Par ses outils développés pour Gallica et son récent DataLab, la Bibliothèque nationale de France propose une nouvelle génération de services pour l'exploration des ressources numériques et la production de connaissances.

La BnF et les services à la recherche *à l'heure des humanités numériques*

Par la richesse et la particularité de ses collections, les liens qui unissent la Bibliothèque nationale de France (BnF) au monde de la recherche sont nombreux et anciens. Cet attachement fort lui permet de suivre et d'accompagner les chercheurs qui, s'appuyant sur la disponibilité de vastes collections, renouvellent les pratiques de recherche en sciences humaines et sociales. L'augmentation massive des collections numériques (numérisations sur Gallica, archives de l'Internet, métadonnées du catalogue, notices d'autorité, etc.) a favorisé l'apparition d'un nouveau patrimoine dématérialisé et ouvert des pistes de recherche : fouille de textes et d'images, lecture distante, data visualisation, réutilisation et alignement de référentiels, entre autres.
Confrontée à des demandes récurrentes de mise à disposition de corpus massifs, la BnF a su au fil du temps adapter ses services pour répondre aux enjeux liés à ces nouvelles pratiques de recherche.

FACILITER LA COMPRÉHENSION DES COLLECTIONS

Depuis sa création en 1994, la BnF s'est interrogée sur l'articulation entre accroissement exponentiel des collections et diffusion des savoirs : devant une masse sans cesse grandissante de documents, comment diffuser au mieux et au plus grand nombre ? Comment rendre cette masse non seulement disponible mais également exploitable pour la production des savoirs ?
La dématérialisation et le recours au réseau pour communiquer et valoriser les collections ont été mis en oeuvre dès 1997 avec la création de Gallica, la bibliothèque numérique de la BnF et de ses partenaires, qui comptait à son lancement 2 500 ouvrages et 10 000 images.
Depuis, les collections numériques de la BnF n'ont cessé de s'étoffer et représentent aujourd'hui environ 6 pétaoctets. Elles sont caractérisées par une variété considérable : documents numérisés (9 millions de documents disponibles dans Gallica), documents nativement numériques, numériques sur support (CD, DVD, jeux vidéo), archives de l'Internet, métadonnées bibliographiques et données d'autorité. Autant d'ensembles de données hétérogènes, qui ont toutes des structures, formats, qualité, contextes de production, fonctions et contenus différents. La compréhension de ces collections exige des traitements spécifiques et par conséquent des compétences et expertises particulières, aussi bien pour les conserver ou les communiquer que pour les analyser.
La BnF travaille non seulement à mettre à disposition ces collections mais aussi à en faciliter la compréhension et la manipulation par la mise en place d'une gamme d'outils dédiés aux opérations suivantes : aide à la constitution de corpus (interface de recherche par proximité, par occurrence, par similarité) ; extraction des documents (portail API et jeux de données, export csv du Catalogue général) ; recherche avancée dans les métadonnées (data.bnf.fr et le *sparql endpoint*).

