Quand le dépôt légal devient numérique : épistémologie d'un nouvel objet patrimonial

Emmanuelle Bermès

Quand le dépôt légal devient numérique : épistémologie d'un nouvel objet patrimonial

Revue Quaderni - Emmanuelle Bermès - Juin 2018

Selon le Code du patrimoine, la BnF reçoit par dépôt légal tout document édité, importé ou diffusé en France : sont ainsi concernés les livres, périodiques, documents cartographiques, la musique notée, les documents graphiques et photographiques, mais aussi les documents sonores, les vidéogrammes, les documents multimédia, les logiciels et bases de données, ainsi que, depuis 2006, les sites web. Aucun jugement de valeur, qu'il soit moral, esthétique ou social n'est appliqué pour sélectionner les documents appelés à faire patrimoine, au contraire d'une politique documentaire classique telle que la pratiquent les autres types de bibliothèques, qu'elles s'adressent à un public de recherche comme les bibliothèques universitaires ou plus diversifié comme les bibliothèques publiques. La conséquence de cette particularité est l'existence, dans les collections de la BnF, de masses considérables de documents que l'on ne trouve nulle part ailleurs, qui reflètent l'esprit de leur époque et présentent un potentiel d'usage allant bien au-delà de la documentation : ils peuvent servir de source, de matériau brut pour étudier la société qui les a produits. Avec le numérique, ce potentiel est démultiplié : la disponibilité de ce matériau s'accroît, toujours dans le respect du droit d'auteur, et il devient possible de lui appliquer des méthodes de lecture « distante » afin de l'appréhender de manière globale et non plus unitaire.

1) Vers un dépôt légal numérique

Mission régalienne de la Bibliothèque nationale de France depuis l'ordonnance de Montpellier de 1537, le dépôt légal vise à constituer un patrimoine national représentatif de la production éditoriale de son temps. À l'obligation de dépôt pour les éditeurs s'assortit, pour la bibliothèque, celle de signaler ce patrimoine et d'y donner accès, dans ses emprises, aux chercheurs accrédités. À travers le temps, le dépôt légal s'est adapté aux évolutions des formes et supports de la production éditoriale : à l'imprimé se sont ajoutés la photographie, le son, l'image animée puis le multimédia. L'objectif de cette évolution restait de constituer un patrimoine national représentatif et autant que faire se peut, exhaustif de la production de la nation, qu'elle soit scientifique, littéraire, artistique, intellectuelle ou commerciale, sous toute ses formes.

1.1. Ce que le numérique fait au dépôt légal

En 2005, le département du dépôt légal recevait entre 150 et 300 livres par jour, soit un peu plus de 60.000 livres pour l'année 2005 (Heller, 2006). Déjà à cette époque, les chiffres démentaient le mythe selon lequel la production de papier avait vocation à être enrayée par le numérique : ce chiffre représentait un record, le nombre de dépôts éditeur enregistré étant de 42.000 en 1995 et moins de 30.000 en 1975. Par la suite, les dépôts de monographies ont continué à croître jusqu'à dépasser pour la 1e fois les 80.000 ouvrages en 2017 (Observatoire, 2017). Principale cause identifiée de cette croissance : il n'a jamais été aussi facile de produire et diffuser des documents imprimés ou audiovisuels de toute nature. Il n'est plus nécessaire d'être un professionnel de

l'édition pour mettre en page, imprimer et diffuser un livre : l'autoédition représente environ 10% des titres en 2010) ; en 2017, elle atteint 17 % et représente 45% des nouveaux déposants.

Cependant, si l'on s'intéresse à d'autres supports que le livre, on voit leur part se réduire au profit de contenus échangés en ligne. Le nombre de titres de périodiques est ainsi en baisse depuis plusieurs années, avec une transition constatée vers une version numérique du journal ou de la revue mais également vers des formes complètement différentes : site web, lettre d'information envoyée par courriel, réseaux sociaux… Le même constat s'applique aux supports de diffusion de la musique et du film (CD, DVD, Blu-Ray…)

1.2. Ce qu'internet fait au dépôt légal

La transition d'un certain nombre de contenus vers une diffusion en ligne, tendance déjà sensible dans les années 2000, ainsi que le constat de l'importance croissante de la publication sur le web comme phénomène de société ont constitué les deux principales motivations, pour la BnF, de se lancer à grande échelle dans l'archivage de l'internet. La loi dite « DADVSI »(Droits d'Auteurs et Droits Voisins dans la Société de l'Information) de 2006 étend le dépôt légal aux « signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique » ce qui permet au dépôt légal d'embrasser désormais le web mais aussi les autres formes de publications dématérialisées (ebooks, plateformes musicales, VOD etc.), celles-ci devant cependant faire l'objet d'un décret spécifique encore en cours d'élaboration à l'heure où nous écrivons ces lignes.

