Le patrimoine numérique national à l'heure de l'intelligence artificielle. Le programme de recherche Corpus comme espace d'expérimentation pour les humanités numériques

Emmanuelle Bermès, Eleonora Moiraghi

Le présent article a été soumis en 2018 à l'appel à contributions au numéro spécial "Intelligence artificielle et humanités numériques" de la Revue d'Intelligence Artificielle (RIA), publiée à l'époque aux éditions Lavoisier.
La construction de ce numéro spécial a suivi le processus de sélection d'une revue à comité de lecture avec une relecture anonyme par deux rapporteurs.
Le numéro spécial va paraître prochainement.

Le patrimoine numérique national à l'heure de l'intelligence artificielle.

Le programme de recherche Corpus comme espace d'expérimentation pour les humanités numériques.

Emmanuelle Bermès[1], Eleonora Moiraghi[2]

1. Bibliothèque nationale de France
Quai François Mauriac, 75706 Paris Cedex 13, France
emmanuelle.bermes@bnf.fr

2. Bibliothèque nationale de France
Quai François Mauriac, 75706 Paris Cedex 13, France
eleonoramoiraghi@gmail.com

RÉSUMÉ. Dans un contexte d'augmentation des volumétries des données et de réduction des temps de traitement, la Bibliothèque nationale de France est confrontée à plusieurs défis et évolutions. Afin de collecter, préserver, décrire et permettre l'étude d'ensembles de données massifs et hétérogènes, elle fait non seulement appel aux méthodes relevant des sciences de l'information mais elle recourt aussi aux techniques issues de l'informatique, de plus en plus développées dans le domaine de l'intelligence artificielle. Cette nécessité croissante de convoquer des compétences complémentaires, s'ajoutant aux opportunités ouvertes par les collections numériques pour la recherche, notamment en sciences humaines et sociales, induit pour la Bibliothèque la définition d'un espace pour le développement des humanités numériques.

ABSTRACT. In a context of increasing volumes of data and reduced processing times, the National Library of France is facing several challenges and developments. In order to collect, preserve, describe and enable the study of massive and heterogeneous data sets, the Library uses not only methods of information sciences but also techniques developed in the field of computer science, especially in artificial intelligence. This growing need to convene complementary skills, combined with the research opportunities opened by these digital collections, has led the Library to create a space for supporting digital humanities.

MOTS-CLÉS : patrimoine numérique, fouille de données, apprentissage profond, intelligence artificielle, humanités numériques, sciences de l'information, mégadonnées.

KEYWORDS: digital heritage, digital corpora, data mining, artificial intelligence, machine learning, deep learning, digital humanities, information science, digital scholarship, big data.

Comment constituer, communiquer et connaître le patrimoine aujourd'hui ? Comment renouveler le dialogue entre institutions patrimoniales et milieu de la recherche ? Comment adapter les pratiques et l'offre de services d'une bibliothèque aux besoins et aux usages de l'*homo numericus* et à l'infosphère qui caractérise le XXI^{e} siècle ?

À l'heure d'une progressive mise en données, voire en réseau, du monde et d'une complémentarité croissante entre humain et systèmes intelligents, une institution patrimoniale telle que la Bibliothèque nationale de France (BnF) est confrontée à de nombreuses évolutions, à plusieurs étapes de son activité allant de la collecte, de la description, du classement, du stockage, de la conservation et du signalement de ses collections numériques à la recherche, l'analyse et la communication de l'information.

Le présent article propose d'abord une définition préliminaire, au prisme d'une bibliothèque, des éléments fondamentaux convoqués dans le corps de l'article — sciences de l'information, intelligence artificielle et humanités numériques — pour ensuite se resserrer sur trois exemples de projets de constitution et d'analyse de corpus numériques conduits à la BnF. Ces exemples viennent illustrer les objectifs du programme Corpus, qui vise à construire une offre de services pour les chercheurs autour des collections numériques de la Bibliothèque. En conclusion seront abordées les possibilités qui découlent d'une collaboration renouvelée entre bibliothèques et milieu académique ainsi que les perspectives ouvertes par les expérimentations menées dans le cadre du programme de recherche Corpus notamment en matière d'intelligence artificielle.

1. La Bibliothèque face au numérique : périmètre et définitions

1.1 Une approche interdisciplinaire pour l'analyse et l'étude de corpus numériques

« Sciences de l'information », « humanités numériques » et « intelligence artificielle » sont trois concepts qui se sont développés à partir de la deuxième moitié du XX^{e} siècle. Le terme « intelligence artificielle » a été forgé en 1956 par l'américain John McCarthy à l'occasion de sa demande de subvention au NSF (National Science Foundation) pour l'école d'été au Dartmouth College. L'expression « sciences de l'information » (*information science*) est utilisée aussi à partir du milieu des années 1950 aux États-Unis. Malgré cette proximité chronologique, ces deux champs sont cependant restés longtemps disjoints.

Les sciences de l'information, qui ont pour objet d'étude l'information dans ses dimensions de production, de gestion, d'utilisation et de communication, fournissent à la Bibliothèque des méthodes et des techniques pour l'organisation et l'administration des données. La bibliothéconomie, pilier de l'expertise de toute bibliothèque, en est une application concrète. Plus précisément, elle mobilise la

recherche d'information, qui désigne les méthodes et techniques employées afin de retrouver de l'information dans un ensemble de documents ou de données, ainsi que la structuration et la description de l'information à travers l'élaboration et l'implémentation de modèles de données et de métadonnées. Les bibliothèques ont ainsi élaboré un certain nombre de standards, de formats, de protocoles d'accès appropriés à la gestion de leur domaine spécifique. En termes informatiques, ces éléments se sont traduits par des infrastructures et interfaces de collecte, de conservation et d'accès aux collections. Le numérique a eu pour effet de rendre encore plus prégnante cette omniprésence des technologies de l'information et de la communication dans les métiers des bibliothèques, mais dans un premier temps, essentiellement pour faciliter des usages qui restaient relativement inchangés : accès aux documents de façon unitaire, consultation sur place ou à distance, et dissémination des résultats de la recherche essentiellement à travers des publications dans lesquelles la bibliothèque n'était pas impliquée.

Les *humanities computing* émergent dès la deuxième moitié du XX^{e} siècle, grâce aux travaux de pionniers comme Josephine Miles et Roberto Busa. Avec l'irruption du web, la notion de *digital humanities* apparaît et se popularise progressivement à partir du milieu des années 2000. En tant qu'ensemble de pratiques de recherche en sciences humaines et sociales, arts et lettres « mobilisant les outils et les perspectives singulières du champ du numérique » (cf THATCamp Paris, 2010), les humanités numériques ont profondément fait évoluer les usages de recherche portant sur les collections patrimoniales, s'appuyant notamment sur la disponibilité de collections numériques massives que les bibliothèques s'étaient organisées pour collecter et produire depuis plus de dix ans. Les humanités numériques abordent des questions qui auparavant relevaient au sens strict des sciences de l'information, telles que la description, la gestion et l'analyse d'objets numériques, ainsi que de nouvelles modalités de communication, médiation et valorisation des collections patrimoniales et des recherches dont elles font l'objet. Mais de façon certainement encore plus importante, elles apportent de nouvelles questions scientifiques, liées directement au potentiel offert par l'outil informatique d'analyse massive, quantitative, des collections numériques.