DES OUTILS POUR MIEUX APPRÉHENDER LES DOCUMENTS DANS GALLICA

D'abord numérisés en mode image, les documents textuels ont, depuis 2005, systématiquement été convertis en mode texte par un logiciel OCR, autorisant la recherche plein texte et l'extraction de corpus. À ce jour, une partie seulement de la collection numérique est accessible en mode texte. Pour y remédier, la BnF a engagé un ambitieux programme de rétroconversion.
De plus, toutes les collections textuelles ne sont pas éligibles à l'OCR. C'est le cas bien sûr des manuscrits, mais aussi d'une partie des collections de la presse dont les caractères, la mise en page, l'état de conservation rendent difficile la reconnaissance des caractères. L'un des axes de la feuille de route 2021-2026 de la BnF autour de l'intelligence artificielle se concentre sur la reconnaissance automatique des écritures manuscrites (HTR), dont les technologies et outils arrivent aujourd'hui à maturité.
Au-delà de la transcription du texte, ces technologies permettront de naviguer à l'intérieur du document et de reconnaître les zones de contenu sur la page (segmentation), d'extraire la structure logique du document grâce à la technologie OLR (*Optical Layout Recognition*), d'identifier les entités nommées (REN). La catégorisation de ces objets dans des classes permettra d'améliorer les fonctionnalités de consultation et de présentation de Gallica.
L'interface de consultation de Gallica a fortement évolué depuis sa création. Décrits dans un format Dublin Core, les champs de la recherche depuis l'interface de Gallica sont moins riches que ceux du Catalogue général. Cependant, l'océrisation des documents permet une recherche plein texte, et même une recherche sémantique par proximité de termes. Les résultats ainsi obtenus peuvent être analysés, comparés et extraits grâce au rapport de recherche, dans des formats divers. Grâce aux API (IIIF, API documents, métadonnées), en service depuis 2017, les usagers peuvent extraire les contenus de Gallica à distance et lancer des requêtes sur les corpus. Si les images représentent un corpus important dans Gallica, il n'est pour l'instant pas encore possible de faire de la recherche iconographique à l'intérieur des documents, comme dans les corpus de presse, par exemple. Des prototypes, comme par exemple *GallicaPix* et le programme *GallicaSnoop,* développé en partenariat avec l'INA-Inria, travaillent à transformer la bibliothèque numérique Gallica en banque d'images grâce à des technologies de segmentation et de reconnaissance de forme.
Ces outils, déjà opérationnels ou en phase de développement, ont été pensés pour aider l'usager -chercheur, professionnel des bibliothèques ou simple lecteur - à naviguer au mieux dans les collections riches et complexes de la BnF. Cependant, la maîtrise et l'appropriation des collections numériques, soumises à la coordination de multiples expertises (sur les collections, les catalo-

gues, les outils d'extraction, les formats de données), imposent aux chercheurs comme aux professionnels des bibliothèques une évolution de leurs pratiques.
C'est pour répondre aux nouveaux usages et besoins des chercheurs et assurer une coordination des expertises à l'échelle de l'établissement que la BnF a mis en place un nouveau service, le BnF DataLab.

LE BNF DATALAB, UN LABORATOIRE EN RECHERCHE ET DÉVELOPPEMENT

Inauguré le 18 octobre 2021, le BnF DataLab est à la fois un espace physique et un ensemble de services destinés à accompagner les chercheurs qui souhaitent travailler sur les collections numériques de la BnF. Plus qu'un simple point d'information ou une offre de services, le BnF DataLab a été imaginé comme un laboratoire, un lieu d'échange, d'interaction et de formation entre pairs, permettant des expérimentations et des collaborations entre équipes de recherche en résidence. Il a été pensé pour travailler en partenariat avec des chercheurs de différents profils et rattachés à différentes institutions scientifiques. Le partenariat privilégié avec Huma-Num, infrastructure de recherche dédiée aux humanités numériques, a pour objectif d'articuler l'offre de services d'Huma-Num avec celle du BnF DataLab afin de permettre un accompagnement complet autour de la recherche en humanités numériques.
L'accueil d'un ingénieur d'étude ainsi que l'appel à projet conjoint Huma-Num/BnF lancé en juin 2021 sont les premières formes concrètes de ce partenariat.
Ce partenariat, comme ceux qui suivront, marque la volonté de valoriser les résultats de la recherche. Les outils développés dans le cadre de programmes de recherche ou d'appels à projets seront conservés dans une sorte de boîte à outils et réutilisables pour d'autres projets.

CRÉER UNE COMMUNAUTÉ DE RECHERCHE AUTOUR DES COLLECTIONS DE LA BNF

L'enjeu est de créer une communauté de recherche, dans un lieu identifié au sein de la BnF et autour de ses collections, où chercheurs, bibliothécaires et ingénieurs peuvent échanger et porter ensemble des projets grâce à des expertises complémentaires. Une programmation de manifestations scientifiques (atelier de démonstration d'outils, colloques, séminaires, valorisation de travaux) a été mise en place dès 2020 pour faire vivre cette communauté et favoriser les échanges entre les différentes cultures professionnelles.

