La Bnf, notamment au travers du projet Corpus, met toute la richesse de ses métadonnées à la disposition des chercheurs.

BnF : des métadonnées au service de projets de recherche innovants

{BnF

Les collections numériques de la Bibliothèque nationale de France (BnF) représentent aujourd'hui plusieurs pétaoctets de données, avec une grande diversité de structures, de formats, de qualité, de contextes de production, de fonctions et de contenus : documents numérisés consultables dans Gallica, documents nativement numériques sur support (jeux vidéo, CD et DVD,…) ou dématérialisés et collectés via les archives de l'Internet, métadonnées bibliographiques et d'autorité décrivant les personnes, lieux, organisations, concepts, etc. Ces nouvelles collections, disponibles sous forme numérique et dès lors susceptibles d'être analysées par des outils informatiques de plus en plus performants, peuvent désormais être assimilées malgré leur diversité à un ensemble homogène : les données. On parle ainsi dans le monde anglo-saxon de « *collections as data* »[1]. Elles se définissent principalement par leur usage, leur nature numérique ouvrant des opportunités inédites pour la recherche, notamment en sciences humaines et sociales, ce qu'on appelle les « humanités numériques ».

UN PROJET FONDATEUR SUR LA GRANDE GUERRE

Cette tendance se vérifie, dans les bibliothèques nationales en particulier, depuis plusieurs années. À la BnF, nombre de projets de recherche portant sur des corpus numériques mobilisent conjointement l'expertise des chercheurs, des compétences bibliothéconomiques pour la mise à disposition des données (y compris par la numérisation de corpus massifs), et des compétences informatiques pour les exploiter. Certains de ces projets ont fait l'objet d'un partenariat avec la BnF, qui s'y engage dans l'espoir d'améliorer ses propres outils de production, de gestion et d'accès, d'augmenter la visibilité de ses collections, ou encore de mieux connaître les usages de ses publics.

Le projet « Le devenir du patrimoine numérisé en ligne : l'exemple de la Grande Guerre » a fait partie de ces projets fondateurs. Conduit de 2013 à 2016 dans le cadre du Labex « Les passés dans le présent » et porté par la BnF, Télécom ParisTech et la BDIC (devenue depuis « La contemporaine »), il visait à étudier les pratiques d'appropriation des documents numérisés mis en ligne par les institutions patrimoniales[2]. Pour cela, une analyse automatique du réseau des sites web français concernant la Grande Guerre et une cartographie des liens entre ces sites ont été réalisées, s'appuyant sur une collecte d'archives web réalisée spécifiquement par la BnF. En parallèle, le projet avait vocation à développer des outils et à proposer des méthodes reproductibles pour analyser un corpus d'archives web.

LE PROJET CORPUS

Confrontée à une montée en puissance de ce type de projets autour des collections numériques et des données, la BnF s'est lancée en 2016 dans

[1] https://collectionsasdata.github.io

[2] *Le web français de la Grande Guerre : réseaux amateurs et institutionnels.* Sous la direction de Valérie Beaudouin, Philippe Chevallier et Lionel Maurel. Presses universitaires de Paris Nanterre, 2018.

Partie centrale de la voie lactée

© X-RayNASACXCUMassD. Wang et al.

la construction d'une nouvelle offre de services aux chercheurs. Inscrit au plan quadriennal de la recherche de la BnF pour 2016-2019, le projet Corpus[3] avait pour objectif de faciliter l'accès des chercheurs aux collections numériques de la BnF, en vue de permettre leur exploitation sous forme de données : fouille de textes et d'images, data visualisation, réutilisation et alignement de référentiels, etc. Des corpus issus de trois principaux ensembles – les archives de l'Internet, les documents numérisés et les métadonnées – ont ainsi été étudiés de manière expérimentale et itérative.

La première année, un partenariat avec l'équipe du projet ANR Web90 a débouché sur l'élaboration d'une application « Archives web Labs » proposant l'indexation en plein texte de deux corpus : les « incunables du web » (1996-2000) et la collecte « attentats » de 2015. En 2018, ces nouvelles fonctionnalités ont été déployées sur tous les postes d'accès aux ressources numériques de la bibliothèque de recherche. Un nouveau corpus, la collecte « Actualités » (2010-2017), est venu s'y ajouter dans le cadre d'un autre projet de recherche appelé « Neonaute »[4]. En 2017, la BnF s'est associée au projet « Giranium », conduit par une équipe du CELSA (Laboratoire de sciences de l'information et de la communication de Sorbonne Université), dont l'objectif était l'étude des premières industries culturelles et médiatiques en France à travers le prisme d'Émile de Girardin, personnalité emblématique du journalisme français du XIXe siècle. Outre la numérisation d'un corpus de presse du XIXe siècle, ce partenariat a été l'occasion d'expérimenter d'autres dimensions concrètes de ce que pourrait être une potentielle offre de services autour des collections numériques : le besoin d'espaces de travail dédiés (pour le travail en groupe) et la conduite d'ateliers méthodologiques sur les humanités numériques (notamment autour des formats, standards, pratiques de structuration, normalisation et pérennisation des informations)[5].Enfin, la troisième année du projet a porté sur la réutilisation des métadonnées bibliographiques et d'autorité à des fins de recherche, avec le projet ANR « Foucault Fiches de Lecture » qui avait pour objectif de numériser, mettre en ligne et enrichir les notes de lecture manuscrites de Michel Foucault, en utilisant une plate-forme numérique de travail collaboratif.