Dans ce contexte, l'intelligence artificielle, en tant que discipline informatique qui vise à élaborer des machines ou des outils simulant les fonctions cognitives, apporte de plus en plus à la Bibliothèque des possibilités techniques aussi bien pour l'automatisation des traitements documentaires que pour offrir aux chercheurs en humanités de nouvelles modalités d'exploration, d'analyse et de gestion des collections ou ensembles cohérents de données massives. L'apprentissage automatique, à travers le développement et l'implémentation de méthodes statistiques et algorithmiques permettant à un ordinateur d'apprendre à réaliser des tâches, présente des cas d'usage essentiels pour les bibliothèques d'une part, pour les chercheurs en humanités d'autre part. Trois champs sont ainsi particulièrement concernés :

- les traitements d'analyse de l'image conduisant à la création de contenu structuré et exploitable, notamment de contenu textuel (OCR ou reconnaissance optique de caractères, OLR ou reconnaissance automatique

de la mise en page, HCR ou reconnaissance automatique de l'écriture manuscrite, OMR ou reconnaissance automatique de l'écriture musicale…) ou permettant d'accéder à l'image par le contenu (reconnaissance de formes …) ;

- la fouille de données (*text and data mining*) permettant de faire émerger des tendances ou des motifs à partir de masses importantes de données, notamment en passant par une étape de visualisation de données ;
- les traitements sémantiques, qui permettent d'opérer des rapprochements automatisés entre des données similaires (alignements de données…) ou des documents similaires (*clustering*…), ou d'extraire des informations sémantiques à partir d'informations brutes (annotation de texte, d'image fixe ou animée…)

Dans le cadre de sa transition numérique initiée dès 1994, la BnF s'est fixée pour objectif d'explorer, en participant à des projets d'humanités numériques, les territoires de recouvrement entre la recherche, notamment en sciences humaines et sociales, le domaine de l'informatique, notamment pour ce qui concerne les techniques d'apprentissage automatique et profond, et l'expertise de la Bibliothèque notamment en matière de systématisation des traitements, structuration, normalisation et préservation des données. Ainsi, sciences de l'information, humanités numériques et intelligence artificielle se croisent, s'influencent, se mêlent, et participent de concert à des projets de recherche dont la Bibliothèque est partie prenante.

1.2 Un besoin croissant d'automatisation pour analyser et gérer le patrimoine numérique national

De plus en plus confrontée à des niveaux de volumétrie et de vélocité typiques des mégadonnées (*big data*), les collections numériques de la BnF, qui occupent aujourd'hui environ six pétaoctets, sont caractérisées par une variété considérable. Documents numérisés, tels que par exemple les livres et manuscrits consultables dans Gallica — la bibliothèque numérique de la BnF — ; documents nativement numériques comme les œuvres d'art vidéo, les logiciels, les bases de données, les archives de l'Internet ; métadonnées bibliographiques et données d'autorité décrivant les personnes, lieux, organisations, concepts… autant d'ensembles de données diverses en termes de structures, formats, qualité, contextes de production, fonctions et contenus. Ces ensembles ont des histoires différentes, issues des changements des supports et des multiples strates de pratiques documentaires accumulées au fil du temps. Leur hétérogénéité exige des traitements spécifiques et par conséquent des compétences et des méthodes particulières, aussi bien pour les conserver ou les communiquer que pour les analyser (cf Moiraghi, 2017). Cette hétérogénéité des données, qui découle de l'amplitude chronologique et de la vocation à l'encyclopédisme caractéristiques des bibliothèques nationales, s'ajoute à l'accroissement de la quantité des données en entrée et à l'accélération conséquente des temps de traitement. La tendance traditionnelle des bibliothèques à

la systématisation des procédures doit dès lors trouver son équilibre face à la spécificité des données mais aussi des questions scientifiques propres aux projets de recherche qui les exploitent.

Dans ce contexte de tension entre l'accroissement des volumétries des données et la réduction des temps d'analyse et de traitement, la BnF développe une expertise poussée dans le domaine de l'informatique documentaire et de la gestion de collections numériques. On peut ainsi citer la préservation numérique dans son système SPAR, l'archivage de l'Internet qui représente près d'un pétaoctet de données pour des milliards d'URL, ou encore la publication de ses métadonnées bibliographiques sur le web des données qui ouvre la porte à des alignements semi-automatiques avec d'autres jeux de données (cf Moiraghi, 2018). Cependant, innover dans le domaine des sciences de l'information requiert parfois de mobiliser des compétences nouvelles ou des champs inexplorés de la connaissance. Pour atteindre cet objectif, la BnF expérimente, souvent en partenariat avec des équipes de recherche, dans le cadre de projets aux échelles variées, des techniques issues de l'informatique et de plus en plus de l'intelligence artificielle, pour automatiser la gestion, la communication et l'analyse de son patrimoine numérique.

La génération automatique de contenu textuel à partir d'images numériques a été la première technologie issue de l'intelligence artificielle à rejoindre les dispositifs informatiques régulièrement employées par la BnF dans le cadre de la gestion de ses collections numériques. Quatre-vingt-neuf ans après la « machine à lire » de Gustav Tauschek et soixante et un an après l'encombrant et tentaculaire Perceptron[1], la Bibliothèque effectue une reconnaissance automatique de caractères (OCR, *Optical Character Recognition*) dans la majorité des documents imprimés qu'elle détient afin que les contenus puissent être recherchés et exploités dans le format texte. Dans ce domaine, elle ne se limite pas à l'état de l'art, mais mobilise des partenariats de recherche pour repousser les limites de la technique et obtenir des résultats toujours plus performants, comme dans le cadre du projet Europeana Newspapers qui portait sur l'extraction automatique de la mise en page (OLR, *Optical Layout Recognition*) et de la structure logique des documents (cf Moreux, 2016). Plus récemment, elle favorise également la recherche dans le champ de la reconnaissance automatique de l'écriture manuscrite (HWR, *Handwriting Recognition*) par exemple en soutenant le projet européen Himanis qui se propose de comprendre la réalité du gouvernement royal français à partir des registres de la chancellerie royale des XIV^{e} et XV^{e} siècles conservés aux Archives nationales et à la BnF.

D'autres applications de l'intelligence artificielle sont ensuite venues rejoindre la boîte à outils de la BnF pour exploiter ces contenus textuels : elle expérimente ainsi une indexation automatique des contenus textuels via la reconnaissance d'entités nommées (NER, *Named-Entity Recognition*) avec le moteur sémantique Exalead, utilisé dans sa bibliothèque numérique Gallica. Au-delà du texte, elle se donne pour objectif d'ici 2020 d'étudier la faisabilité de l'application de solutions

[1] La machine à lire de Tauschek (1929) et le Perceptron (1957) peuvent être considérées comme les ancêtres précurseurs de l'OCR.

d'apprentissage profond pour l'indexation d'images et de nouvelles interfaces pour la recherche et l'analyse de documents iconographiques (cf Moiraghi et Moreux, 2018). Enfin, pour que les contenus puissent être explorés et analysés non plus de manière unitaire (par document) mais de manière globale (par corpus) via des outils numériques et avec des méthodes relevant notamment du *data mining* (fouille de données), elle est en train de construire une offre de services autour de ses collections numériques.

La mise au point de ces outils informatiques de plus en plus intelligents pour analyser les collections numériques ouvre des opportunités inédites pour la recherche, notamment en sciences humaines et sociales. La BnF a pu constater depuis 2015 une augmentation du nombre de projets de recherche portant sur des corpus numériques et impliquant non seulement l'expertise des chercheurs dans leur domaine scientifique mais aussi la mobilisation de compétences en sciences de l'information et en informatique, y compris en matière d'intelligence artificielle.

La partie suivante illustre via trois projets de recherche menés à la BnF comment des questionnements scientifiques issus des humanités numériques peuvent déboucher sur l'expérimentation d'outils relevant du champ de l'intelligence artificielle pour explorer et traiter les données numériques de la Bibliothèque, et les perspectives que ces expérimentations ouvrent pour l'évolution de son système d'information. Dans les trois cas, il s'agit de projets de recherche dont la Bibliothèque a été à l'initiative. Elle ne s'est pas limitée à jouer le rôle de commanditaire ou de fournisseur des données mais s'est autorisée, à titre d'expérimentation, à être partie prenante de la démarche de recherche. C'est ainsi au croisement de compétences diverses, mobilisées par la Bibliothèque autour de ses collections, qu'émergent les lignes de force de nouveaux usages numériques.