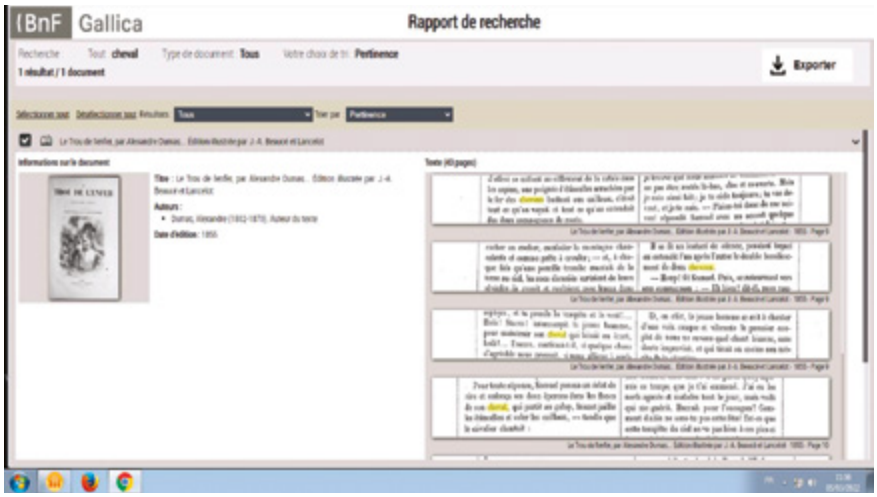

Image libre de droits

Le rapport de recherche dans Gallica

En partant des cas d'usages, un catalogue d'une vingtaine de services, basé sur le cycle de vie des projets de recherche, a été élaboré pour accompagner les projets depuis la formulation du besoin, la constitution et le travail sur le corpus jusqu'à la valorisation des résultats.
Les services autour de l'accueil et de l'orientation constituent une phase cruciale : définir une problématique en termes de recherche bibliographique, déterminer ce qui est disponible, ce qui ne l'est pas, comment le rendre disponible et sous quel format sont autant d'étapes indispensables dans la construction du corpus. Les collections et les catalogues étant complexes, établir un corpus requiert souvent des recherches croisées et nécessite déjà plusieurs niveaux d'expertise. Il importe donc de centraliser les demandes pour construire un parcours cohérent et mobiliser les bons services. Le BnF DataLab est donc avant tout un service de coordination, qui doit permettre de coconstruire un parcours avec les chercheurs et les services internes : aide à la recherche, formation aux outils d'extraction des documents ou des données, à la manipulation des formats.
Une fois le corpus constitué, il paraissait important d'offrir un environnement de travail adapté à la fouille de données, en particulier pour les corpus sous droits (archives du Web, Gallica Intramuros, jeux vidéo) qui ne peuvent être consultés que dans les espaces recherche de la BnF. Un portail dédié, le DataLab Center, a été conçu pour empêcher la fuite des données sous droit, tout en autorisant les usagers à accéder à une machine virtuelle (VM), à stocker et travailler sur les corpus. Un accès multimode offre aux usagers du BnF DataLab la possibilité de basculer aisément d'une VM sans accès à Internet, lieu de stockage du corpus, à une VM connectée à l'internet, permettant ainsi la mise à jour des outils ou le chargement de bibliothèques de scripts.
Si les humanités numériques transforment profondément le travail des chercheurs, elles modifient tout autant le métier des professionnels des bibliothèques. Introduire les humanités numériques dans la bibliothèque est un défi qui nécessite à la fois de nouvelles compétences métiers (experts, ingénieurs, etc.) et de nouvelles infrastructures en termes d'espaces et d'équipements : serveurs, machines virtuelles, logiciels, déploiement d'API.
C'est une nouvelle génération de services en bibliothèques que préfigure le BnF DataLab, articulant espace physique et virtuel, accueillant des pratiques mixtes, individuelles et collectives, tournées à la fois vers l'exploration des ressources numériques et la production de connaissance et d'outils. Cette évolution des usages et des métiers, qui ne concernent encore qu'un nombre restreint de chercheurs et de bibliothèques, n'en change pas moins le rapport aux collections et à l'espace même de la bibliothèque qui, à l'instar du BnF DataLab, se réinvente au plus près des besoins des usagers.

Marie Carlin
Conservatrice, coordinatrice du BnF DataLab
marie.carlin@bnf.fr

Arnaud Laborderie
Chef de projet Gallica, département de la Coopération, Bibliothèque nationale de France
arnaud.laborderie@bnf.fr