de modélisation thématique afin de découvrir des projets de recherche, des publications et des ensembles de données connexes.
– Résumé automatique et simplification des résumés des publications en *open access* à l'aide de grands modèles de langage pré-entraînés afin d'améliorer la compréhensibilité et la visibilité pour les différentes parties prenantes (par exemple, pour que les élèves et les citoyens puissent comprendre le contenu des articles de recherche).

• La plateforme de découverte

Une plateforme de découverte, actuellement utilisée à des fins de présentation, utilise l'API Connectôme et *Insights as a Service* pour améliorer l'expérience de recherche et de découverte des utilisateurs finaux. Les liens sémantiques visent à permettre aux utilisateurs finaux de rechercher des personnes, des organisations, des projets, des publications et des ensembles de données. Les fonctions d'aperçu permettent de retrouver des similitudes, des recommandations, des analyses de réseau et des visualisations. En outre, l'équipe travaille sur des simplifications automatisées et des résumés de publications, le calcul de scores d'ouverture et la recommandation de ressources sur la base de l'analyse de textes/documents.

UN OUTIL ADAPTÉ POUR LES BIBLIOTHÈQUES

Deux bibliothèques suisses sont actuellement impliquées dans la cocréation des cas d'utilisation sélectionnés. Le premier cas d'utilisation implique un processus de mappage, basé sur RML, des métadonnées d'archives de la base de données de patrimoine numérique de la BCU Lausanne, Patrinum[5] (MARCXML) aux structures de données ouvertes liées RiC-O et RESCS (RDF) avec le soutien de la Haute école spécialisée des Grisons dans le cadre d'un SWITCH Innovation Lab. Ces processus génèrent une base de qualité nécessaire pour fournir du *linked open data* (LOD) et du *data insight* à Patrinum et des enseignements pertinents pour les autres archives et bibliothèques. Un autre cas d'utilisation vise à récolter et filtrer les métadonnées bibliographiques de la *Swiss Library and Service Platform* - SLSP pour fournir du LOD aux autres fournisseurs de données et de services (par exemple la *Swiss AI Research Overview Platform*).

Les bibliothèques sont les championnes de la collecte, de la conservation, de la gestion durable et de la réutilisation des métadonnées archivistiques, bibliographiques et de recherche. Une infrastructure de recherche ouverte telle que le Connectôme vise à aider les bibliothèques et leurs utilisateurs à enrichir qualitativement les métadonnées existantes (par exemple par des extractions d'entités, des enrichissements avec des données provenant d'autres fournisseurs), à interconnecter les métadonnées bibliographiques avec les données (de recherche) ouvertes et à réutiliser les données ouvertes liées pour élaborer de nouvelles fonctionnalités de recherche, de découverte et de réutilisation pour les utilisateurs finaux des plateformes et des services des bibliothèques.

Jeannette Frey
Directrice de la Bibliothèque cantonale et universitaire de Lausanne
jeannette.frey@bcu.unil.ch

[1] www.switch.ch
[2] www.switch.ch/connectome
[3] Voir par exemple sur le Blue Brain Nexus : www.semantic-web-journal.net/content/blue-brain-nexus-open-secure-scalable-system-knowledge-graph-management-and-data-driven
[4] www.sairop.swiss par SATW
[5] https://patrinum.ch/?ln=de

L'intelligence artificielle au service du **traitement des archives**

Utilisée aux Archives nationales depuis 2015, l'intelligence artificielle permet de traiter les fonds de manière inédite, principalement grâce à la reconnaissance d'écritures manuscrites, au bénéfice des archivistes comme des lecteurs.

ARCHIVES NATIONALES

L'intelligence artificielle, en plus d'être une notion de plus en plus divulguée, est devenue un outil à part entière dans les administrations pour tirer parti des vastes réservoirs de données qu'elles produisent. Le monde de la recherche, de la culture et du patrimoine est lui aussi concerné par ce mouvement de fond : désormais prédomine le sentiment que le stade de l'expérimentation est dépassé pour entrer dans une nouvelle phase des pratiques professionnelles. Aux Archives nationales, c'est à partir de 2015 que l'intelligence artificielle a été employée pour traiter les fonds de manière inédite, principalement en ayant recours à la reconnaissance d'écritures manuscrites, ou *handwritten text recognition* (HTR). Cette technologie s'appuie sur les processus de *machine learning,* qui consistent, à partir de jeux de données, à entraîner une machine à effectuer des actions humaines : ici, en l'occurrence, la transcription de textes manuscrits. Cette technologie répond au rêve longtemps caressé de pouvoir traiter les documents à l'instar de ce qu'était capable d'accomplir l'OCR sur les imprimés. Elle a frayé plusieurs pistes pour le traitement, la diffusion et l'utilisation des archives.

ACCÉDER AU TEXTE DES ARCHIVES : HIMANIS ET ENDP

Lancé fin 2014 par l'IRHT (Institut de recherche et d'histoire des textes) sous l'égide de Dominique Stutzmann, le projet HIMANIS[1] présenta la singularité de choisir comme « terrain de jeu » les registres de la chancellerie royale des XIII[e] et XV[e] siècles conservés dans le Trésor des chartes. Si l'HTR était alors couramment utilisé sur les écritures contemporaines, il n'avait jamais été employé pour transcrire des écritures si anciennes, comportant de nombreuses abréviations. Le corpus à traiter, soit 199 registres pour plus de 83 000 pages, était en outre suffisamment volumineux pour évaluer la maturité de la technologie.

Les premiers résultats ont été obtenus rapidement grâce aux images déjà disponibles au début du projet et aux éditions électroniques d'actes royaux qui fournissaient une « vérité de terrain »[2], c'est-à-dire une transcription d'une partie du corpus assez exacte pour entraîner l'intelligence artificielle à comprendre l'écriture et ses mécanismes d'abréviations. Les données obtenues sont des lemmes alignés sur leur image d'origine et sont interrogeables *via* une interface de recherche[3].

Le chercheur a la possibilité de retrouver directement au cœur du document les termes de son choix et d'effectuer aussi bien des études d'ordre historique que philologique en s'appuyant sur la statistique lexicale. Du point de vue de l'archiviste, ce moteur de recherche pallie l'absence d'inventaire complet de ces registres. On notera néanmoins qu'un tel outil ne peut remplacer un travail d'indexation classique. Face à la matière brute du texte, c'est à l'utilisateur final de retrouver les termes du Moyen Âge qui traduisent un phénomène ou un concept, de connaître les formes lexicales du latin et de l'ancien français pour retrouver toutes les occurrences pertinentes, d'envisager tout ce qu'une indexation matière pouvait lui suggérer.

Dans la continuité d'HIMANIS, il faut aussi évoquer le projet eNDP, débuté en 2020 et qui porte sur les registres des décisions du chapitre de Notre-Dame de Paris. Autre exemple de coopération entre institutions de conservation et structures de recherche, le but de ce projet, qui a actuellement réussi le stade de la transcription par HTR, vise à explorer le contexte social, économique et urbain dans lequel évoluait le chapitre cathédral[4].

ACCOMPAGNER LE TRAITEMENT DES ARCHIVES : LECTAUREP ET SIMARA

Une autre possibilité d'exploiter l'intelligence artificielle consiste à récupérer, à partir des documents d'archives, des informations afin de les réutiliser dans le travail de description documentaire. LectAuRep, projet financé par le ministère de la Culture, mené par le Minutier central des notaires parisiens et l'INRIA, a été conduit de 2018 à 2021[5]. Il porte sur les répertoires de notaires, qui sont les clés d'accès fondamentales à leurs minutes, en donnant la date, l'objet de l'acte et le nom des parties. L'HTR a permis de transcrire ces informations. Elles fournissent un ensemble de métadonnées utiles aussi bien pour décrire les documents que pour se prêter à des explorations statistiques. L'équipe est aussi allée plus loin en ayant recours à la technologie de

[1] L'acronyme signifie : « *HIstorical MANuscripts Indexing for user-controlled Search* ». Carnet de recherche : www.himanis.hypotheses.org

[2] Notamment P. Guérin et L. Celier, *Recueil des documents concernant le Poitou contenus dans les registres de la chancellerie de France,* 14 vol., Poitiers, 1881 ; édition électronique par l'École des chartes : www.corpus.enc.sorbonne.fr/actesroyauxdupoitou

[3] Hébergement par Huma-Num : www.himanis.huma-num.fr/app

[4] Présentation du projet : https://lamop.hypotheses.org

[5] LectAuRep signifie : lecture automatique de répertoires. Carnet de recherche du projet : www.lectaurep.hypotheses.org

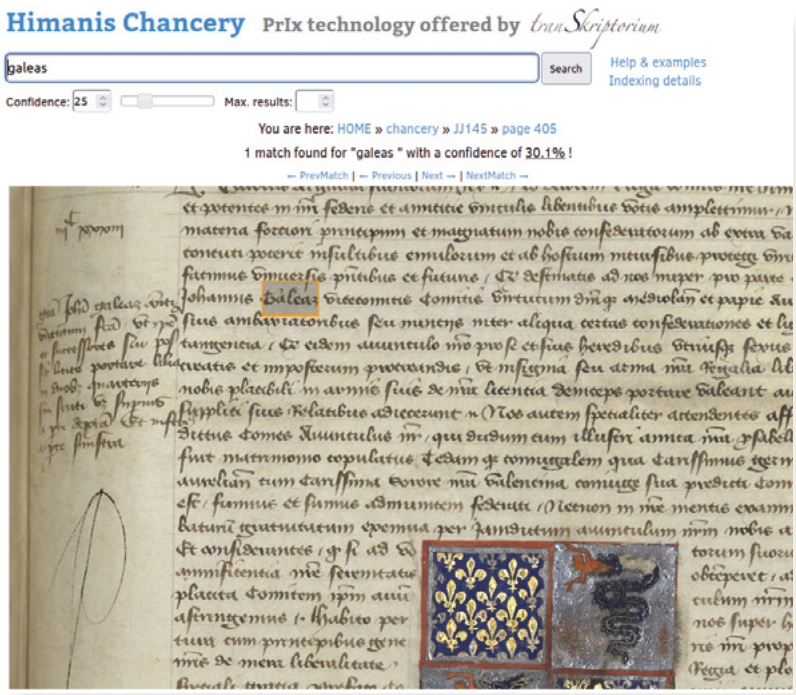
reconnaissance des entités nommées (NER), afin d'accompagner davantage l'archiviste dans le travail de constitution d'index.