2. Trois exemples de projets d'humanités numériques conduits à la BnF

Les trois projets présentés brièvement ici, en assumant le point de vue de la Bibliothèque, portent chacun sur des ensembles de données différents : corpus sur la Grande Guerre extrait des archives de l'Internet, données (ou *logs*) de connexion à la bibliothèque numérique Gallica et ressources iconographiques issues de toutes les collections dans Gallica couvrant la période 1914-1918. Ils adoptent trois approches d'exploration, d'analyse et d'exploitation de corpus numériques, chacune reposant sur des méthodes et des techniques différentes en raison de la nature des données explorées et des finalités scientifiques propres à chaque projet. Ils partagent cependant la mobilisation, à différents niveaux, de compétences et méthodes en sciences de l'information, informatique et sciences humaines et sociales. Enfin, les deux premiers projets en particulier montrent que chaque acteur en autonomie n'aurait pas pu parvenir aux mêmes résultats, qui découlent, non sans difficultés, d'un travail collectif et interdisciplinaire.

2.1 Un projet fondateur : « Le devenir du patrimoine numérisé en ligne : l'exemple de la Grande Guerre »

Le projet « Le devenir du patrimoine numérisé en ligne : l'exemple de la Grande Guerre » a été lancé en 2013 dans le cadre du Labex « Les passés dans le présent » et porté par la BnF, le département de Sciences économiques et sociales de Télécom ParisTech et la Bibliothèque de documentation internationale contemporaine (BDIC). D'une durée de trois ans (2013-2016), son objectif était multiple : d'abord étudier « les pratiques sociales en ligne visant à construire une représentation du passé et à perpétuer la mémoire de la Grande Guerre » (cf Beaudouin et Pehlivan, 2017); mesurer l'impact des institutions patrimoniales dans la circulation et dans l'appropriation des documents massivement numérisés et mis en ligne ; puis, à partir des archives de l'Internet de la BnF, analyser de manière automatique le réseau des sites web français concernant la Grande Guerre et cartographier les liens entre ces sites internet. En parallèle, le projet avait vocation à développer des outils et à proposer des méthodes reproductibles pour analyser un corpus issu des archives de l'Internet de la BnF comme du web en général.

2.1.1 Une collection des archives de l'Internet de la BnF à l'origine de la deuxième phase du projet de recherche

L'étude des archives de l'Internet n'était pas au premier abord au cœur des préoccupations de l'équipe de recherche : c'est le besoin exprimé par les chercheurs de disposer d'un corpus web fiable, documenté, légal et permettant la reproductibilité des traitements qui a conduit à faire appel à cette collection patrimoniale constituée par la BnF. Il semblait en effet que le web « vivant »[2] ne permettait pas de définir un corpus présentant ces caractéristiques, et qu'il fallait travailler sur un web archivé. La BnF disposait depuis 2006 d'un cadre juridique, le dépôt légal, l'autorisant à reproduire et archiver des sites Internet et à les communiquer à un public de chercheurs accrédités. En outre, pour la Bibliothèque, le fait de travailler sur les archives de l'Internet présentait l'intérêt de fournir un terrain d'expérimentation autour d'une collection patrimoniale encore peu connue et peu exploitée et d'envisager le développement d'outils qui pourraient être réutilisés à terme dans d'autres projets. C'est ainsi que l'étude d'un corpus d'archives de l'Internet est devenue l'une des étapes fondamentales du projet et a demandé le recrutement d'une ingénieure informatique. Le travail réalisé par cette dernière, conjointement avec la sociologue qui pilotait la partie scientifique du projet, a porté d'une part sur la création d'un graphe de visualisation des liens entre les sites web constituant le corpus, et d'autre part sur l'extraction des données (fouille) d'un forum en ligne, le forum 14-18, afin d'analyser les méthodes utilisées par les amateurs pour identifier et partager des contenus culturels numérisés. À la BnF, c'était la première fois que de tels outils informatiques étaient utilisés pour appréhender le contenu d'un corpus d'archives de l'Internet.

---

[2] Des outils comme Hyphe ou Webrecorder sont souvent utilisés par les chercheurs en sciences sociales pour constituer un corpus à partir du web vivant.

2.1.2 Le dialogue de multiples compétences au cœur du processus de constitution et d'analyse du corpus numérique

Le rapport de Beaudouin et Pehlivan (2017) détaille les défis particuliers posés par ce choix. En effet, les archives de l'Internet, entrées dans le champ du dépôt légal en 2006, font à la BnF l'objet de collectes selon deux modalités : la première porte annuellement sur un très grand nombre de sites internet (4,5 millions de domaines en 2017) identifiés à partir des listes de bureaux d'enregistrement, et la deuxième consiste en des collectes ciblées, plus fréquentes et/ou plus profondes, d'un nombre plus restreint de sites internet (environ 20 000), sélectionnés par des bibliothécaires ou des partenaires en fonction de plusieurs thématiques. La collecte « Grande Guerre » faisait partie depuis novembre 2013 de cette seconde catégorie. Elle s'est enrichie au fil de la durée du projet, à travers les sélections effectuées par la BnF et ses partenaires, occasionnant une variation importante de couverture du corpus entre le début du projet et les derniers mois de l'analyse. En outre, les modalités de la collecte, qui repose sur des robots, a débouché sur un effet de « bruit » important, nuisant à l'interprétation des visualisations de données.

Ces difficultés ont entravé le processus idéal qui aurait dû, en théorie, fonctionner comme une chaîne dans laquelle les experts d'une thématique opèreraient d'abord la sélection et la description des sources, les experts techniques procèderaient ensuite à la collecte et à l'archivage des données, et enfin le corpus serait fourni aux équipes de recherche pour qu'il puisse être étudié et analysé. En réalité, un dialogue constant s'est établi tout au long de la recherche, occasionnant de multiples itérations entre experts des collections, experts des formats, informaticiens et équipes de recherche. Les interprétations des graphes générés par les traitements ont évolué avec la compréhension progressive des modalités de collecte et d'archivage mises en place par la BnF, et la mise en œuvre de solutions adaptées pour corriger les biais inhérents au matériau source. En tant que partie prenante du projet, la Bibliothèque n'a donc pas seulement contribué à l'identification des sources (collecte « Grande Guerre ») et à la délimitation du corpus pour répondre aux questions scientifiques du projet : elle a aussi fourni son expertise autour des formats de fichiers (ARC/WARC[3], DAT/WAT[4]) ; elle a mis à disposition ou employé ses outils, comme la base de données « Bcweb » (BnF Collecte du Web), ainsi que ses procédures (*crawl logs*[5]) pour constituer, nettoyer et enrichir le corpus.

2.1.3 Les conclusions et les résultats d'un travail de recherche collaboratif

En conclusion, ce projet, en plus de déboucher sur la création d'outils et l'élaboration de méthodes pour l'analyse de corpus issus des archives de l'Internet, a contribué à démontrer l'intérêt de travailler conjointement entre bibliothécaires, informaticiens et chercheurs sur ces nouveaux objets afin d'en fonder l'approche épistémologique. Il a également montré qu'une approche linéaire et dissociée n'était

[3] http://www.bnf.fr/fr/professionnels/dlweb_boite_outils/a.dlweb_formats_fichiers.html
[4] https://webarchive.jira.com/wiki/display/ARS/WAT+Overview+and+Technical+Details
[5] Fichiers qui contiennent les traces de l'activité des robots de collecte pendant le processus de *crawl* ou capture des sites internet.

pas suffisante, et que le succès de projet en humanités numériques portant sur des collections numériques patrimoniales massives et complexes requérait une organisation adéquate avec des mesures itératives afin de produire une analyse fiable. L'ouvrage « Le web français de la Grande Guerre. Réseaux amateurs et institutionnels » (cf Beaudouin, Chevallier, Maurel, 2018) retrace cette démarche pluridisciplinaire et synthétise les conclusions qu'il a été possible de tirer de ce travail de recherche collaboratif. En croisant démarches quantitatives et qualitatives, sociologie et sciences des données numériques, ce projet a permis d'éclairer la manière dont les sources documentaires numérisées circulent et dont les réseaux s'organisent sur le web à partir ou autour de ces sources. Il a montré l'apport des espaces amateurs de discussion sur le web en tant que vecteurs de valorisation de recherches individuelles, mais aussi d'acquisition de compétences et d'élaboration d'une conscience collective et de nouvelles connaissances. Du côté de la Bibliothèque, le projet a initié la création d'outils qui ont ensuite servi d'autres projets et fait évoluer globalement l'approche documentaire de ces collections (voir chapitre 4).