À la BnF, les archives du web remontent à 1996 (si l'on inclut les collections rétrospectives, ou « incunables du web », qui furent rachetées à Internet Archive a posteriori). Leur constitution est automatisée : contrairement aux autres types de dépôt, ce n'est pas l'éditeur qui envoie sa production à la BnF, mais la bibliothèque qui prend en charge la constitution de la collection au moyen de robots qui parcourent le web en effectuant des copies des pages et des ressources qu'ils découvrent (Game, 2006 ; Illien, 2008). Ces archives résultent de trois types de collecte :

- La « collecte large », une capture annuelle de l'ensemble du web français, réalisée à partir d'une liste de sites en .fr ou dont le producteur est identifié comme français. Cette liste est établie en partenariat avec des acteurs comme l'AFNIC ou encore des hébergeurs privés. Elle permet de capturer en surface des millions de sites chaque année et a représenté en 2017 plus de 100 téra-octets de données. Massive et automatisée, elle constitue ce qui se rapproche le plus de l'exhaustivité, dans l'esprit du dépôt légal.
- Les « collectes ciblées » : des collectes réalisées tout au long de l'année dont la cible, la fréquence et la profondeur sont déterminées par des bibliothécaires experts du domaine thématique correspondant. Ces collectes peuvent porter sur des événements planifiés comme les élections, des événements imprévus comme l'attentat contre Charlie Hebdo en 2015, un type de contenu comme la collecte « Actualités » qui porte sur les sites de presse et médias, une thématique comme les « Alsatiques en ligne » sélectionnés en partenariat avec la Bibliothèque nationale et universitaire de Strasbourg… Elles visent souvent, pour un site donné, à tenter d'en préserver la totalité du contenu, soit en raison de son intérêt intellectuel et scientifique, soit dans un souci d'assurer la continuité avec les collections sur support papier ou autre, soit enfin dans un esprit propre au dépôt légal de représentativité des contenus diffusés en ligne comme dans le cas des journaux intimes ou des réseaux sociaux.
- Cas particulier, la collecte de la presse quotidienne en ligne emprunte également le canal technique de l'archivage du web mais à des fins bien distinctes. 41 titres régionaux et 2 titres nationaux sont concernés, représentant des centaines d'éditions locales qu'il était impossible de collecter et conserver dans leur intégralité sous forme papier. L'archivage de la version en ligne vient donc en partie se substituer à la collecte traditionnelle.

L'ensemble de ces collections d'un genre nouveau représente près d'un péta-octet de données pour des milliards d'URL. Elles sont accessibles dans les conditions définies par le code du patrimoine : dans les emprises de la bibliothèque, pour un public de chercheurs accrédités. L'arrêté du 16 septembre 2014 du ministère de la Culture et de la Communication fixe la liste des 26 bibliothèques et autres organismes en région habilités à donner également accès à ces archives dans leurs emprises aux mêmes conditions : ce sont, pour la plupart, les bibliothèques de « dépôt légal imprimeur » (Martin, 2017).

1.3. Nouvelles perspectives de recherche

À l'image de la société qui est de plus en plus saturée d'informations de toute nature, la BnF se trouve ainsi confrontée à des flux énormes de collections numériques qu'il lui faut collecter, préserver, décrire et communiquer. Mais ces nouvelles collections et les métadonnées qui les décrivent présentent également une opportunité exceptionnelle pour la recherche. Elles constituent un matériau qu'il devient possible de fouiller avec des outils numériques, qui permettent d'appréhender la masse bien au-delà de ce qui était possible pour l'œil ou le cerveau humain.

Ces dernières années ont ainsi vu émerger des usages du dépôt légal qui s'intéressent à sa globalité plus qu'ils ne reposent sur l'étude de tel ou tel document. Alors que les interfaces d'accès, qu'il s'agisse de numérisation, d'archives du web ou même des métadonnées du catalogue, ont jusque-là été conçues pour imiter au plus près l'expérience de lecture originale et donner accès aux documents un à un, comme on le faisait avec les collections traditionnelles, la disponibilité de ces masses de contenus et d'informations sous forme numérique conduit à les envisager comme un tout. L'outil informatique, de l'intelligence artificielle à la fouille de texte et de données, permet désormais de les analyser en masse et d'en extraire de nouveaux types de connaissances : statistiques, représentations visuelles, etc. On peut comparer cette démarche à la « lecture distante » que Franco Moretti propose d'appliquer à la littérature (Moretti, 2013).