MIEUX CERNER LES BESOINS

Parallèlement à ces expérimentations menées en collaboration avec les chercheurs, une étude a été conduite en 2017 afin de mieux cerner les besoins des équipes de recherche, notamment en termes d'espaces dédiés. Fondée sur une enquête qualitative par entretiens, des observations informelles et un atelier participatif, l'étude relève le besoin des équipes de recherche de disposer des collections numériques à distance et explore la valeur ajoutée potentielle d'un espace physique à la BnF[6]. Outre la nécessité d'un tel espace pour la consultation et l'analyse de corpus sous droits, la proximité des experts de la BnF est perçue comme la principale valeur ajoutée d'un lieu physique. Dans la logique d'un dialogue renouvelé entre milieu de la recherche et bibliothèques, ce futur espace favoriserait la formulation de nouvelles questions scientifiques portant sur les collections numériques, la bibliothèque apportant une expertise sur les collections, les questions juridiques et les aspects techniques. Une infrastructure et des outils logiciels, dédiés notamment à la fouille de données, y seraient déployés.

Le site *Api et jeux de données*[7], ouvert en 2017 à l'occasion du 2e hackathon de la BnF, apporte le volet numérique de l'offre en documentant les API d'accès aux collections numériques et en redistribuant les données enrichies par les chercheurs.

Le département de l'Orientation et de la recherche bibliographique (ORB) et sa salle de lecture du rez-de-jardin, la salle X, sont pressentis pour accueillir à partir de 2020 ce nouvel espace et le service associé[8]. Fin 2018, un accompagnement de l'Acco{Lab, structure interne à la BnF qui mobilise des méthodologies relevant de l'innovation participative pour faciliter la conduite de projets au sein de l'établissement, a permis aux agents du département ORB de commencer à s'approprier cette mission nouvelle et, en particulier, de définir les contours d'un parcours de formation allant des humanités numériques à la construction d'un corpus test, parcours qui sera mis en place progressivement en 2019 et 2020.

ÉLABORER DES PARTENARIATS

La dernière étape de la construction de ce nouveau service aux chercheurs réside dans l'élaboration de partenariats. Qu'il s'agisse du CNRS (via l'INSHS, la TGIR Huma-Num ou des laboratoires comme le Lattice[9]) ou d'établissements d'enseignement supérieur (Sorbonne Université, École Polytechnique de Lausanne), plusieurs acteurs ont déjà manifesté leur intérêt pour la dynamique qui pourrait émerger d'un tel projet. Dans le cadre de CollEx-Persée, un groupe de partenaires s'est constitué pour porter l'idée d'un réseau national de compétences, capable de relayer ces services dans les universités du territoire. Au niveau international, des contacts ont été établis avec le groupe *Digital Humanities* de LIBER[10] et le réseau des « *Library labs* » qui se fédère autour de la *British Library*[11]. Autant de perspectives qui confirment l'importance de développer ces nouvelles pratiques autour des collections et données numériques, dans les bibliothèques de recherche en général, et à la BnF en particulier.

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

[3] https://hal-bnf.archives-ouvertes.fr/hal-01739730

[4] Le projet « Neonaute » portait sur la réalisation d'un moteur de recherche et d'études terminologiques sur les néologismes dans la langue française.

[5] Les comptes rendus des ateliers sont en ligne sur le carnet de recherche de la BnF : https://bnf.hypotheses.org

[6] Eleonora Moiraghi, *Le projet Corpus et ses publics potentiels. : Une étude prospective sur les besoins et les attentes des futurs usagers*. Paris : Bibliothèque nationale de France, 2018. En ligne : https://hal-bnf.archives-ouvertes.fr/hal-01739730

[7] http://api.bnf.fr

[8] Éloi, Catherine, Moiraghi, Eleonora et Rose, Virginie. « Un espace pour les humanités numériques à la BnF », *Bulletin des bibliothèques de France,* 2019, n°17, p. 90-95 : http://bbf.enssib.fr/consulter/bbf-2019-17-0090-009

[9] www.lattice.cnrs.fr

[10] https://libereurope.eu/strategy/digital-skills-services/digitalhumanities

[11] https://pro.europeana.eu/post/building-library-labs-what-do-they-do-and-who-are-they-for