2.2 Une approche technique basée sur l'intelligence artificielle : l'analyse des traces d'usage de Gallica

Alors que le projet « Le devenir du patrimoine numérisé en ligne : l'exemple de la Grande Guerre » s'était emparé des collections numériques de la BnF, à travers les archives de l'Internet, parce qu'elles présentaient une opportunité pour répondre à la question scientifique posée par le projet, ce deuxième exemple visait spécifiquement à expérimenter des méthodes informatiques issues de la fouille de données et de l'intelligence artificielle, en complément d'autres méthodes visant également à appréhender les usages de Gallica (des entretiens, un questionnaire en ligne administré à plus de 7000 gallicanautes, un dispositif d'observation vidéo ethnographique). L'idée était d'évaluer l'apport des méthodes automatisées pour les études d'usage des bibliothèques numériques en s'appuyant sur l'emploi d'un type de données particulier : les « logs » ou traces d'usage. L'ensemble du dispositif scientifique, mis en œuvre en 2016 dans le cadre du Bibli-Lab[6], partenariat de recherche entre la BnF et l'école d'ingénieurs Télécom ParisTech, constituait une vaste étude des usages de Gallica, aux multiples facettes complémentaires, dont les résultats ont été présentés le 3 mai 2017 lors de la journée d'étude « Quels usages aujourd'hui des bibliothèques numériques ? Enseignements et perspectives à partir de Gallica » (cf Bibliothèque nationale de France, 2017).

---

[6] Bibli-Lab est un partenariat de recherche initié en 2013 entre la BnF et l'école Télécom ParisTech. Il vise à étudier les usages en ligne du patrimoine numérique des bibliothèques, URL : http://www.bnf.fr/fr/la_bnf/pro_publics_sur_place_et_distance/a.bibli-lab.html ; http://actions-recherche.bnf.fr/BnF/anirw3.nsf/IX01/A2017000006_bibli-lab-laboratoire-d-etude-des-usages-du-patrimoine-numerique-des-bibliotheques

2.2.1 La BnF en dialogue avec d'autres acteurs et compétences pour étudier les comportements de ses publics

Le volet de l'étude intitulé « Analyse des traces d'usage de Gallica » proposait une approche inédite d'analyse des parcours-types d'usagers de Gallica, la bibliothèque numérique de la BnF, fondée sur des méthodes d'apprentissage automatique (*machine learning*). À partir des fichiers de connexion aux serveurs de Gallica, l'objectif principal du projet consistait à identifier des sessions-types, c'est-à-dire des parcours similaires en termes d'enchaînement d'actions et de consultation de documents de la bibliothèque numérique. Le projet, d'une durée de quinze mois, a été conduit par Adrien Nouvellet, chercheur en traitement du signal en contrat postdoctoral à l'école Télécom ParisTech, qui était encadré par deux enseignants-chercheurs en sciences économiques et sociales (Valérie Beaudouin et Christophe Prieur) et deux enseignants-chercheurs en traitement du signal et des images (Florence D'Alché-Buc et François Roueff) de la même école. La transversalité du projet, qui faisait se rencontrer deux équipes distinctes de Télécom ParisTech, ainsi que l'implication d'un chercheur informaticien extérieur au domaine culturel, faisaient partie des aspects intéressants du projet d'un point de vue méthodologique.

2.2.2 L'avancement collaboratif et itératif pour la préparation et le traitement des données

Afin de découvrir des tendances dans l'utilisation de Gallica via l'application de méthodes de type fouille de données (*data mining*), on a choisi d'exploiter les données de connexion aux serveurs de la bibliothèque numérique. Ces données sont appelées communément « logs de connexion » et contiennent les requêtes effectuées depuis une adresse IP (qui identifie généralement un utilisateur). Un exemple de ligne de logs de connexion à Gallica est proposé ci-dessous.

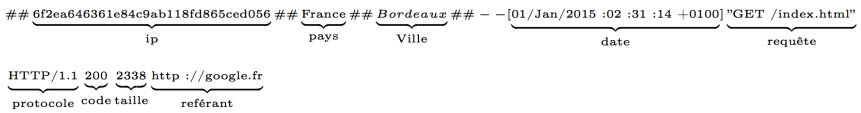


Figure 1. Exemple de ligne de logs de connexion à Gallica, extrait de Nouvellet et. al. *2017.*

Outre les requêtes HTML, les requêtes SRU[7] et les identifiants ARK[8] sont aussi présents dans les logs de connexion. Ces deux types d'information ont permis

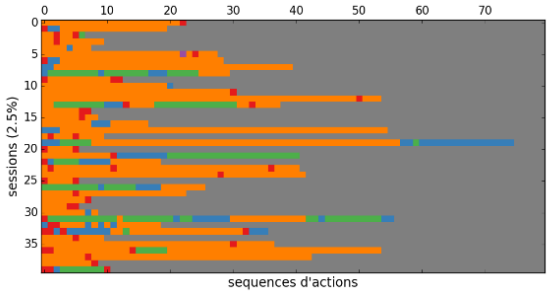
[7] SRU : Search/Retrieve via URL est un protocole de type REST (Representational State Transfer) utilisé pour formuler des recherches notamment dans le contexte des données de bibliothèque et obtenir des résultats. Il s'agit d'un standard reconnu par le consortium OASIS et maintenu par la Bibliothèque du Congrès aux États-Unis.

[8] ARK : Archival Resource Key est un standard d'identification utilisé par la BnF et maintenu par la California Digital Library.

respectivement d'identifier les différentes actions d'un usager sur le site Internet et d'identifier les ressources consultées.

Ces logs ont été nettoyés et enrichis avec d'autres données en fonction des objectifs de la recherche, étape cruciale et préliminaire à tout processus d'analyse dont dépendent les objectifs scientifiques ; d'autant plus dans le cas de données comme les logs de connexion qui n'ont pas été conçus pour l'étude et l'analyse de comportements d'usagers. À plusieurs reprises, le DSI de la BnF (Département des Systèmes d'Information) a modifié ou complété les données en fonction des besoins de l'étude : anonymisation, application de correctifs, ajout de champs nécessaires pour la recherche... Les logs ont ensuite été liés aux métadonnées des documents consultés grâce au lien entre l'identifiant unique ARK, qui identifie un document, et la notice bibliographique du document correspondant collectée grâce au protocole OAI-PMH[9] : la normalisation des données bibliographiques et l'utilisation de l'identifiant unique ARK pour chaque document ont permis une utilisation et un enrichissement rapide des données ce qui contribue à démontrer l'intérêt des données structurées et de qualité dans une perspective de recherche. Enfin, autre type de données exploité dans le cadre du projet : l'ensemble des liens vers Gallica extraits des contenus du blogue et de la page Facebook consacrée à la bibliothèque numérique. Ces données ont fait l'objet d'une analyse dans la dernière phase du projet pour déterminer l'impact des activités de médiation sur la consultation de documents dans Gallica.

Afin d'atteindre l'objectif principal consistant à identifier des similitudes ou des tendances parmi les sessions et donc les comportements des usagers, un algorithme de classification non supervisée (*clustering*) fondé sur un mélange de modèles de Markov a été utilisé. Les descriptions du modèle de Markov et l'algorithme employé sont présentés en détail dans le chapitre 3.1 du rapport du projet (cf Nouvellet *et al.*, 2017). Ce traitement a permis, à travers des visualisations de données, de représenter visuellement des types de comportements récurrents, regroupés en clusters, ce qui a servi de support à l'analyse sociologique des usages.



[9] OAI-PMH : Open Archive Initiative Protocol for Metadata Harvesting. Protocole standard utilisé dans le domaine culturel et scientifique pour collecter et mettre à disposition de manière asynchrone des métadonnées issues de plusieurs silos.