Le projet SIMARA[6], soutenu par le plan France Relance, porte quant à lui sur la conversion d'inventaires d'archives anciens en données. Il existe en effet une masse importante d'inventaires manuscrits, allant du XVIIIe au XXe siècle, qui demeurent à ce jour peu accessibles aux lecteurs et qui n'ont fait l'objet d'aucun traitement, hormis une numérisation en mode image. SIMARA vise à traiter environ 800 000 fiches et 100 000 pages. Il s'agit d'une plateforme Web développée par la société Teklia qui l'a adossée à ses infrastructures de traitement HTR. Son but est d'accomplir deux tâches chronophages auparavant réalisées séparément et manuellement par les archivistes : la saisie bureautique de l'inventaire et la structuration des informations en XML EAD. Ces deux opérations sont effectuées conjointement par la plateforme : les outils de segmentation et d'HTR se chargent de transcrire le texte manuscrit, tandis que des outils d'identification des contenus répartissent les informations dans des champs correspondant à des éléments et des attributs XML. L'archiviste peut ainsi se concentrer uniquement sur la relecture des données, en corrigeant les transcriptions, contribuant au passage à améliorer les modèles de reconnaissance d'écritures. Il peut ensuite récupérer le tout au format EAD pour le publier. Ce travail de relecture demeure forcément long et exigeant mais la prise en charge par l'intelligence artificielle des tâches les plus fastidieuses de saisie et d'encodage permettent de réduire significativement le temps de traitement.

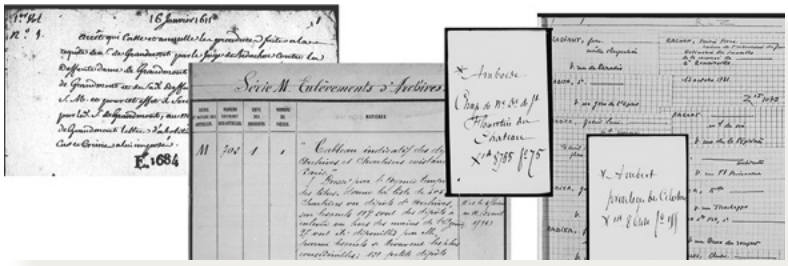
UN PROJET DE PLATEFORME COLLABORATIVE

Face au succès de la solution, une plateforme collaborative pour associer le public à la relecture des transcriptions automatiques sera mise en œuvre (projet GIROPHARES). Les expériences conduites aux Archives nationales ont essentiellement exploité la reconnaissance d'écritures manuscrites, afin d'aller plus loin dans la transformation des fonds d'archives en données. Cette transformation se fait tant au bénéfice des archivistes, pour produire métadonnées et inventaires, que des lecteurs qui réutilisent ces données pour évaluer les phénomènes historiques. De tels projets se multiplient aussi dans le réseau des archives départementales, notamment avec le projet SOCFACE[7] qui, dans la même logique, explore les recensements de population de 1836 à 1936.

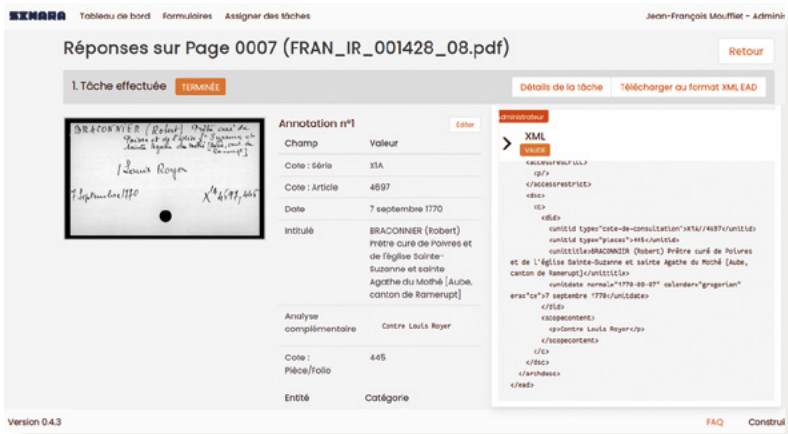
JEAN-FRANÇOIS MOUFFLET
Conservateur en chef du patrimoine, responsable de fonds au département du Moyen Âge et de l'Ancien Régime, Archives nationales
jean-francois.moufflet@culture.gouv.fr



Recherche et reconnaissance d'un même terme au cœur du document.



Les registres des actes notariés sont l'une des sources historiques les plus consultées aux Archives nationales.



Structuration du texte dans un document XML conforme aux spécifications de la *Text Encoding Initiative*.

[6] Acronyme signifiant : Saisie d'Inventaires Manuscrits Assistée par Reconnaissance Automatique. Présentation en ligne pour Etalab : www.speakerdeck.com/etalabia/20220203-datadrink-simara

[7] Présentation du projet : www.socface.site.ined.fr

Acclimatation de l'IA à l'Abes : *la période des semis*

Disposant de données riches et structurées, l'Abes a vocation à accueillir l'intelligence artificielle dans ses outils. Plusieurs initiatives encourageantes sont déjà en cours.

L'Abes et ses réseaux produisent ou agrègent beaucoup de données, souvent riches et bien structurées, qu'il s'agit d'exploiter pour offrir des services. Depuis quelques années déjà, il ne fait plus de doute que les techniques de l'intelligence artificielle relèvent de l'état de l'art, et plus seulement de la recherche scientifique. À ce titre, elles doivent rejoindre la boîte à outils de l'Abes, en répondant à de grands axes de réflexion : les chantiers IA ont vocation à s'aligner sur les missions permanentes de l'Abes et les priorités du projet d'établissement, en tenant compte des moyens et compétences dont l'agence dispose et en s'appuyant sur les besoins et les expertises des établissements de ses réseaux, de ses partenaires et de ses homologues à l'international.

L'IA ne remplacera pas le catalogage ni d'autres formes de curation de données. Au contraire, il faut des données de qualité, validées par des professionnels, pour faire apprendre les machines. L'IA peut aider là où le travail humain serait trop fastidieux ou incapable de digérer la masse des documents à décrire et analyser.

IDENTIFIER DES OBJECTIFS PRIORITAIRES ET FAIRE ÉMERGER UNE COMPÉTENCE INTERNE

À partir de ces principes directeurs, de nombreux défis se posent :

Faire émerger une compétence interne, sans recourir exclusivement à l'externalisation. Il faut à la fois essayer de rééquilibrer les spécialisations informatiques actuelles (conception et développement d'applications, gestion d'infrastructures) vers l'ingénierie des données, et compter sur la remarquable capacité des bibliothécaires et des informaticiens de l'Abes à embrasser de nouvelles technologies. Les études du Labo visent à construire avec eux des mises en pratique orientées par des besoins précis, pour que demain on ne parle plus d'IA mais d'outils précis, spécialisés et banalisés.

Identifier les objectifs prioritaires et accessibles

L'étiquette « IA » recouvre différents types de tâches génériques qui ont toutes une application potentielle dans notre contexte : classification automatique (Rameau, Dewey, codes de fonction, types de document, langue), reconnaissance d'entités (repérer une personne dans une mention de responsabilité, un organisme dans une affiliation), liage automatique (d'un nom à IdRef ou Orcid, d'une fonction à un code, d'une manifestation à une œuvre) ; mesure de similarité entre entités (moteur de recommandation, dédoublonnage) ; détection de clusters (repérer des groupes de bibliothèques ou de documents dans la masse des localisations, par exemple). Les possibilités sont immenses, les difficultés inégales.

DEUX ÉTUDES QUI UTILISENT LE *MACHINE LEARNING* POUR AMÉLIORER LES NOTICES BIBLIOGRAPHIQUES SUDOC

Depuis 2021, c'est principalement dans le cadre du Labo, sous forme d'études, que l'Abes fait ses premiers pas.

En 2021 et 2022, le labo de l'Abes a accueilli pendant 6 mois deux étudiants du master IASD (Intelligence artificielle, Systèmes, Données) de l'université de Montpellier, sous la direction du Professeur Pascal Poncelet. Deux études, sur deux thématiques distinctes, ont pu être menées.

La première, effectuée par Min Young Yang, portait sur l'indexation Rameau automatique. Avec l'identification des auteurs, l'indexation automatique est un défi majeur pour la gestion des bases bibliographiques, confrontée à l'inflation de la production scientifique. Dans le cadre de ce stage, il s'agissait de créer un premier prototype qui prédit des concepts Rameau pertinents à partir du titre et du résumé d'un document. Défi redoutable étant donné le caractère subjectif de cette opération intellectuelle et la taille du vocabulaire (100 000 termes). En 2023, l'Abes poursuivra ce travail à travers une collaboration avec une société spécialiste en IA. À terme, si la preuve de concept est concluante, il s'agira d'appliquer l'algorithme aux données de l'Abes mais aussi de le partager sous la forme d'un web service et de code ouvert (éventuellement intégrable dans un outil générique comme Annif[1]).

La seconde étude, menée par Thomas Zaragoza, consistait dans le repérage des auteurs et de leur fonction dans les mentions de responsabilité des notices bibliographiques du Sudoc. Dans plusieurs centaines de milliers de notices du Sudoc, la transcription des mentions de responsabilité n'est pas en accord avec les entrées Auteur : certains auteurs ne sont pas listés séparément dans les zones dédiées, ou bien sans précision de leur rôle. Thomas Zaragoza

[1] https://annif.org

a travaillé à identifier automatiquement dans le texte des mentions de responsabilité les chaînes de caractère qui correspondent à une personne (PER) ou à une fonction (FONCT) :

Hawking / Stephen Finnigan, réalisation ; Stephen Hawking, Stephen Finnigan, Ben Bowie, scénario ; Joe Lovell ; Tina Lovell ; Arthur Pelling [et al.] acteurs

Cette opération de reconnaissance d'entités (NER) a nécessité l'annotation manuelle de milliers de notices, à travers une interface dédiée, pour fournir à l'algorithme d'apprentissage des données de qualité en entrée. Il fallait ensuite aligner les mots évoquant une fonction avec le bon code, ce qui est souvent bien plus difficile que dans l'exemple ci-dessus. L'équipe labo entend approfondir le travail sur ce dernier point.