Les métadonnées associées au dépôt légal constituent par elles-mêmes un reflet du patrimoine de la nation. Créées par la BnF ou agrégées d'autres sources (dont les éditeurs eux-mêmes), elles constituent la *Bibliographie nationale française*, une publication devenue numérique dans les années 2000 (http://bibliographienationale.bnf.fr). Elles couvrent le dépôt légal de manière aussi complète que possible : les archives du web sont la seule collection de dépôt légal à ne faire l'objet d'aucun catalogage, mais des listes de sites sont toutefois disponibles sous licence ouverte, comme le reste des métadonnées. Avec son *Observatoire du dépôt légal*, un rapport annuel publié en ligne qui décrit le paysage de l'édition vu à travers le prisme du dépôt légal, la BnF est le premier utilisateur de ces données (Pajou, 2016). Conjointement avec cette publication, des jeux de données brutes sont également mis à disposition dans Data.gouv.fr pour permettre à d'autres, chercheurs, acteurs du monde de l'édition, etc. d'en faire autant.

Les archives de l'internet constituent un autre corpus d'intérêt pour ces nouvelles méthodes. Le projet « Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre » conduit dans le cadre du Labex « Les passés dans le présent » par deux bibliothèques, la BnF et la BDIC (Bibliothèque de documentation internationale contemporaine) de 2013 à 2016, a été le premier cas concret d'utilisation de méthodes d'analyse globales pour étudier les archives de l'internet (Baudouin, Pehlivan, 2017). Un corpus de sites portant sur la 2e guerre mondiale a été collecté pour analyser la manière dont les communautés d'amateurs réutilisent sur internet les contenus institutionnels numérisés diffusés par les bibliothèques. Les résultats ont été de deux types : des visualisations de graphes représentant les liens entre les sites du corpus d'une part, et l'analyse détaillée des données du forum 14-18 d'autre part. Le projet a montré qu'utiliser un corpus d'archives web comme source donnait des résultats plus fiables et maîtrisables que de travailler sur le web vivant, en raison du travail de sélection et d'analyse réalisé en amont par les bibliothécaires. Le fait de s'appuyer sur les données issues du dépôt légal présentait en outre l'avantage de résoudre les problématiques juridiques qui interdisaient, dans un autre contexte, la pérennisation du corpus.

Enfin, en plus de déboucher sur la création d'outils pour l'analyse de corpus dans les archives de l'internet et sur la montée en compétences des équipes de la BnF, ce projet a contribué à démontrer l'intérêt de travailler conjointement entre bibliothécaires et chercheurs sur ces nouveaux objets afin d'en fonder l'approche épistémologique.

2) Le projet Corpus

Nouvelles collections, nouveaux outils, nouvelles problématiques de recherche, nouvelles méthodologies : il revient aux bibliothèques et aux chercheurs de construire ensemble une nouvelle histoire de la constitution des savoirs au XXIe siècle. En 2016, la BnF s'est lancée dans le cadre de son plan quadriennal de la recherche dans le projet Corpus, dont l'objectif est de « préfigurer un service de fourniture de corpus numériques »[1] à destination de la recherche. Confrontée de plus en plus fréquemment à des demandes de partenariats émanant d'équipes de recherche qui souhaitaient l'associer en tant que fournisseur de données ou de contenus à leurs réponses à des appels à projets nationaux ou européens, la BnF s'est donnée 4 ans pour conduire des expériences concrètes, observer leurs besoins et leurs méthodes de travail, et en déduire les éléments d'une nouvelle offre de services dédiée à l'étude des corpus numériques par les chercheurs.

2.1. Le projet Corpus et les archives de l'internet

La 1e année du projet était consacrée aux archives de l'internet : en partenariat avec une équipe de l'Institut des sciences de la communication du CNRS en charge du projet ANR Web90, une application expérimentale nommée « Archives web Labs » a été développée (Stirling, 2017). Celle-ci proposait l'indexation en plein texte de deux corpus sélectionnés par l'équipe de recherche partenaire : les « incunables du web » de 1996 à 2000 et la collecte « attentats » de 2015. Cette nouvelle fonctionnalité permet de rechercher tous les mots présents dans les pages, au lieu de n'y accéder que par l'adresse URL comme c'est le cas dans l'interface standard. Dans la continuité du projet précédent sur la Grande Guerre (Baudouin, Pehlivan, 2017), la plateforme proposait également un certain nombre de métadonnées extraites de ces corpus et pouvant être utilisées par exemple pour des statistiques ou des visualisations de données. Enfin, elle offrait aux chercheurs partenaires du projet un certain nombre de fonctions de personnalisation : enregistrement de leurs requêtes, exports de résultats… Ce travail a conduit à l'organisation d'une manifestation scientifique organisée conjointement avec l'équipe ANR Web90 et avec l'Ina : « Il était une fois dans le web, 20 ans d'archives de l'internet en France »[2].