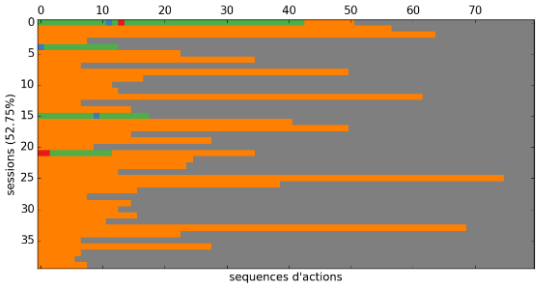


Figure 2. Exemple de visualisation des clusters représentant les parcours des usagers, extrait de Nouvellet et al. *2017*

2.2.3 Les conclusions et les résultats d'une « lecture distante »

Comme dans l'exemple du projet Grande Guerre, les corpus analysés étaient constitués de données nativement numériques massives, qui ne pourraient pas être appréhendées seulement par l'œil ou le cerveau humain. L'opportunité que ces corpus offrent à la recherche ne réside pas, dans la plupart des cas, dans une lecture unitaire basée sur le document mais dans une lecture distante et globale de l'ensemble. En ce sens, cette lecture augmentée, cette hyper-lecture de données liées, en constituant la prothèse d'un œil augmenté, d'un hyper-œil qui s'ajoute à l'œil organique, augmente aussi la faculté de la recherche à extraire des connaissances. Cette approche technique pour l'étude des usages de Gallica a permis de confirmer le caractère siloté des usages de Gallica, en démontrant que la majorité des sessions d'utilisation de la bibliothèque numérique portait sur un seul document et que les utilisateurs qui ouvraient plus de cinq documents consultaient seulement des documents d'un ou deux types. Concernant la valorisation sur les réseaux sociaux, l'analyse d'audience des publications sur la page Facebook a montré qu'un lien illustré d'une vignette engendrait vingt-cinq fois plus de visites qu'un simple lien textuel : un constat qui a immédiatement conduit la BnF à modifier sa stratégie de communication et médiation autour de Gallica.

Cependant, de tels résultats ne parviennent à faire sens qu'au prix d'une mise en œuvre itérative de ces méthodes, au cours de laquelle se confrontent les différentes compétences réunies par le projet, qu'il s'agisse des sciences de l'information avec les normes et formats utilisés par la BnF, de l'apprentissage automatique avec les opérations de traitement, ou des sciences humaines et sociales lorsqu'il s'agit de formuler les hypothèses et d'interpréter les résultats obtenus et les visualisations de données. Comme dans le projet précédent, ce processus d'interaction entre un trinôme de compétences n'était pas linéaire mais faisait l'objet d'itérations successives tout au long du projet. Le résultat de l'analyse, ensuite confronté aux autres méthodes d'études en jeu dans le dispositif plus complet incluant

questionnaire, entretiens et vidéo-ethnographie, est le fruit de cette approche transversale au croisement des trois champs qui font l'objet du présent article.

2.3 Une expérimentation mettant l'apprentissage profond à l'épreuve : le moteur de recherche iconographique GallicaPix

À la différence des deux projets précédemment présentés, « Gallica.Pix » n'est pas un projet de recherche proprement dit, mais un démonstrateur qui a été développé pour proposer de nouvelles méthodes de recherche dans les documents iconographiques présents dans les collections de la BnF. Cette preuve de concept (PoC, *Proof of Concept*) a été réalisée en 2017 par Jean-Philippe Moreux, actuellement expert scientifique de Gallica avec la collaboration de Guillaume Chiron (L3i, université de la Rochelle). Elle met en œuvre une approche d'indexation sémantique sur un corpus de ressources iconographiques de Gallica contemporaines de la Première Guerre mondiale.

2.3.1 À l'origine du projet, un chercheur autonome aux compétences variées

Toutefois, ce prototype ne constitue pas, comme les deux projets précédents, un projet de recherche institutionnel dans lequel la Bibliothèque est partie prenante : conduit de manière autonome par un chercheur alors en poste à la Bibliothèque, il démontre la faisabilité de construire des projets d'exploitation des données de la BnF en utilisant ses API (*Application Programming Interface*), sans que celle-ci n'intervienne dans la réalisation ou ne produise des outils *ad hoc*. Il aurait aussi bien pu être réalisé par un chercheur en dehors de la BnF, avec les mêmes données et les mêmes outils. En cela, « GallicaPix » est aussi un démonstrateur de ce que les chercheurs pourraient construire avec les données de la BnF sans forcément mettre en place de partenariat : c'est pourquoi il est intégré dans Gallica Studio[10] et utilisé ici comme exemple de perspective offert par le projet Corpus.

2.3.2 Les trois phases de réalisation : de l'extraction des données à l'élaboration de l'interface de recherche

La réalisation de ce démonstrateur a impliqué essentiellement trois phases de recherche et d'expérimentation : la première consistant à repérer et à extraire les illustrations à l'aide des API mises en place par la BnF ; la deuxième visant à les enrichir avec des métadonnées permettant leur recherche ; la dernière portant sur l'élaboration d'une interface de recherche web interrogeant une base de données XML.

La première phase de la réalisation a montré encore une fois l'importance d'avoir des métadonnées bibliographiques complètes et des vocabulaires normalisés, et le défi que représente la gestion de l'hétérogénéité lors de l'extraction des données à partir de documents, comme par exemple une illustration publicitaire dans un journal. Dans le cadre de la deuxième phase du développement, celle de

[10] http://gallicastudio.bnf.fr

l'enrichissement des métadonnées, plusieurs méthodes et techniques ont été expérimentées sur le corpus d'images extrait des serveurs de Gallica via l'API Image IIIF[11] (*International Image Interoperability Framework*) : Inception-V3[12], réseau de neurones artificiels convolutionnels et *open-source* de la société Google qui a été réentrainé sur les genres iconographiques présents dans le corpus (photographie, gravure, carte, dessin de presse, etc.) ; la bibliothèque *open source* OpenCV/dnn[13] et les API IBM Watson Visual Recognition[14] et Google Cloud Vision[15] pour l'indexation sémantique.

Parmi les principaux obstacles techniques constatés dans la tâche de classification des genres, on pourra citer : la confusion sur des genres visuellement proches comme photogravure-gravure et celle liée à l'identification des publicités illustrées de la presse quotidienne, qui relèvent d'un mode de communication et non d'une forme graphique visuellement homogène. L'entraînement des réseaux de neurones artificiels proposés par les offres commerciales sur des ressources iconographiques majoritairement contemporaines s'avère inadapté aux besoins des institutions patrimoniales, puisque cette approche montre ses limites même sur un corpus du XX^{e} siècle. Une collaboration étroite avec des équipes de recherche permettrait de traiter la spécificité des ressources iconographiques historiques. Enfin, la volumétrie importante des métadonnées générées pose d'autres défis en termes d'architecture technique, de puissance de calcul et d'espace de stockage ainsi qu'en termes de normalisation et d'interopérabilité des métadonnées de classification générées par les API de reconnaissance visuelle. Malgré ces limites, l'application web « GallicaPix », en interrogeant tout à la fois les métadonnées bibliographiques, les métadonnées de reconnaissance visuelle et l'OCR des documents et en mobilisant des techniques d'intelligence artificielle, permet de satisfaire de nombreux cas d'usage en matière de ressources iconographiques.