Yann Nicolas
Responsable du Labo de l'Abes
nicolas@abes.fr

Crédit Adobe stock

QUALINKA : DE L'IA « À L'ANCIENNE » POUR AUTOMATISER LE LIAGE AUX AUTORITÉS

Dès 2012, en participant au projet ANR Qualinka, l'Abes a voulu trouver des solutions techniques pour automatiser le travail de liage et de diagnostic qualité des liens entre notices bibliographiques et notices d'autorité. Les efforts opérationnels ont porté jusqu'à présent sur les liens entre les personnes référencées dans les notices bibliographiques du Sudoc et les notices d'autorité de type personnes physiques du référentiel IdRef. Depuis 2019, le programme Qualinka issu de ces travaux est disponible pour les catalogueurs du Sudoc dans l'application paprika.idref.fr, où il offre une aide à la décision précieuse pour corriger et créer de nouveaux liens.

Un programme d'IA symbolique

Ce courant de l'IA, plus ancien que les méthodes de *machine learning*, repose sur la modélisation des connaissances et des raisonnements humains pour expliciter un ensemble de règles au sein de programmes informatiques capables, au moyen d'approches logiques, de les exécuter pour prendre des décisions. Pour concevoir Qualinka, il a fallu reproduire les différentes étapes qui permettent à un humain de créer des liens bibliographiques. Par exemple, un catalogueur sait d'expérience qu'une personne a tendance à coécrire avec les mêmes personnes ; s'il doit, à partir d'un document, identifier cette personne à une notice d'autorité, il choisira celle dont les documents liés ont les mêmes coauteurs que le document de départ. Évidemment, il n'y a pas toujours de coauteurs et le catalogueur doit, en fait, souvent croiser différentes informations et opérer des pondérations, en particulier pour les cas d'homonymie (plusieurs notices d'autorité avec les mêmes noms et prénoms).

Qualinka prend en entrée un sous-ensemble de références de personnes issues de notices documentaires et de notices d'autorité. Ce sous-ensemble a été préalablement construit à partir d'une recherche sur le nom et le prénom. À chaque référence sont associés des attributs tels que le titre du document, les cocontributeurs, les sujets, la date de publication, les dates de vie, etc. Ces attributs sont comparés au travers de critères prédéfinis, eux-mêmes combinés ensemble dans des règles logiques permettant à Qualinka de décider quelles références correspondent à la même personne.

SudoQual, cadre de développement de nouveaux scénarios d'application

Il n'est pas possible qu'un seul programme puisse traiter tous les problèmes de liage entre divers types d'entités, car les connaissances et les raisonnements impliqués sont différents. En revanche, un cadre technique et méthodologique a été créé et utilisé dans un premier temps pour mettre au point Qualinka. Ce cadre, baptisé SudoQual, permet de réaliser les différentes étapes de configuration d'autres programmes : modélisation, formalisation en règles logiques, élaboration d'algorithmes de décisions, sans oublier l'évaluation du comportement produit et l'ajustement des phases précédentes. SudoQual est d'ores et déjà disponible en *open source*[2]. Cela doit permettre, à l'Abes ou à d'autres, de créer des scénarios d'applications adaptés à différents cas d'usage orientés par un type d'entité, un contexte documentaire, un référentiel cible ou encore une modalité particulière de liage.

Aline Le Provost
Analyste de données bibliographiques, l'Abes
le-provost@abes.fr

[2] https://github.com/abes-esr/sudoqual-framework

L'IA et la fouille de textes à l'INIST : *l'IA à portée de tous ?*

L'Inist a pour objectif de développer des outils de fouille de textes intégrables dans des chaînes de production mais également utilisables par tous les acteurs travaillant avec des publications numériques.

Si la fouille de textes a toujours été présente à l'Inist[1], ce n'est qu'avec le lancement du projet Istex[2] que des méthodes d'intelligence artificielle vont être développées pour être appliquées en grandeur nature sur de gros volumes de données, dans un processus industrialisé.

Initié en 2012 par le MESR dans le cadre du Programme d'investissements d'avenir[3], Istex comprend dès le début un volet « enrichissement » *via* des méthodes de fouille de textes. Alors que l'IA n'était pas encore un mot-clé passé dans le langage commun, l'équipe Istex-RD d'alors commence à développer des méthodes d'enrichissement de données à partir notamment de techniques d'apprentissage automatique (*machine learning* ou ML), afin de déterminer automatiquement le domaine scientifique des articles. D'autres méthodes, comme l'extraction de mots clés ou d'entités nommées vont également voir le jour sous forme de modules intégrés à la chaîne de production, permettant le traitement des corpus au fur et à mesure de leur intégration[4].

À travers quatre axes de travail (structuration des documents, indexation automatique, reconnaissance d'entités nommées, catégorisation des documents) nous avons alors répondu aux trois principaux défis rencontrés :

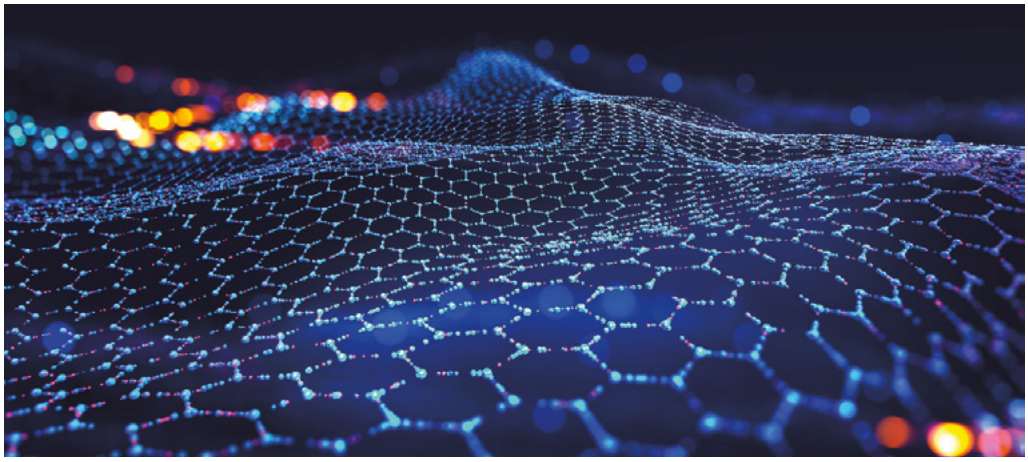
- Mise au point et intégration d'outils de TDM : entraînement, adaptation, mise en production.
- Passage à l'échelle : 25 millions de documents à traiter.
- Reversement des données : modélisation des données, réintégration, mise à disposition.

Si cette approche a donné de bons résultats, elle a montré un certain nombre de limites : développer et mettre en place un nouveau traitement est un processus complexe à mettre en œuvre, et surtout cela rend très difficile l'utilisation de ces programmes en dehors de la chaîne Istex.

D'UNE IA INTÉGRÉE DANS UN PROCESSUS DÉFINI À UNE IA INDÉPENDANTE DES DONNÉES

Dans une démarche « science ouverte », on constate la mise en ligne croissante, *via* GitHub ou GitBucket, de programmes permettant de traiter des données. Mais si tout un chacun a accès à ces algorithmes, leur mise en œuvre souvent complexe n'incite pas les non informaticiens à les utiliser. Nous nous inscrivons bien dans ce mouvement, en publiant tous nos codes, cependant nous voulons aller plus loin en faisant en sorte que qui que ce soit puisse les utiliser, quelles que soient ses compétences. Cela doit répondre aux demandes d'utilisateurs, documentalistes ou chercheurs, qui souhaitent pouvoir utiliser ces programmes sur leurs propres données de façon simple, et en pouvant choisir eux-mêmes les traitements dont ils ont besoin.

Nous avons fait le choix de créer et déployer des applications d'IA sous forme de web services (WS), intégrables dans une chaîne de production comme Istex, mais également directement utilisables par tout utilisateur désirant traiter ses propres corpus. Ainsi nous passons d'une IA intégrée dans un processus défini à une IA indépendante des données,



Crédit Adobe stock

[1] www.inist.fr

[2] www.istex.fr

[3] https://franceuniversites.fr/wp-content/uploads/2012/04/Projet_Istex.pdf

[4] Cuxac P., Thouvenin N. (2017) : Archives numériques et fouille de textes : le projet Istex. *Atelier TextMine, conférence EGC,* 24 janvier 2017, Grenoble, France. https://textmine.sciencesconf.org/data/pages/TextMine17.pdf

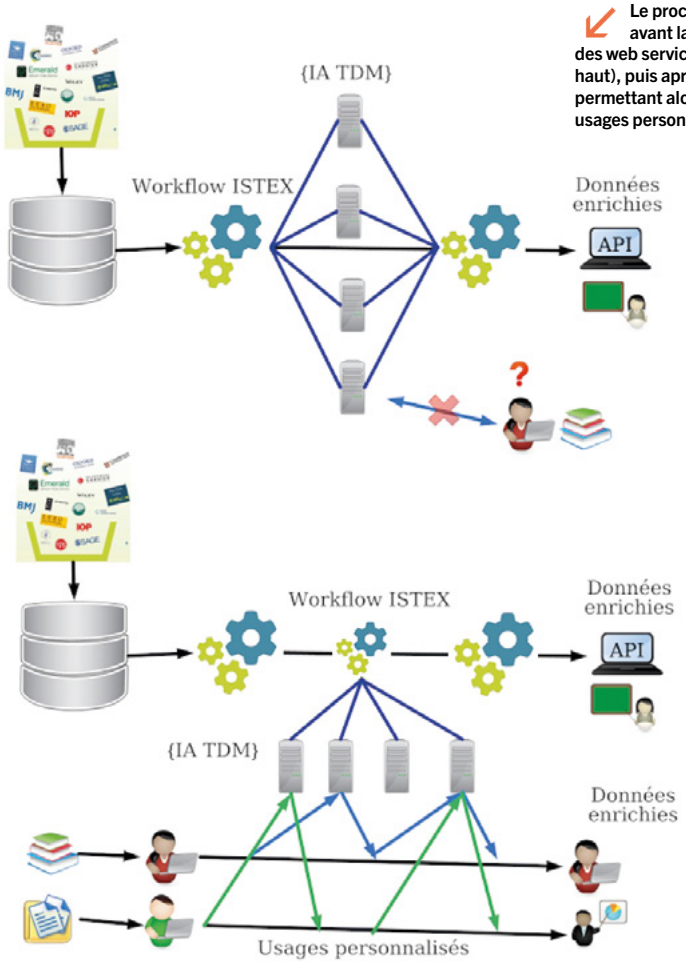
avec des contraintes minimales, utilisable par des non spécialistes, et largement extensible pour répondre à de nouveaux besoins.

Les méthodes implémentées peuvent être complexes, mettant en œuvre des réseaux neuronaux élaborés, avec un nombre élevé de paramètres à optimiser. Afin de faciliter au maximum leur usage, les web services doivent répondre à un certain nombre d'exigences :

- Chaque service ne doit répondre qu'à un seul besoin.
- Il n'y a pas de paramétrage par l'utilisateur.
- Il doit y avoir un seul format d'entrée/sortie très simple (ex : un JSON identifiant/valeur).
- Ils doivent être utilisables *via* une interface graphique (dans l'outil de visualisation Lodex[5])

Les modèles de ML sont construits par des spécialistes TDM, avec l'aide d'experts pour la constitution des corpus d'apprentissage et la validation des algorithmes, puis utilisés par les WS mis à disposition : l'IA devient alors accessible à tous et facilement applicable aux données bibliographiques, que ce soit sous forme de métadonnées ou de texte intégral[6]. Pour aider l'utilisateur, le site Internet objectif-TDM[7] recense les services en production : il permet à la fois d'identifier le service correspondant à ses besoins, de connaître son URL et d'avoir une aide sur son utilisation. À partir de là, le service est utilisable suivant les compétences de chacun : *via* une interface graphique dans Lodex (outil *open source* de visualisation de données structurées[8]), en ligne de commande, ou intégré dans un programme informatique.

Les nouveaux services proposés permettent l'utilisation de méthodes apportant une forte valeur ajoutée aux données traitées sans qu'il soit nécessaire de mobiliser des compétences en informatique, ou en *data mining*. Le système est suffisamment souple pour rapidement mettre en production de nouveaux services répondants à de nouveaux besoins ou à des données ayant des caractéristiques propres. En effet, les nouvelles approches font appel à des modèles de langage, modèles statistiques capables par exemple, à partir de la distribution de séquences de mots, de deviner une succession de mots. Avec le développement des approches neuronales (et le *deep learning*), on trouve d'importantes ressources pré-entrainées sur de gros volumes de données et dans des langues différentes. Ces gros modèles « génériques » peuvent avoir un intérêt mais peuvent aussi entraîner des biais importants[9] ; nous l'avons constaté, par exemple, dans le cas d'analyses de publications de la fin XIXe siècle au début XXe en français, où le style est fondamentalement différent de celui de nos jours. C'est le cas également dans le cas de corpus scientifiques spécialisés où le vocabulaire scientifique est très mal appréhendé *via* des modèles génériques. Pour faire face à ces cas-là, nous devons pouvoir nous adapter en créant nos propres modèles, adaptés aux données à traiter.



Le processus IA avant la création des web services (en haut), puis après (en bas), permettant alors des usages personnalisés

À l'Inist-CNRS, nous avons mis en place un environnement approprié facilitant le déploiement de ces services à partir d'algorithmes d'IA (plus spécifiquement de TDM). Cela permet une grande souplesse quant à la modification, l'adaptation ou la création de web services.

Cette nouvelle offre de service est donc là pour répondre à de multiples finalités et s'adresse à tous les professionnels de l'IST qui ont besoin, par exemple, de détecter des thématiques scientifiques, de classer des documents, ou encore de les enrichir pour faire de la bibliométrie. Elle propose des services assez génériques pour être utiles au plus grand nombre, mais est également capable de s'adapter aux besoins exprimés, et ainsi d'évoluer continuellement pour répondre à de nouveaux usages.

Pascal Cuxac
Responsable du Service Text & Data Mining, Inist-CNRS
pascal.cuxac@cnrs.fr

[5] www.inist.fr/projets/lodex

[6] Bonvallot V., Parmentier F., Bourguignon L., Clauss I. and Gregorio S. (2022) : Le TDM pour tous grâce à des web services au sein de LODEX, outil libre de visualisation, *Revue des Nouvelles Technologies de l'Information, Extraction et Gestion des Connaissances,* RNTI-E-38, 2022, 445-452 https://editions-rnti.fr/render_pdf.php?p=1002758

[7] https://objectif-tdm.inist.fr

[8] Gregorio, S., A. Collignon, F. Parmentier, and Thouvenin N. (2019) : LODEX : des données structurées au web sémantique https://hal.archives-ouvertes.fr/hal-01990444. *Atelier Web des Données Conférence EGC,* 2019, Metz, France.

[9] Bender E.M., Gebru T., McMillan-Major A., and Shmitchell S. (2021) : On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the *2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 610–623.

ISTEX : de la plateforme de référence à l'infrastructure de recherche

En mettant à disposition près de 26 millions de documents, Istex est aujourd'hui le plus vaste réservoir d'archives scientifiques au service de la recherche française, proposant un usage documentaire pour la consultation de documents, et un usage plus avancé de fouille de textes pour l'exploitation et le traitement de lots de documents.

Né d'une volonté nationale, le projet Istex (Initiative d'excellence de l'information scientifique et technique) s'inscrivait dans le programme « Investissements d'avenir », initié alors par le ministère de l'Enseignement supérieur et de la Recherche. L'idée était d'acquérir massivement des collections rétrospectives de la littérature scientifique dans toutes les disciplines et de se doter d'un outil innovant d'exploitation des données. En s'inspirant du modèle de la Fondation allemande pour la recherche (DFG) qui avait amorcé une démarche d'indépendance envers les éditeurs, ce projet visait un accès pérenne aux publications via une plateforme hébergée sur le territoire national, afin de gagner une certaine autonomie vis-à-vis des éditeurs scientifiques, souverains jusque-là en matière d'accès aux publications.

Quatre acteurs principaux, reliés par un accord de consortium, ont mis en œuvre, chacun avec un rôle spécifique, ce projet doté à sa création le 19 avril 2012 d'un budget de 60 millions d'euros. Le CNRS était porteur du projet, l'Inist avait pour mission de développer l'infrastructure matérielle et logicielle, le Consortium universitaire de publications numériques (Couperin) avait comme mission principale le recueil des besoins et les négociations avec les éditeurs, tandis que l'Agence bibliographique de l'enseignement supérieur (Abes) prenait en charge les acquisitions et le signalement des collections dans les outils documentaires nationaux. Quant à la Conférence des présidents d'université (aujourd'hui France Universités), représentée par l'université de Lorraine, elle avait pour rôle de faire le lien avec les communautés de recherche, en pilotant notamment les projets de services à valeur ajoutée et les chantiers d'usage.

UN PROJET EN DEUX ÉTAPES

La première étape a consisté en une politique volontariste et massive d'achats centralisés d'archives scientifiques sous forme de licences nationales. Celles-ci ont été déterminées en fonction des besoins recensés dans les différentes communautés notamment *via* une enquête de grande ampleur à laquelle quelque 7 000 professionnels de la recherche ont répondu. Un comité de pilotage représentatif de l'ensemble des communautés a ensuite validé les choix et hiérarchisé les priorités d'acquisitions en veillant aux équilibres disciplinaires. S'appuyant sur l'expérience de consortia étrangers, l'Abes et le consortium Couperin ont mené les négociations avec les éditeurs, dans le cadre de contrats d'acquisition innovants.

Portée par l'Inist, la seconde étape du projet était la création de la plateforme destinée à héberger l'ensemble des données, construite en méthode Agile en lien avec les partenaires et utilisateurs.

Un autre choix a été fait, et pas des moindres : celui de ne pas créer d'interface mais plutôt de s'intégrer dans les systèmes existants. Cela a commencé par des widgets, intégrés dans les portails documentaires des établissements, avant de devenir un bouton Istex visible sur les plateformes utilisées par les chercheurs. Depuis mars 2021, les ressources Istex sont accessibles via l'extension unifiée *Click & Read* installable sur les principaux navigateurs Internet.

BONIFIER LES DONNÉES POUR LA FOUILLE DE DONNÉES

Les données reçues n'étant pas toujours de qualité optimale pour l'exploitation par les utilisateurs finaux, un des plus gros défis a été de les nettoyer pour les homogénéiser et les rendre ainsi aptes à être « ingérées ». Pour cela, des feuilles de style ont été créées afin de structurer les données.

Un *workflow* a été mis en place en étroite collaboration avec l'Abes pour les échanges avec les éditeurs et la restructuration des métadonnées mises à disposition par ceux-ci.

Le premier chargement de données s'est déroulé en 2014 avec 6 millions de documents. Le processus s'est ensuite généralisé : enquête, négociation, livraison et chargement pour proposer aujourd'hui plus de 25,5 millions de documents provenant de 32 sources différentes.

En parallèle, des étapes d'enrichissement des données se sont mises en place pour ajouter de nouvelles métadonnées telles que des entités nommées, des références bibliographiques structurées, une indexation ou encore une catégorisation par domaine scientifique.

Grâce aux services Istex qui ont été développés, il est possible d'explorer, d'analyser des données et de faire de la fouille de textes. L'API Istex permet de faire de la recherche documentaire (facilitée grâce à la revue de sommaires : **https://revue-sommaire.istex.fr**), les résultats étant téléchargeables de façon massive avec Istex-DL. Lodex intervient ensuite pour l'exploration et la visualisation des corpus. Sans oublier Data.Istex qui regroupe des exemples de corpus prêts à l'emploi.

Récemment, Istex s'est affiché parmi les 108 infrastructures retenues dans la feuille de route nationale des Infrastructures de recherche 2021, dans la catégorie projet, éditée par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Les principaux objectifs stratégiques exposés pour ce projet d'infrastructure sont :

- Ouvrir la collection aux ressources nativement publiées en accès ouvert et poursuivre son alimentation grâce à une politique d'acquisition ambitieuse
- Faciliter la constitution de corpus cohérents et enrichis, directement exploitables pour du TDM
- Promouvoir le développement de services avancés avec la communauté des chercheurs en TAL
- Offrir des services d'exploration et d'exploitation de corpus accessibles à tous.

Outre le caractère novateur de la réalisation technique, Istex a ouvert la voie à de nouvelles collaborations entre des acteurs de l'IST mutualisant leurs efforts et compétences au service de la communauté ESR. Il est aussi une ressource pour de la fouille de données grâce à la mise à disposition de textes intégraux documentés par des métadonnées riches et téléchargeables massivement.

Alexandra Petitjean-Monnin
Chargée de communication, Inist-CNRS
alexandra.petitjean@inist.fr

Raluca Pierrot
Responsable du service Documentation électronique, Abes
pierrot@abes.fr

Cécilia Fabry
Responsable communication, Inist-CNRS
cecilia.fabry@inist.fr

En juin 2022 s'est créé un chapitre francophone dans ai4LAM[1], une communauté participative et internationale autour de IA dans les bibliothèques, les musées et les archives. Explication avec les trois membres du secrétariat de ce chapitre[2].