2.3.3 L'apport d'un prototype pour explorer de nouveaux usages

Cette expérimentation démontre la maturité croissante des techniques à base d'intelligence artificielle pour le traitement d'images, notamment contemporaines. Elle ouvre de nouvelles perspectives comme la création de jeux de données iconographiques à destination du chercheur ou de l'ingénieur dans le cas de projets impliquant de l'apprentissage profond, et elle confirme l'intérêt de l'utilisation de protocoles standards tels que IIIF. Ce projet d'application web mobilise à nouveau des compétences de différente nature : en utilisant des réseaux de neurones artificiels pour la reconnaissance et l'indexation automatique de formes et d'images, il s'appuie sur des méthodes et des techniques relevant de l'intelligence artificielle ; en

[11] https://iiif.io
[12] https://www.tensorflow.org/tutorials/images/image_recognition
[13] https://docs.opencv.org/3.3.0/d2/d58/tutorial_table_of_content_dnn.html
[14] https://www.ibm.com/watson/services/visual-recognition
[15] https://cloud.google.com/vision/

employant des protocoles standards et des outils, comme les protocoles SRU[16] et IIIF développés dans le milieu des bibliothèques, il s'appuie également sur des méthodes et des techniques relevant des sciences de l'information ; en élaborant une interface de consultation permettant une expérience de recherche améliorée et enrichie, il propose une plateforme relevant des humanités numériques, qui pourrait donner lieu à la formulation de nouvelles questions scientifiques. Cependant, pour que l'efficacité du prototype soit testée et perfectionnée, l'intervention d'experts du contenu des collections et de chercheurs en sciences humaines et sociales semble indispensable. L'élaboration d'une plateforme technique de ce type peut être vue comme un outil au service des chercheurs et non comme une fin en soi ; elle ne peut révéler son utilité et sa pertinence qu'au prisme de compétences scientifiques éprouvées.

3. Le programme de recherche Corpus porté par la BnF

Les trois projets ci-dessus illustrent l'interaction entre les trois domaines et les trois compétences que sont les sciences de l'information, l'informatique (dont l'intelligence artificielle) et les sciences humaines et sociales. Si les motivations pour travailler sur les collections numériques de la BnF et les modalités de cette collaboration diffèrent, il n'en reste pas moins qu'aucun de ces trois projets n'aurait pu être mené à son terme sans cette conjonction de compétences spécifiques. Au croisement de trois domaines dotés chacun de leurs apports, les humanités numériques ouvrent ainsi un champ des possibles pour la Bibliothèque : celui de l'exploration de ses collections au moyen de techniques nouvelles, permettant la formulation de questions scientifiques originales et l'émergence de nouvelles connaissances.

Ils montrent également que l'existence de données numériques disponibles en masse et d'outils aptes à les traiter constitue une opportunité pour le développement de nouvelles recherches que la BnF a vu émerger depuis plusieurs années et auxquelles elle a participé. Plusieurs motivations distinctes ont pu conduire la Bibliothèque à s'y engager : tantôt l'espoir d'améliorer ses propres outils de production, de gestion et d'accès, tantôt le souhait d'augmenter la visibilité et l'étude de ses collections. Confrontée à ce contexte non pas de révolution mais plutôt d'évolution notable pour sa rapidité, la Bibliothèque a décidé d'envisager la construction d'une nouvelle offre de services aux chercheurs, en adoptant une approche par projets expérimentaux pour ensuite opérer une généralisation des processus.

Initié en 2016 et inscrit dans le cadre du plan quadriennal de la recherche de la BnF pour la période 2016-2019, le projet Corpus s'est donné pour objectif de construire un service de fourniture de corpus permettant la fouille de textes et de données à destination de la recherche. En procédant de manière expérimentale, itérative, collaborative et transversale, ce programme de recherche de quatre ans focalise son

---

[16] http://www.bnf.fr/fr/professionnels/recuperation_donnees_bnf_boite_outils/a.service_SRU.html

attention sur des corpus issus de trois principaux ensembles cohérents de données numériques : les archives de l'Internet, les documents numérisés et les métadonnées.

3.1. Une nouvelle dimension pour l'accès aux archives de l'Internet

En parfaite continuité avec les expérimentations menées pour le projet « Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre », un autre projet de recherche portant sur les archives de l'Internet a été conduit dans le cadre de ce programme. En 2016, le partenariat avec l'équipe de l'Institut des sciences de la communication du CNRS en charge du projet ANR Web90 a porté sur l'élaboration d'une application expérimentale nommée « Archives du web Labs » ainsi que sur l'indexation en plein texte de deux corpus : les « incunables du web » (1996-2000) et la collecte « attentats » de 2015. Cette application représente la continuation du travail entrepris pour la réalisation de l'interface pour le projet sur la Grande Guerre car elle offre, en plus de l'extraction des métadonnées des corpus, la possibilité de rechercher tous les mots présents dans les pages du corpus ainsi que plusieurs fonctions de personnalisation telles que l'enregistrement de requêtes et l'export de résultats de traitements.

La mise en place de la fonctionnalité de recherche plein texte constituait un défi technique pour la BnF qui n'avait pas les moyens de la déployer à l'échelle de l'ensemble des collections des archives de l'Internet. L'approche par corpus, grâce à la collaboration étroite avec les équipes de recherche qui s'intéressent à ces données, est donc une manière de lever cet obstacle. Elle permet d'envisager le passage à l'échelle et la proposition aux lecteurs de la BnF de services à valeur ajoutée : en 2018, l'essentiel des fonctionnalités de l'application « Archives web Labs » ont fait l'objet d'un déploiement sur tous les postes d'accès aux ressources numériques de la bibliothèque de recherche, alors qu'auparavant elle n'était offerte qu'à l'équipe de recherche partenaire, sur un seul poste informatique dédié[17]. Un nouveau corpus, la collecte « Actualités » (2010-2017), a également été ajouté et indexé en plein texte pour servir les besoins d'un autre projet de recherche, le projet « Neonaute »[18].

---

[17] Certaines fonctionnalités comme la possibilité d'accéder aux métadonnées dans la ont été retirées dans la deuxième version qui est accessible au public, l'accès à ces métadonnées étant limité aux chercheurs sous convention. Ce cadre contractuel fixent des conditions pour respecter les stipulations du Code du patrimoine, mais également la législation sur la propriété intellectuelle et la protection des données.

[18] UMR 7030 – LIPN - Université Paris 13 Sorbonne Paris Cité. Le projet « Neonaute » a été retenu dans le cadre de l'Appel à Projets de la Délégation Générale à la Langue Française et aux Langues de Frances (DGLFLF), Langue et numérique 2017. Il porte sur la réalisation d'un moteur de recherche et d'études terminologiques s'appuyant sur le corpus « Actualités » issu du dépôt légal du web de la BnF.

3.2. Des ateliers pour explorer de nouvelles dimensions des humanités numériques et de l'intelligence artificielle

En 2017, dans le cadre de la deuxième année du programme Corpus, le projet « Giranium », conduit par une équipe du GRIPIC du CELSA (laboratoire de sciences de l'information et de la communication de Sorbonne Université), a permis d'inclure dans la réflexion l'exploitation numérique de corpus numérisés. En lien avec les recherches menées par Jean-Philippe Moreux (2016) autour des approches innovantes pour l'étude de la presse ancienne numérisée, le projet « Giranium » visait à mieux comprendre l'apparition des premières industries culturelles et médiatiques en France à travers le prisme d'Émile de Girardin, personnalité emblématique du journalisme français du XIX^{e} siècle, tout en mettant en œuvre des pratiques relevant des humanités numériques. Outre la numérisation d'un corpus de presse du XIX^{e} siècle et son océrisation, le projet a demandé à la Bibliothèque d'explorer d'autres aspects liés à une potentielle offre de services autour des collections numériques tels que le besoin d'espaces de travail dédiés dans les espaces physiques de la bibliothèque (pour le travail en groupe) et d'ateliers méthodologiques sur les humanités numériques (autour notamment des formats, des standards, des pratiques de structuration, normalisation, pérennisation et liage des informations).

Un de ces ateliers, intitulé « Explorer des corpus d'images. L'IA au service du patrimoine » (cf Moiraghi, Moreux, 2018), a été l'occasion d'inclure de manière très explicite l'intelligence artificielle dans la réflexion menée dans le cadre du programme Corpus. Neuf projets d'humanités numériques impliquant à différents niveaux la reconnaissance automatique d'écritures ou bien la reconnaissance automatique d'image par le contenu étaient présentés au fil de cet après-midi dans une logique de partage d'expérience entre institutions patrimoniales et milieu académique. Pour la Bibliothèque, prendre connaissance, étudier et expérimenter ces techniques de reconnaissance automatique de textes ou d'images, impliquant notamment l'utilisation de réseaux de neurones, constitue une nouvelle opportunité et un enjeu considérable pour réduire les temps de traitement des collections et améliorer le travail de recherche. Pour les équipes de recherche, les collections de la BnF constituent un terrain idéal pour éprouver l'efficacité des outils et mesurer la maturité des technologies sur des matériaux historiques.