« NOTRE OBJECTIF EST DE RASSEMBLER LES PROFESSIONNELS FRANCOPHONES AUTOUR DE L'IA »

Arabesques : Qu'est-ce que la communauté ai4LAM ?
C'est une communauté participative et internationale, à la structure assez informelle, qui vise à partager et faire connaître les usages des technologies liées à l'intelligence artificielle appliquée aux domaines des bibliothèques, archives et musées. La communauté s'est montée en 2018 autour de l'université de Stanford, de la Bibliothèque nationale de France et de la Bibliothèque nationale de Norvège. Cette dernière a accueilli la première conférence ai4LAM en 2018, Stanford la seconde en 2019, la BnF et l'université Paris-Saclay la dernière conférence en décembre 2021[3].

En juin dernier est né le projet de créer un « chapitre » francophone au sein de cette structure. En quoi consiste ce projet ?
Lors de la conférence *Les futurs fantastiques* de 2021, une réunion a été organisée pour réfléchir collectivement à l'opportunité de créer un chapitre francophone, aux enjeux d'une telle création et aux grandes lignes d'un programme d'action. Partage d'expériences, traduction des contenus existants, formations ont été les principaux axes de travail suggérés pour le chapitre. Depuis, nous avons élaboré avec quelques autres personnes[4] une charte pour ce chapitre francophone[5], qui précise notamment le programme d'action du chapitre pour un an et ses modalités générales d'organisation. Cette charte a été approuvée par une cinquantaine de participants lors de la deuxième réunion en juin 2022. Le chapitre s'est donné pour objectifs de décloisonner les communautés, de lever les obstacles linguistiques, de faciliter le partage d'expériences et la mutualisation. Il travaillera dans le respect des principes FAIR[6] et CARE.

Comment fonctionnera ce chapitre et quelle sera son articulation avec la structure mère ?
Ai4LAM est une structure assez informelle, sans statut juridique. Pour officialiser la création d'un chapitre, le bureau de ai4LAM demande juste une charte et la liste des premiers membres du chapitre. Nous lui avons envoyé ces documents. Si la demande est validée, un espace sur le site Web ai4LAM et des outils de travail seront ouverts au nouveau chapitre.
Le chapitre est animé par un secrétariat, qui sera assuré par nous-mêmes pendant ses premiers mois d'existence. À compter de la deuxième année, les membres du secrétariat seront élus par les membres du chapitre. Le chapitre a décidé de s'intéresser dans un premier temps, dans le cadre de groupes de travail informels, au recensement des projets HTR (reconnaissance d'écriture manuscrite) et à la question des formations (état des lieux et besoins).

Existe-t-il déjà une communauté de professionnels français autour de ces questions ?
Plusieurs établissements ont développé des expertises, lancé des projets dans certains domaines de l'IA. Les projets de *machine learning* appliqués à la reconnaissance des écritures manuscrites représentent l'essentiel des travaux, en particulier en archives et en bibliothèques. Il existe aussi, à notre connaissance, des projets relatifs à la reconnaissance d'entités nommées, à la classification ou l'indexation automatique de textes et d'images, à la création d'agents conversationnels. Mais il n'y a pas pour l'instant d'espace qui transcende les barrières sectorielles et géographiques et permette de partager véritablement expériences et états de l'art, voire de mutualiser les ressources (vérités terrain, modèles d'apprentissage etc.) et idéalement les infrastructures qui accueillent les outils.

Y a-t-il une spécificité française ou francophone ?
États de l'art, modèles et outils dépendent souvent de la langue et du système d'écriture des corpus ainsi que du contexte culturel dans lequel les corpus s'inscrivent. Aussi, structurer des communautés linguistiques qui manipulent des corpus dans un contexte francophone est un enjeu fort d'efficacité de ces travaux. L'omniprésence de l'anglais dans la littérature et les outils relevant de l'IA peut être dissuasive pour certains collègues. Nous souhaitons produire ensemble, en français, et rendre accessibles une documentation plus large et des outils d'apprentissage adaptés, ce qui donnera aussi de la visibilité aux projets et à la production francophone.

Quels établissements ou professionnels sont concernés ?
Tous les acteurs qui souhaitent travailler dans un cadre francophone peuvent devenir membres du chapitre sur simple message à l'adresse ***ai4lamfre@gmail.com.*** Il leur sera juste demandé de prendre connaissance de la charte et de confirmer leur adhésion au chapitre et à ses principes. Selon les règles d'ai4LAM[7], à l'inverse des institutions publiques qui peuvent être membres, les acteurs privés ne peuvent participer qu'à titre individuel. Sont potentiellement concernés les acteurs venant de la recherche, publique ou industrielle, les sociétés de services spécialisées, les établissements conservant des biens culturels et toute personne intéressée par ces sujets.

Quel est le calendrier des actions et les prochaines étapes ?
Les prochaines étapes sont la mise en place des espaces de travail et d'une liste, l'organisation de webinaires en 2022, la constitution des groupes de travail, des réunions du chapitre et un événement plus important en 2023.

[1] https://sites.google.com/view/ai4lam

[2] Luc Bellier, directeur adjoint des bibliothèques de l'université Paris-Saclay ; Florence Clavaud, responsable du Lab des Archives nationales ; Antoine Courtin, chargé de développement des ressources numériques du Centre de ressources et de recherches de l'Établissement public du musée d'Orsay et du musée de l'Orangerie Valéry Giscard d'Estaing.

[3] www.bnf.fr/fr/les-futurs-fantastiques

[4] Parmi lesquelles Céline Leclaire (BnF) et Pauline Charbonnier (Archives nationales).

[5] https://docs.google.com/document/d/1mWP2qf0Y0xGFiTxz0gdcdEJjItIkvyrR1A6hMMO4e1Q/edit?usp=sharing

[6] Acronyme de *Findable, Accessible, Interoperable, Reusable.* Voir www.go-fair.org/fair-principles

[7] https://drive.google.com/file/d/1oSfiBwHEax9N-JpSk9AJ4hj1BUEzGWzh/view

Loin d'être le privilège des plus grands établissements, l'intelligence artificielle peut être mobilisée dans toute bibliothèque universitaire. Illustration pratique avec l'expérience du SCD de l'université Côte d'Azur.

Pour une approche décomplexée de l'IA

UNIVERSITÉ CÔTE D'AZUR

Si l'on devait résumer les quelques sujets qui nous ont collectivement animés ces dernières années, l'intelligence artificielle (IA) en est curieusement absente, ou du moins marginalement traitée : transition bibliographique, Web de données, *open access*, ou encore services aux chercheurs ont la part belle, mais point ou peu de *machine learning* et autres réseaux de neurones. En même temps que les applications et services à base d'IA se déploient couramment dans notre vie quotidienne, alors même que ces « nouvelles » technologies de traitement de la donnée commencent concrètement à infuser dans les projets de recherche scientifique (notamment en humanités numériques), force néanmoins est de constater que l'IA tarde à se manifester dans nos quotidiens de bibliothécaires. Sans doute en raison du fait que, pour paraphraser Philippe Le Pape à propos de la transition bibliographique, il n'y aura pas de grand soir de l'IA en bibliothèque mais une succession de petits matins au gré des expérimentations et retombées des projets menés, notamment, à l'Abes et à la Bibliothèque nationale de France. Cela dit, même si ces opérateurs paraissent les mieux placés en terme de compétences techniques et de maîtrise des données pour aborder ces problématiques d'IA, sur le papier rien ne s'oppose à ce que tout un chacun dans son établissement se forme et s'essaie à tester, expérimenter, prototyper, voire utiliser en production des algorithmes d'apprentissage machine, ceux-ci étant largement documentés et leurs implémentations disponibles dans des librairies *open source*.

L'IA OUTIL DE PRÉDICTION

Fondamentalement, les dispositifs d'IA s'appuient sur des logiques inhabituelles pour des bibliothécaires férus de modélisation de haut niveau, dont découle la norme qui encadre la production de données. Inversement, la spécificité des techniques d'apprentissage machine est d'être totalement empirique et de partir des données afin d'en induire les motifs implicites et corrélations sous-jacentes. Sous réserve de disposer d'assez de données et de leur appliquer les traitements adéquats à l'aide de techniques statistiques d'échantillonnage et d'outils de *data mining*, l'apprentissage automatique ou profond consiste à découvrir des éléments de structuration invisibles et à modéliser les *patterns* implicites contenus dans ces observations d'apprentissage afin d'en extraire de l'information et de faire des prédictions. À la base du *machine learning*, rien de magique donc, mais l'application balisée et rigoureuse de techniques statistiques et géométriques connues depuis longtemps pour effectuer de la régression, classification ou « clusterisation » sur des jeux de données simples. Pour la réalisation de tâches plus complexes et non linéaires, telles que la reconnaissance vocale ou la reconnaissance de formes dans des corpus d'images, les algorithmes de *deep learning* s'appuient, eux, sur les technologies à base de réseaux de neurones pour modéliser des transferts auto-apprenants d'informations de plus en plus abstraites entre couches de neurones, à l'image de ce qu'il se produit dans les cerveaux des mammifères.

EXPLORER LES COLLECTIONS D'UN POINT DE VUE NOUVEAU

Afin d'illustrer des possibilités de prises en main dans le cadre de problématiques simples d'établissement, voici trois exemples, dont deux ont été abordés dans une démarche de preuve de concept sans objectif opérationnel, tandis que le dernier cas s'intègre, lui, dans un circuit de traitement ayant donné lieu à la mise en production concrète d'une application (le baromètre Science ouverte de l'université Côte d'Azur développé par le SCD). La première expérimentation consiste à prototyper un système de recommandations de documents intégré avec l'Opac[1] afin d'évaluer dans quelle mesure l'enrichissement des résultats fournis par le moteur de recherche du catalogue grâce à des notices « proches » complémentaires favorise la découvrabilité de contenus au cours de l'expérience usager. Pour ce faire, le cœur du processus consiste à travailler sur un jeu de données bibliographiques et un jeu de données d'usage des documents afin d'opérer des calculs de :

- Similarité entre contenus pour mesurer la proximité entre chaînes de caractères de chaque paire de notices à partir d'une analyse type NLP[2] sur les champs jugés les plus signifiants.

[1] https://bibliotheques.wordpress.com/2019/11/04/ia-et-opac-mettre-en-place-un-moteur-de-recommandation-dans-un-opac-1-4-quelques-bases-conceptuelles

[2] Traitement automatique du langage naturel

• Similarité par l'usage qui permet de déduire des associations de documents à partir des prêts effectués par les lecteurs (les lecteurs qui ont emprunté ceci ont aussi emprunté cela).
On est ici dans le cas d'un apprentissage basé sur des observations à partir duquel l'algorithme est ensuite généralisable à de nouveaux exemples en calculant des proximités à partir des données apprises. Ainsi ce prototype[3] d'application Web, qui requête à la fois l'API de l'Opac et le modèle obtenu, permet d'afficher, pour chaque notice de la liste de résultat, les notices similaires calculées.

Le deuxième exemple proposé part d'une idée toute simple et illustre explicitement l'objectif fondamental de l'IA – faire des prédictions à partir de jeux de données –, cette fois-ci dans le cadre d'un apprentissage par modèle : dans le stock de documents détenus par une bibliothèque, certains ont été empruntés au moins une fois et d'autres jamais. Peut-on dégager des variables pertinentes qui déterminent le fait pour un livre d'être emprunté ou pas, et ainsi, prévoir pour des nouveaux documents leur probabilité d'être empruntés en fonction de leurs métadonnées ?
Réponse : oui à 75 %. Plus précisément, après avoir entraîné plusieurs algorithmes de classification binaire[4] sur un fichier csv contenant les métadonnées de tous les exemplaires de monographies créés entre 2015 et 2021 au SCD, on obtient pour le plus performant d'entre eux un taux de précision de 75 % (l'algorithme a raison 3 fois sur 4 lorsqu'il prédit que les exemplaires du jeu de test seront empruntés) et arrive à détecter 75 % des exemplaires réellement prêtés (mesure du rappel)[5]. Au-delà de la résolution algorithmique du problème posé, l'intérêt d'une telle démarche est aussi d'être amené, durant l'étape de préparation des données[6], à explorer les collections d'un nouveau point de vue en déployant une analyse centrée sur l'usage qui va potentiellement au-delà des analyses traditionnelles de politique documentaire, à l'image de ce prototype de tableau de bord[7] qui reprend la plupart des visualisations créées pour la génération du modèle.

L'IA AU SERVICE DU BAROMÈTRE *OPEN ACCESS* DE L'UNIVERSITÉ

Enfin, le troisième cas d'usage concerne le set de métadonnées de publications collectées puis enrichies avec les données d'Unpaywall qui sert de base au baromètre *open access* de l'établissement. En plus des indicateurs d'ouverture par date de publication, date d'observation et structures de recherche, nous souhaitions également produire des vues et données chiffrées par disciplines, à l'instar du baromètre national. Afin d'indexer le millier de publications UCA qui n'étaient pas

Crédit Adobe stock

présentes dans le jeu de données du ministère, et pour lesquelles nous ne disposions donc pas de variable discipline, la solution la plus évidente à mettre en œuvre a consisté à entraîner plusieurs algorithmes de multi-classification sur le jeu de données national déjà étiqueté (lui aussi grâce à un algorithme de *machine learning,* d'ailleurs) puis à « prédire », grâce au meilleur modèle obtenu, la discipline associée à chaque publication du *dataset* local.
Alors, certes le ticket d'entrée pour se familiariser avec ce type d'usages de la donnée est un peu élevé quand la plupart des métiers de la data (*data scientist, data engineer, data miner…*) sont des profils historiquement absents de nos organigrammes. Mais le retour sur investissement, même au niveau d'un établissement, peut se révéler très positif, en termes de connaissances métier supplémentaires qui ne sont pas produites par le travail courant sur les collections ou les reporting classiques, ou encore en produisant des algorithmes remplaçant avantageusement des règles métiers difficiles à maintenir et souvent appliquées manuellement (pour des opérations de monitoring sur la qualité des données par exemple).

Géraldine Geoffroy
SCD Université Côte d'Azur
geraldine.geoffroy@univ-cotedazur.fr

[3] http://azur-scd.com/apps/poc-recommandations-engine

[4] https://github.com/azur-scd/poc-loan-predict/blob/main/notebooks/ml_algorithmes.ipynb

[5] L'algorithme est entraîné sur 80 % du *dataset* de départ, et évalué sur les 20 % restants d'observations qui lui sont inconnues en mesurant l'écart entre valeurs réelles et valeurs prédites. La précision mesure la qualité des prédictions positives tandis que le rappel prend en compte les faux négatifs pour évaluer le taux de couverture des prédictions positives.

[6] https://github.com/azur-scd/poc-loan-predict/blob/main/notebooks/data_processing.ipynb

[7] https://azurscd-poc-loan-predict.herokuapp.com

TRANSKRIBUS : l'intelligence artificielle au service du patrimoine documentaire

Lancée en 2015, Transkribus est la première plateforme de reconnaissance automatique des écritures manuscrites. C'est aussi une interface de traitement du patrimoine documentaire.

Transkribus est la première plateforme de reconnaissance automatique des écritures manuscrites (RÉM ou HTR pour *Handwritten Text Recognition*) développée pour mettre en valeur le patrimoine documentaire. Elle a été lancée en 2015, dans le cadre du projet READ (*Research and Enrichment of Archival Documents*) mené par l'université d'Innsbruck (en collaboration avec un consortium de 13 autres universités et centres de recherches européens) et financé par la Commission européenne dans le cadre de l'initiative Horizon 2020 (2016-2019). Aujourd'hui prise en charge par la coopérative READ-COOP SCE, dont les membres sont essentiellement des centres d'archives, des bibliothèques, des universités ou des laboratoires de recherche, Transkribus compte plus de 80 000 utilisateurs partout dans le monde[1].

Tablant sur les avancées de la recherche en intelligence artificielle, Transkribus permet la reconnaissance d'écritures de tous les types (manuscrits ou imprimés), de toutes les époques et dans toutes les langues. Pour ce faire, Transkribus exploite une approche d'apprentissage machine basée sur des réseaux de neurones profonds (*deep neural network*) pour localiser avec précision les lignes de texte dans une image numérique[2] et pour reconnaître chaque caractère de ces lignes en les comparant statistiquement avec les données d'entraînement fournies par l'utilisateur. Avec une centaine de pages transcrites, les utilisateurs peuvent ainsi créer un modèle de reconnaissance spécifiquement adapté à la graphie et à la langue des textes qu'ils souhaitent travailler.

Jusqu'ici, quelque 12 000 modèles de reconnaissance ont été entraînés par les usagers, qui ont permis de transcrire plus de 31 millions de pages[3], incluant des imprimés (ouvrages et journaux) anciens et modernes et des documents manuscrits ou hybrides. Parmi ces modèles, une centaine sont publiquement accessibles à tous les utilisateurs et ce, dans 24 langues différentes, du XI[e] au XXI[e] siècles[4].

DES TAUX D'ERREUR TRÈS FAIBLES

La « performance » des modèles varie évidemment selon la nature des données qui ont servi à les entraîner et la nature des documents à transcrire. Ainsi, il est possible, à partir de quelques dizaines de pages de transcriptions fiables (vérifiées attentivement), d'entraîner pour les imprimés anciens des modèles dont les taux d'erreur se situent sous la barre des 1 %. Avec les manuscrits, la variété des graphies et la qualité des images demeurent les principaux enjeux à relever mais pour des documents dont la graphie est assez constante (tels que les greffes de notaires), 150 pages de transcriptions suffisent pour atteindre des taux d'erreurs sous les 5 %. Les modèles accessibles à tous, entraînés à partir de données fournies par plusieurs utilisateurs, sont basés sur des corpus de plusieurs milliers, voire de dizaines de milliers de pages. Ces modèles, qui prennent en charge une grande variété de graphies dans une langue donnée, atteignent facilement des taux d'erreurs de 5 % à 10 %.

UN MODÈLE D'AFFAIRE COOPÉRATIF

Transkribus n'est pas simplement un logiciel de transcription automatisée, c'est aussi une interface de traitement du patrimoine documentaire permettant l'enrichissement des transcriptions par le balisage et le partage des données, et offrant de puissants outils de recherche. Entre autres, la plateforme comporte un module de balisage de métadonnées calqué sensiblement sur les principes de la TEI, qui facilite l'enrichissement des transcriptions avec des données concernant les individus, les institutions, les lieux, les dates ou toute autre information jugée utile pour la recherche. Les métadonnées balisées peuvent même être ajoutées aux données d'entraînement des modèles de reconnaissance, si bien qu'une partie du balisage peut déjà être effectuée automatiquement par la machine.

Contrairement à certains outils de RÉM (dont eScriptorium), l'utilisation des algorithmes de reconnaissance de texte (manuscrit comme imprimé) a un coût dans Transkribus, qui varie en fonction du statut de l'utilisateur (membre ou non de la READ-Coop, étudiants aux cycles supérieurs), du volume d'achat et de sa récurrence (achat unique, abonnement annuel ou mensuel). Néanmoins, le modèle d'affaire coopératif assure que les revenus sont réinvestis dans l'entretien et l'amélioration des serveurs (extrêmement puissants) et le développement continu de la plateforme et de ses algorithmes, ainsi que dans l'ajout d'outils facilitant la diffusion des contenus transcrits (*read@search*) ou la transcription participative (*citizens&science*).

MAXIME GOHIER
Professeur d'histoire, université du Québec à Rimouski et directeur de Nouvelle-France numérique
Maxime_gohier@uqar.ca

[1] https://readcoop.eu/transkribus

[2] Max Weidemann, *et al., HTR Engine Based on NNs P2: Building Deep Architectures with TensorFlow,* READ-H2020 Project, 2017.

[3] Données à jour en août 2022 (https://readcoop.eu/transkribus).

[4] https://readcoop.eu/transkribus/public-models

[5] https://nouvellefrancenumerique.info

eScriptorium : une application libre pour la transcription automatique des manuscrits

Développée en 2019, l'application eScriptorium dote le logiciel Kraken d'une interface graphique et facilite la conduite de campagnes de transcription automatique.