Dans le cadre de la troisième année du projet, d'autres expérimentations ont été menées en lien notamment avec l'exploration et la réutilisation des métadonnées bibliographiques que la Bibliothèque collecte ou crée dans le cadre de son activité. Ces données, qui ont été placées sous licence ouverte de l'État en 2014, sont essentielles pour la gestion et la recherche de l'information mais peuvent constituer aussi un terrain d'enquête en elles-mêmes pour des projets de recherche. Elles peuvent être questionnées par exemple pour des études démographiques sur les auteurs (cf Langlais, 2017).

Comme dans le cadre de la deuxième année, une équipe de recherche a contribué à l'avancement du programme. Le projet ANR « Foucault Fiches de Lecture » a pour objectif de numériser, mettre en ligne, indexer, décrire et enrichir les notes de lecture

manuscrites de Michel Foucault, en utilisant une plate-forme numérique de travail collaboratif. Cette plateforme donne accès aux fiches de lecture numérisées, permet l'enrichissement des métadonnées par un système de *mashup* et d'alignement avec les données bibliographiques et biographiques de *data.bnf.fr* et fournit une transcription de chaque fiche. Cette transcription semi-automatique est obtenue à l'aide du logiciel Transkribus qui, basé sur une technologie d'intelligence artificielle, après une phase d'apprentissage via des réseaux neuronaux, permet la reconnaissance d'écritures manuscrites ainsi qu'une recherche par mots clés. Malgré la nécessité d'un travail minutieux ligne par ligne, l'équipe a constaté un taux moyen de réussite de reconnaissance de l'écriture de 92%, une fois l'entraînement effectué. Les échanges avec l'équipe, notamment lors de l'atelier « Penser, classer, modéliser. L'exemple du projet Foucault Fiches de Lecture » (Moiraghi et Ventresque, 2018) a confirmé l'efficacité croissante de ce type d'approche et contribue à préfigurer les enjeux qui découleront du projet Corpus en matière de reconnaissance automatique des écritures manuscrites.

3.3 Une volonté d'échange et de dialogue avec les publics potentiels

Parallèlement aux expérimentations menées en collaboration avec les chercheurs, une étude a été conduite la deuxième année du programme Corpus afin de mieux cerner les besoins des équipes de recherche notamment en termes d'espaces dédiés. Fondée sur une méthodologie mêlant une enquête qualitative par entretiens, des observations informelles effectuées lors de deux ateliers autour de thématiques liées aux humanités numériques et un atelier participatif utilisant la méthode UX (*User Experience design*) des personas, l'étude relève le besoin des équipes de recherche de disposer des collections numériques à distance et en mobilité mais explore également la valeur ajoutée potentielle d'un espace physique à la Bibliothèque consacré à l'étude et l'analyse de corpus numériques.

Outre la nécessité d'un tel espace pour la consultation et l'analyse de corpus sous droit (selon le Code du patrimoine, les documents sous droit issus du dépôt légal ne sont accessibles que dans les emprises physiques de l'établissement), la possibilité d'avoir un accès immédiat aux différentes expertises de la Bibliothèque 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, le modèle qui se profilerait dans cet espace autoriserait la formulation par les équipes de recherche de questions scientifiques et techniques aux agents de la Bibliothèque, et l'apport de ces derniers serait une expertise sur les fonds, sur les questions juridiques et sur les aspects techniques, notamment de formats et d'outils. Une infrastructure et des outils logiciels, notamment dédiés à la fouille de données, y seraient déployés. Ce modèle convoquerait donc les trois éléments — sciences de l'information, informatique ou intelligence artificielle et sciences humaines et sociales — précédemment mentionnés et illustrés par les trois exemples de projets de recherche.

Un autre facteur, pragmatique mais notable, identifié comme favorisant la fréquentation d'un espace physique à la BnF, est l'actuelle pénurie de locaux dans les universités parisiennes. Tel que défini dans le rapport de l'étude (cf Moiraghi,

2018), ce futur espace à la Bibliothèque se profile comme facile d'accès, convivial, capable d'abriter des formations, des événements, des présentations de travaux de recherche et capable d'évoluer au rythme de l'innovation et du progrès technologique.

À la suite de ce rapport, le programme Corpus avance sur plusieurs axes de recherche : la conception et la mise en place de l'offre de services autour des collections numériques sur place ; la continuation et l'amélioration des dispositifs dans le cadre de la politique de dissémination des données à distance et en ligne ; l'élaboration d'une infrastructure sécurisée permettant la constitution et l'analyse de corpus numériques ; l'articulation de ces deux offres complémentaires de services à la recherche (en ligne et sur place) ; la cartographie des compétences ; la systématisation des processus et des procédures ; l'élaboration d'une feuille de route autour de l'intelligence artificielle ; le positionnement institutionnel et stratégique dans l'écosystème de la recherche aussi bien français qu'international.

4. En conclusion : vers un lieu et un modèle de collaboration scientifique pour la connaissance du patrimoine numérique national

Traditionnellement une bibliothèque a pour vocation de collecter, préserver, décrire et communiquer les objets qui sont appelés ensemble à constituer un patrimoine. La Bibliothèque nationale de France, en raison du dépôt légal, n'opère aucun jugement de valeur, ni moral ni esthétique ni social, pour sélectionner les documents appelés à faire partie des collections nationales, au contraire d'une politique documentaire classique telle que la pratiquent les autres types de bibliothèques (universitaires ou publiques). La conséquence de cette spécificité est l'existence, dans les collections de la BnF, de masses considérables de documents introuvables ailleurs, qui reflètent l'esprit de leur époque et qui peuvent servir de source pour étudier la société qui les produit. Avec le numérique, ce potentiel d'étude et de connaissance est démultiplié car il devient possible d'appliquer à ces matériaux numériques des méthodes de lecture distante (cf Moretti, 2013).

Ces approches d'hyper-lecture, fondées sur l'automatisme et le quantitativisme, continuent de soulever un certain scepticisme, notamment dans les milieux académiques liés aux humanités, depuis les premières expériences d'histoire quantitative dans la deuxième moitié du XX^{e} siècle. Qu'elles mènent à l'élaboration d'interfaces, au développement et à l'amélioration d'algorithmes, à des statistiques ou à des visualisations, ces approches sont souvent accusées de parvenir à des évidences déjà connues ou de trop dépendre des biais présents dans les algorithmes ou dans les corpus convoqués pour les analyses. Avec la conscience de ces limites, la Bibliothèque a vu augmenter les demandes d'accès à des corpus numériques depuis 2015 et a décidé de participer à des projets de recherche en ne considérant pas ces approches comme objets de recherche en tant que tels mais plutôt comme des moyens à mettre en œuvre, parmi d'autres, pour répondre à des problématiques scientifiques.

Grâce à ces multiples expériences de collaboration avec des équipes de recherche, la Bibliothèque a pu constater un triple intérêt pour le déploiement d'une activité autour des humanités numériques et, plus généralement, de tous les usages liés aux collections numériques. Tout d'abord, ce nouveau champ porte une promesse de renouvellement, voire de reconquête du public des chercheurs pour l'étude des collections et la connaissance du patrimoine. Par ailleurs, la mutualisation et la capitalisation sur des méthodes, des outils et des techniques notamment à base d'intelligence artificielle ouvrent des perspectives aussi bien pour la connaissance du patrimoine numérique que pour sa gestion par la Bibliothèque. Enfin, l'utilisation de méthodes issues de l'intelligence artificielle pour l'enrichissement de contenus numériques (comme par exemple l'OCR) ou pour la médiation et la valorisation du patrimoine permet d'imaginer un nouveau cadre de travail scientifique sur les collections, où les compétences des bibliothécaires, des informaticiens et des chercheurs s'associent et se complètent pour mieux faire connaître le patrimoine national.