Cela fait longtemps que la transcription automatique des documents imprimés (OCR) et manuscrits (HTR) intéresse le monde de la recherche et celui des institutions patrimoniales. Le développement de processus s'appuyant sur l'intelligence artificielle et l'augmentation des capacités de calcul ont récemment ouvert de nouvelles perspectives. Dès le début des années 2000, des campagnes d'OCR ont été mises en place pour traiter les imprimés. Pour les manuscrits en revanche, ce n'est qu'à partir du milieu des années 2010 que les choses ont commencé à changer avec l'apparition de logiciels disponibles en ligne sur abonnement comme Transkribus ou en *open source* comme eScriptorium. C'est le groupe de recherche SCRIPTA PSL[1] qui développe, depuis 2019, l'application eScriptorium dont la vocation principale était de doter le logiciel Kraken[2] d'une interface graphique facilitant son utilisation. Kraken est un logiciel de transcription automatique développé en *open source* en 2015 par Benjamin Kiessling et conçu initialement pour proposer une meilleure prise en charge des textes non latins, en particulier arabes. Aujourd'hui, le groupe bénéficie des contributions d'autres infrastructures ou projets de recherche qui ont adopté l'application. Ce fut le cas du projet LectAuRep (Inria/Archives nationales) jusqu'en 2022 ou encore du groupe OpenITI (université du Maryland).

UN ESPACE DE TRAVAIL POUR GÉRER LES ÉTAPES ESSENTIELLES D'UNE CAMPAGNE DE TRANSCRIPTION

L'application eScriptorium sert d'espace de travail pour gérer les étapes essentielles d'une campagne de transcription. Celles-ci sont relativement simples : charger des images (y compris en les extrayant d'un fichier PDF ou d'un serveur IIIF), analyser la mise en page en localisant des ensembles de lignes de texte auxquelles on peut assigner des types, et enfin transcrire. Ces deux dernières étapes peuvent être réalisées à la main ou bien à l'aide de Kraken. À l'issue du processus, des triades composées d'une image, des coordonnées des lignes ou des ensembles, et de la transcription peuvent être exportées dans des formats standards (XML ALTO et PAGE) et servir à générer par exemple des éditions numériques. Ce sont aussi ces triades qui permettent de créer des modèles à l'aide de Kraken, avec ou sans l'intermédiaire d'eScriptorium. Les modèles sont des fichiers qui enregistrent une représentation abstraite des informations telles qu'elles ont été apprises par le logiciel au contact d'exemples de transcription. Cette abstraction permet à un logiciel comme Kraken de générer un texte à partir de l'analyse d'une image. En plus de ces actions essentielles, eScriptorium propose d'autres fonctionnalités pour la gestion de projet : création d'équipe, partage des transcriptions, images et modèles, regroupement des images en « documents », eux-mêmes rangés dans des « projets », étiquetage des documents, suivi de la progression, etc.

LA PRODUCTION DE MODÈLES, PRINCIPAL DÉFI À RELEVER

Pour utiliser eScriptorium, l'application doit être déployée sur un serveur Web installé sur un ordinateur personnel ou sur une machine dédiée. Les capacités de calcul du matériel employé font ensuite la différence au moment de faire tourner Kraken, en particulier lors des entraînements. Certaines institutions ou infrastructures de recherche proposent d'ouvrir des comptes sur leur serveur eScriptorium, mais il est difficile de les recenser toutes. Heureusement, il est aisé de déplacer ses données d'une instance à une autre puisque tout peut être téléchargé.

À l'heure actuelle, eScriptorium propose un modèle de segmentation par défaut efficace mais n'en propose pas pour la transcription : il faut en créer un soi-même ou trouver sur Internet ceux que d'autres utilisateurs de Kraken/eScriptorium ont créés. Certains sont déposés sur Zenodo[3] et des initiatives comme HTR-United[4] permettent de trouver des données à partir desquelles générer ces modèles. La production de modèles, qu'ils soient spécialistes d'une écriture, d'un type de document ou bien généralistes, est l'un des principaux défis à relever pour faire progresser l'implémentation de l'HTR dans les institutions patrimoniales. L'avantage de l'écosystème ouvert de Kraken/eScriptorium réside justement dans le fait qu'il permet aux utilisateurs de créer en autonomie et en toute transparence ces données et ces modèles.

Alix Chagué

Doctorante en humanités numériques au sein de l'équipe ALMnaCH (Inria – Paris) et du GREN (Université de Montréal)

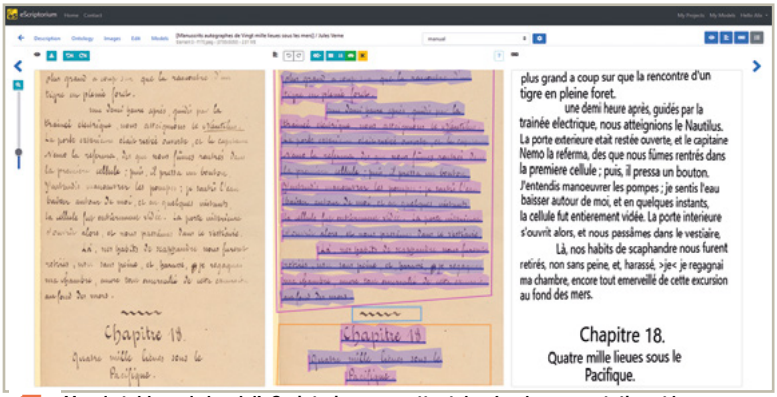
alix.chague@inria.fr

[1] SCRIPTA PSL : https://scripta.psl.eu

[2] Kraken : https://kraken.re

[3] Zenodo : https://zenodo.org/communities/ocr_models

[4] HTR-United : https://htr-united.github.io



Vue du tableau de bord d'eScriptorium permettant de gérer la segmentation et la transcription d'un document manuscrit.

La Bibliothèque Sorbonne Nouvelle : *un projet immobilier d'envergure nationale*

Ouverte le 9 mai 2022 à Paris et signée par l'architecte Christian de Portzamparc, la Bibliothèque Sorbonne Nouvelle renouvelle le paysage des bibliothèques universitaires parisiennes.

Située dans l'Est parisien au cœur du nouveau Campus Nation de l'université Sorbonne Nouvelle, la Bibliothèque Sorbonne Nouvelle - BSN, ouverte depuis le 9 mai 2022, renouvelle le paysage des bibliothèques universitaires parisiennes. Rassemblant toutes les bibliothèques de l'université jusqu'alors multi-sites, ce nouvel équipement, conçu par l'architecte Christian de Portzamparc, est désormais la plus grande BU spécialisée en lettres, arts, sciences humaines et sociales de Paris *intramuros,* labellisée CollEx pour huit fonds.

UNE NOUVELLE BU À L'ARCHITECTURE REMARQUABLE POUR UN NOUVEAU CAMPUS

Une donnée déterminante du projet est qu'il ne concerne pas uniquement le SCD[1] – dénommé DBU[2] à la Sorbonne Nouvelle, mais toute l'université. Celle-ci, anciennement « Paris 3 », émanant directement de la faculté des Lettres de l'université de Paris, occupait depuis sa création en 1970 des locaux d'enseignement et de recherche répartis dans plusieurs quartiers parisiens. Il en était de même pour le SCD composé de 10 bibliothèques[3]. La nécessité de doter l'université de locaux neufs date des années 2000 : ses infrastructures vétustes et énergivores deviennent insuffisantes pour répondre à l'augmentation du nombre d'étudiants et aux nouvelles pratiques pédagogiques. Plusieurs projets d'extension se sont succédé sans aboutir. C'est en 2010 que l'impulsion est donnée avec le « Rapport Larrouturou » à la ministre de l'Enseignement supérieur et de la Recherche, *Pour rénover l'enseignement supérieur parisien*, d'après lequel Paris 3 – Sorbonne Nouvelle « est l'université parisienne qui connaît la situation immobilière la plus critique ». L'« urgence à décider et mettre en œuvre un plan précis » déclenche le projet Campus Nation chapeauté par une MOA[4] publique, l'EPAURIF[5], et une MOE[6] de renom : l'agence Christian de Portzamparc.

LES USAGERS, FIL ROUGE DU PROJET DE LA FUTURE BIBLIOTHÈQUE DU CAMPUS NATION

De 2013, date de conception du programme immobilier, à 2022, année au cours de laquelle où l'université emménage sur son nouveau campus, les bibliothécaires élaborent la future BU en suivant un même cap : les besoins des publics. Le SCD a exploité durant ces 9 années le travail de préfiguration initié dans les années 2000 lors des projets immobiliers successivement avortés afin de cartographier les usages réels et potentiels. Il est à noter le rôle stratégique joué par les congrès internationaux et les visites de bibliothèques en France et à l'étranger[7]. Organisées par la direction du SCD pour son équipe afin d'observer ce qui fonctionne (ou pas !) en vue du futur bâtiment, elles ont aussi été proposées à la gouvernance de l'université pour sensibiliser les interlocuteurs idoines aux enjeux auxquels un nouveau campus se doit de répondre en se dotant d'une BU qui en soit la vitrine. Ont découlé de cette phase de visites les maîtres-mots d'évolutivité, modularité et connectivité que le SCD a inscrits comme caractéristiques-phares de la partie Bibliothèque du Programme immobilier voté en 2014 par le CA de l'université.

Ensuite, la culture des enquêtes et de l'amélioration continue de l'accueil (Marianne, SP+) intégrée à la politique d'établissement du SCD a déterminé les échanges avec les divers acteurs du projet (université, MOA, AMO[8], MOE, entreprises) : études quantitatives auprès des usagers (2014-2015) ; entretiens avec les enseignants-chercheurs (2016) ; enquêtes UX[9] flash Nation (2017-2018) ; atelier avec les enseignants-chercheurs (2019) ; horaires de la future BU (2019) ; quels nouveaux magazines ? (2020) ; quel nom pour la BU Campus Nation ? avant (2020) ; bibliothèques pendant le 2e confinement (2021).

Le principe de l'usager comme fil conducteur a ainsi perduré durant la genèse et la réalisation du projet. Or, que d'évolutions entre 2000 et 2021, parmi lesquelles celles induites par la crise sanitaire et l'essor du distanciel apparus en fin de construction.

UN « ÎLOT OUVERT » DURABLEMENT ÉVOLUTIF, MODULAIRE ET CONNECTÉ

Outre la problématique de s'adapter à des usages qui évoluent dans le temps, il était fondamental que la nouvelle BU soit en

Crédit Bibliothèque Sorbonne Nouvelle

Crédit Bibliothèque Sorbonne Nouvelle

Crédit Bibliothèque Sorbonne Nouvelle

LES CHIFFRES CLÉS

1 100 PLACES RÉPARTIES SUR 4 NIVEAUX PUBLICS :

- **5 000 m²** d'espaces publics
- **2 000 m²** d'espaces internes (bureaux et magasins)
- **38 salles de travail en groupe** insonorisées d'une capacité de 4, 6 ou 8 personnes dont :
  – 3 réservées aux enseignants-chercheurs
  – 1 prévue pour le public en situation de handicap avec une offre de logiciels et de matériels adaptés
- **22 carrels** pour 1 ou 2 personnes
- **2 salles de formation** d'une capacité de 25 personnes
- **1 salle « innovation »** particulièrement modulaire d'une capacité de 25 personnes
- **1 Biblio Café** à l'entrée

JUSQU'À 72H30 D'OUVERTURE HEBDOMADAIRE

8 FONDS LABELLISÉS COLLEX :

- 600 000 livres et revues papier
- 1 100 BD
- 8 800 films et séries en DVD et Blu-ray
- 63 000 ebooks, 14 800 revues en ligne, 194 bases de données, VOD.

mesure de répondre aux besoins exprimés durant les enquêtes et observations menées : se connecter facilement (davantage de prises et d'équipements, meilleur accès à Internet), se concentrer au calme, travailler en groupe, se retrouver et faire des pauses, bref, avoir un cadre de travail mais aussi un lieu de vie convivial au cœur du campus. Quel service, sinon la BU, pour assurer cette mission ?
Le défi : comment faire coexister dans un même lieu des attentes qui évoluent et sont *a priori* concurrentes et traduire ces différents besoins dans une bibliothèque à la superficie somme toute restreinte ? C'est en gardant en ligne de mire les principes d'évolutivité, modularité et connectivité que le SCD a traduit en fonctionnalités et termes immobiliers le panel des besoins utilisateurs constatés : un faux-plancher technique pour ajouter ou déplacer facilement prises, mobiliers, équipements ; des dispositifs phoniques pour une acoustique de qualité ; préférer des salles de travail en groupe de petites et moyennes capacités en plus grand nombre à celles de capacités plus grandes (moins flexibles et moins nombreuses) ; luminaires individuels complémentaires aux suspensions et plafonniers pour un confort visuel à la carte ; nombreuses ouvertures pour favoriser la lumière naturelle plutôt qu'artificielle, dont celles permises par des fenêtres en forme de trapèze, devenues l'emblème architectural du campus – et le logo de l'université depuis l'adoption de sa nouvelle charte graphique en 2019.
La BSN à cet égard est un compromis réussi entre attentes architecturales et attentes fonctionnelles parfois renvoyées dos à dos. Lorsqu'on visite la BSN on comprend ce que signifie le concept d'« îlot ouvert »[10] cher à Portzamparc.
Telle est l'expérience BSN : espace dense mais lumineux et ouvert, cocon hyperconnecté, concentré d'espaces et d'ambiances variés pour rendre possible une multitude d'usages au même endroit : travailler seul ou en groupe avec des équipements facilitant le *coworking* (salles insonorisées, partage d'écran, tableaux), étudier au calme ou discuter, contenter les fervents du « BYOD » pour « *bring your own device* » traduit en français par PAP pour « prenez vos appareils personnels », aussi bien que les usagers victimes de la fracture numérique en équipant chaque place de prises électriques et Ethernet pour accéder à Internet en filaire en complément du wifi, emprunter des matériels (ordinateurs, liseuses, casques, chargeurs), lire une BD, visionner un film ou une série en DVD/Blu-ray sur grand écran, consulter des livres anciens patrimoniaux ou dessins rares, faire des photocopies, scans, impressions, boire un café en lisant la presse, assister à une action culturelle autour des fonds documentaires… et d'autres usages qui émergeront à court ou long terme !
La satisfaction des usagers, finalité visée depuis 2000, paraît atteinte à la lecture des verbatims[11] recueillis depuis l'ouverture de la BSN en mai 2022 dans son Livre d'or, son enquête flash d'ouverture (16-21 mai 2022) ou ses réseaux sociaux[12].

Brigitte Auby-Bucherie
Directrice de la Direction des bibliothèques universitaires de l'université Sorbonne Nouvelle
Brigitte.auby-bucherie@sorbonne-nouvelle.fr

Floriane Berti
Directrice adjointe de la Direction des bibliothèques universitaires de l'université Sorbonne Nouvelle
floriane.berti@sorbonne-nouvelle.fr

[1] SCD : service commun de la documentation, régi par les articles D714-28 à D714-40 du Code de l'éducation.

[2] DBU : direction des bibliothèques universitaires, https://www.dbu.univ-paris3.fr/accueil-dbu

[3] Une BU tête de réseau (site Censier) ; 7 bibliothèques d'UFR intégrées au SCD spécialisées en langues et littératures à Censier, Bièvre, Sorbonne, Monde anglophone et Dauphine (ESIT) ; 2 bibliothèques associées : la Bibliothèque Gaston-Miron – Études québécoises (Délégation générale du Québec) et la Théâtrothèque Gaston Baty (UFR Arts et Médias de l'université).

[4] MOA : maîtrise d'ouvrage.

[5] EPAURIF : Établissement public d'aménagement universitaire de la région Ile-de-France.

[6] MOE : maîtrise d'œuvre.

[7] Paris, Lille, Le Havre, Aix-Marseille, Dijon, Glasgow, Helsinki, Amsterdam, Berlin.

[8] AMO : assistant(s) à la maîtrise d'ouvrage.

[9] UX : *user experience*, « expérience utilisateur »

[10] Îlot ouvert : bâtiments à hauteur de piétons avec des rez-de-chaussée animés par contraste avec les modèles dominants d'urbanisme parisiens hauts et fermés, luminosité et espaces verts pour ouvrir sur l'extérieur malgré l'étroitesse des sites Paris intramuros.

[11] https://www.dbu.univ-paris3.fr/images/DBU/Les_usagers_racontent_leur_1ere_visite_BSN.pdf

[12] Rendez-vous sur Instagram pour voir la BSN comme si vous y étiez, sur Facebook pour les actualités pratiques (@bibliothequesorbonnenouvelle) et sur Twitter pour les infos recherche (@bibliothequeBSN)

Portrait

Gwenaëlle MARCHAIS,

responsable de la Bibliothèque électronique du SCD de l'université de La Réunion

Parlez-nous de vos fonctions actuelles…
Je suis à la fois gestionnaire des ressources numériques, avec un fort accent sur la politique documentaire, et administratrice des logiciels et sites Web du service commun de la documentation de l'université de La Réunion. J'apprécie la technicité de ce travail et l'évolution constante des outils. J'interviens comme formatrice en culture numérique et scientifique sur les publications, l'intégrité scientifique, l'identité numérique. Depuis 2021, j'assure aussi le rôle de coordinatrice Thèses et j'accompagne les doctorants pour le dépôt légal des thèses. Sensibiliser à la diffusion en accès libre et à la qualité des documents fait désormais partie de mon quotidien.

Quelles sont les étapes qui vous semblent les plus importantes dans votre parcours professionnel ?
À ma sortie de l'Enssib en 2004, j'ai pris un poste d'adjointe à la réinformatisation d'un grand SCD parisien. J'ai plongé dans le vif du sujet : travailler avec des équipes variées dont certaines découvraient le Sudoc, implémenter des logiciels d'un nouveau type, dans un contexte de pleine expansion des abonnements numériques. Un rôle d'intermédiaire passionnant qui demandait d'abord d'observer les pratiques pour comprendre les besoins de continuité et d'évolution. C'était très formateur. J'en ai gardé le goût de l'analyse des circuits de travail et de l'animation d'équipes.
J'ai ensuite pris en charge la documentation électronique, des acquisitions à l'ouverture des accès hors campus, un service inhabituel à l'époque. La demande d'assistance a été forte. J'ai adoré le travail d'invention de solutions.
À La Réunion, j'ai découvert un établissement soucieux de valoriser ses collections et de moderniser sa présence numérique. Mon équipe a bénéficié d'une belle autonomie pour monter de nouveaux projets : service de renseignement en ligne, blog de médiation documentaire, animation de comptes de réseaux sociaux.
Et puis comment oublier ma première participation à Cyclobiblio, événement professionnel collectif majeur selon moi ? Je laisse les curieux se renseigner…

Quelles sont vos relations avec l'Abes ?
Je suis une grande consommatrice de la documentation de l'Abes. Nous échangeons aussi régulièrement, *via* le guichet d'assistance ou par e-mail, les collègues sont toujours très réactifs. C'est aussi un vrai réseau. Depuis 2010 et les premières grandes enquêtes, l'Abes anime la réflexion professionnelle. On connaît la suite : SGBm, Bacon, etc.

Quels défis majeurs, d'après vous, aura à relever l'Abes dans les prochaines années ?
Rester attractive et conserver des équipes aussi compétentes et efficaces. Techniquement, les attentes sont fortes sur STAR, l'intranet des thèses, et sur l'évolution du Sudoc. Rester un laboratoire d'expérimentation sur des sujets tels que l'intelligence artificielle, l'exploration de données, le Web sémantique. Et conforter son rôle parmi les grands acteurs nationaux et internationaux de l'information scientifique et technique.

Qu'appréciez-vous le plus dans votre métier ?
Je dois rester en phase avec l'actualité numérique et académique ; mes activités se renouvellent sans cesse et j'apprends constamment. Je suis aussi fière d'appartenir à une profession qui contribue à la construction et à la transmission des savoirs. J'apprécie particulièrement les moments d'échange avec nos publics, débutants comme avancés. Je me sens utile.

Qu'est-ce qui vous énerve le plus ?
Je pense au « solutionnisme » technologique et aux objectifs de « changement » : on confond parfois les moyens (une technologie, un changement) et les fins (les objectifs). Qu'améliorer, mais aussi que préserver ? Au nom de quelles missions ? Je suis attentive à l'adhésion des équipes et à tous ces fonctionnements implicites efficaces, collectifs, qui donnent du sens et du pouvoir d'agir aux collègues.

Quelle image donneriez-vous pour qualifier l'Abes ?
Une accolade. J'allais répondre une poignée de main pour illustrer le contrat que l'Abes remplit chaque jour, qui est de mettre une équipe d'experts au service d'une communauté. Mais la gentillesse des collègues et leur esprit de convivialité pendant les Journées Abes invitent à l'accolade.

Votre expression favorite ?
« Dann oui na poin batay »[1].

[1] Traduction du créole réunionnais : « Dans oui, il n'y a pas de bataille » (Il est facile de dire oui pour éviter la discussion).