En promouvant une complémentarité entre méthodes quantitatives et qualitatives, entre humains et systèmes intelligents, entre sciences de l'information, ingénierie informatique, intelligence artificielle et humanités numériques, l'offre de services qui résultera du programme de recherche Corpus se veut un lieu et un modèle de collaboration scientifique ancré dans la pluridisciplinarité. Ce nouveau modèle représente un moyen prometteur pour la Bibliothèque d'améliorer et d'assurer la constitution et la communication de son patrimoine numérique. Incarné dans un lieu physique, ce modèle permettra de renouveler le dialogue entre institutions patrimoniales et milieu de la recherche. Également virtuel, sous forme d'infrastructure sécurisée, il saura répondre aux besoins de mobilité, de dynamisme et de rapidité de l'*homo numericus*. Dans un contexte international de redéfinition des rapports entre bibliothèques et milieu de la recherche, de définition des enjeux et des questions éthiques autour des GAFA ainsi que des compétences dans le domaine de l'intelligence artificielle au sein des bibliothèques, cette offre de services se configure en faveur de l'ouverture des données, de la science ouverte, de la sociabilité scientifique et d'une économie du savoir fondée sur le partage et la coopération. Grâce au développement de procédures et d'outils intelligents, tels que des moteurs d'indexation automatique de contenus, pour la gestion de flux massifs de données, elle a vocation à perpétuer dans l'infosphère l'équilibre entre la vision microscopique de la recherche, qui réside dans la spécificité des questions scientifiques, et la vision macroscopique d'une bibliothèque à vocation encyclopédique et universelle.

Bibliographie

Beaudouin V. (2016). Forums en ligne : des espaces de co-production de la connaissance et du lien social, *L'ordinaire d'internet*, O. Martin & É. Dagiral (dir.), Paris, Armand Colin, p. 203-225.

Beaudouin V., Pehlivan Z. (2017). *Cartographie de la Grande Guerre sur le Web : Rapport final de la phase 2 du projet "Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre"*, https://hal.archives-ouvertes.fr/hal-01425600

Beaudouin V., Maurel L. (2017). La commémoration de la Grande Guerre sur le web : présence et diffusion du patrimoine numérisé, *Matériaux pour l'histoire de notre temps*, t. 1, n°120, p.10-17.

Beaudouin V., Chevallier P., Maurel L. (à paraître fin 2018). *Le web français de la Grande Guerre. Réseaux amateurs et institutionnels*, Presses Universitaires de Paris Nanterre.

Bermès E. (2017). *Préfiguration d'un service de fourniture de corpus numériques à destination de la recherche*, http://c.bnf.fr/fom

Bermès E. (2017). Text, data and link-mining in digital libraries: looking for the heritage gold, *IFLA Satellite Meeting - Digital Humanities – Opportunities and Risks: Connecting Libraries and Research 2017*, Berlin, Allemagne, https://hal.inria.fr/hal-01643293

Bermès E. (2018). Text, data and link-mining in digital libraries: looking for the heritage gold, *Library Science Talks 2018*, Zurich et Genève, Suisse, https://indico.cern.ch/event/714588/attachments/1617718/2585647/LSTalks-20180326-Bermes_EN_v3.pdf

Bermès E. (à paraître). Quand le dépôt légal devient numérique : épistémologie d'un nouvel objet patrimonial*, Quaderni*, https://journals.openedition.org/quaderni

Bibliothèque nationale de France (2017). *Contrat d'objectifs et de performance 2017-2021*, http://www.bnf.fr/documents/contrat_performance.pdf

Bibliothèque nationale de France (2017). *Il était une fois dans le web : 20 ans d'archives de l'internet en France*, http://c.bnf.fr/fse

Bibliothèque nationale de France (2017). *Quels usages aujourd'hui des bibliothèques numériques ? Enseignements et perspectives à partir de Gallica*, http://c.bnf.fr/fuZ

Bouchard A. (2017). *Présentation du projet CORPUS à la BnF*, https://webcorpora.hypotheses.org/119

Chevallier P. (2017). Web de la mémoire et mémoire du web, *Revue de la BnF*, n° 54, p. 179-193.

Illien G., Sanz P., Sepetjan S., Stirling P. (2011). La situation du dépôt légal de l'internet en France : retour sur cette nouvelle législation, sur sa mise en pratique depuis cinq ans, et perspectives pour le futur, *Actes du 77e congrès de la Fédération internationale des associations de bibliothécaires et d'institutions (IFLA)*, San Juan (Porto Rico), http://conference.ifla.org/past-wlic/2011/193-stirling-fr.pdf

Jacquot O. (2018). Stratégie de recherche de la Bibliothèque nationale de France. *Revue Patrimoines. Enjeux contemporains de la recherche*, n° 137, p. 22-23.

Le Follic A., Stirling P., Wendland B. (2013). *Putting it all together: creating a unified web harvesting workflow at the Bibliothèque nationale de France,* http://netpreserve.org/wp-

content/uploads/IIPC_project-Putting_it_all_together-web_harversting_workflow_at_BnF.pdf

Moiraghi E. (2017). *Décrire, transcrire et diffuser un corpus documentaire hétérogène : méthodes, formats, outils*, https://bnf.hypotheses.org/2214

Moiraghi E. (2017). *Géolocalisation et spatialisation de documents patrimoniaux : trois heures de partage autour de la cartographie numérique*, https://bnf.hypotheses.org/2299

Moiraghi E. (2018). *Données liées et données à lier : quels outils pour quels alignements ?*, https://bnf.hypotheses.org/4128

Moiraghi E. (2018). *Le projet Corpus et ses publics potentiels : Une étude prospective sur les besoins et les attentes des futurs usagers*, https://hal-bnf.archives-ouvertes.fr/hal-01739730

Moiraghi E., Moreux J.-P. (2018). *Explorer des corpus d'images. L'IA au service du patrimoine*, https://bnf.hypotheses.org/2809

Moiraghi E. (2018). *Penser, classer, modéliser. L'exemple du projet Foucault Fiches de Lecture*, https://bnf.hypotheses.org/7445

Moretti F. (2013). *Distant reading*, Verso, Londres/New York.

Moreux J.-P. (2016). *Approches innovantes pour la presse ancienne numérisée : fouille et visualisation de données*, https://bnf.hypotheses.org/208

Moreux J.-P. (2016). Data Mining Historical Newspaper Metadata - Old News Teaches History », *IFLA News Media Section Conference 2016*, Hamburg.

Moreux J.-P. (2018). *Plongez dans les images de 14-18 avec notre nouveau moteur de recherche iconographique GallicaPix*, http://gallicastudio.bnf.fr/bo%C3%AEte-%C3%A0-outils/plongez-dans-les-images-de-14-18-en-testant-un-nouveau-moteur-de-recherche

Moreux J.-P., Chiron G. (2018). Hybrid Image Retrieval in Digital Libraries: A Large Scale Multicollection Experimentation of Deep Learning techniques, *Theory and Practice of Digital Libraries 2018*, Porto.

Moreux J.-P. (à paraître). Recherche d'images dans les bibliothèques numériques patrimoniales - Expérimentation de techniques d'apprentissage profond, *Documentation et bibliothèques*, vol. 64, n° 4.

Nouvellet A., Beaudouin V., D 'Alché-Buc F., Prieur C., Roueff F. (2017). *Analyse des traces d'usage de Gallica : une étude à partir des logs de connexions au site Gallica*, https://hal.archives-ouvertes.fr/hal-01709264

Pardé T., Jacquot O. (2016). Les humanités numériques à la Bibliothèque nationale de France. *Revue Patrimoines. Enjeux contemporains de la recherche*, n° 133, p. 67-69, https://hal-bnf.archives-ouvertes.fr/hal-01379908

Stirling P. (2017). *Le dépôt légal de l'internet dans le projet CORPUS*, https://webcorpora.hypotheses.org/111

Ventresque V. (2018). *Atelier BnF Corpus (II) — Penser, classer, modéliser,* https://ffl.hypotheses.org/1079