La mise en place de cette plateforme Labs a été riche d'enseignements pour la BnF. Elle a permis de mieux comprendre les besoins des chercheurs s'intéressant aux archives de l'internet en analysant les éléments qui leur ont été le plus utiles. En outre, en 2018, elle a 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 dédié. À cette occasion, un nouveau corpus a été indexé, la collecte Actualités (2010-2017).

La mise en place de la recherche plein texte dans les archives de l'internet était un défi technique pour la BnF, qui n'avait pas les moyens de la développer à l'échelle de l'ensemble de la collection. L'approche par corpus, initiée par le projet et développée dans le cadre d'autres partenariats de recherche, est une manière de lever cet obstacle et de proposer aux lecteurs de la BnF des services à valeur ajoutée.

1 Http://c.bnf.fr/fom

2 https://webcorpora.hypotheses.org/conference

2.2. Autres recherches et projets connexes

En parallèle du lancement du projet Corpus, les demandes d'accès à des corpus numériques ont continué à affluer. En 2015, le Labex OBVIL (Observatoire de la Vie Littéraire), qui était déjà un partenaire majeur de la BnF pour numérisation, a demandé l'accès à 135.000 textes de Gallica afin de conduire des analyses de type « big data » avec leur partenaire le laboratoire ARTFL de l'université de Chicago. Celui-ci souhaitait en effet répliquer sur des textes français les résultats du projet Commonplaces (Roe et al., 2016) qui utilise un algorithme pour identifier des passages similaires au sein d'un corpus de 200.000 textes scientifiques et littéraires imprimés en Grande-Bretagne au XVIIIe siècle. L'objectif de cette recherche est d'identifier des « lieux communs », c'est à dire des citations très fréquentes, sans avoir de connaissance a priori de la nature des textes faisant l'objet de citations ou de ceux qui les réutilisent. Le produit de cette analyse est en lui-même un nouveau corpus, une base de données que l'on peut interroger pour effectuer des analyses plus poussées. Une telle recherche ne pourrait être entreprise avec un périmètre aussi étendu et interdisciplinaire sans l'outil numérique : elle rend possible l'identification de motifs qui n'auraient pas nécessairement été envisagés a priori par des chercheurs.

En 2016, une autre étude conduite par Télécom Paristech dans le cadre du Bibli-lab, observatoire des usages des bibliothèques numériques, visait à utiliser des techniques de fouille de données pour analyser les traces d'usage (« logs ») de Gallica et de Data.bnf.fr (Nouvellet, 2017). En 2017, un nouveau projet a été lancé avec le LIPN, un laboratoire de linguistique de l'université Paris 13, afin de suivre le devenir des néologismes dans la langue française à travers un corpus d'archives de l'internet. Dans le cadre du projet Corpus, une équipe du CELSA (un laboratoire de sciences de l'information de Sorbonne Université) s'est associée à la BnF non seulement pour la numérisation d'un corpus de presse du XIXe siècle mais également pour l'accès à des espaces de travail dédiés et l'organisation d'ateliers méthodologiques sur les humanités numériques. Fin 2017, la BnF a commandité au Centre Européen de Valorisation Numérique Valconum (http://valconum.org/) une étude comparative des techniques de recherche dans les images par le contenu, qui a donné lieu à un autre atelier Corpus (Moiraghi et al., 2018).

On pourrait ainsi multiplier les exemples de projets qui démontrent que le besoin d'accès à des corpus numériques mais également à des services autour de ces corpus n'est pas en recul. Cependant, il existe encore aujourd'hui, du côté des chercheurs comme de celui des bibliothécaires, un certain scepticisme à l'égard de ces nouvelles méthodologies de recherche basées sur les outils numériques. Les méthodes quantitatives en sciences humaines sont connues de longues dates ainsi que leurs limites. Des projets comme le Commonplaces d'ARTFL sont parfois accusés de ne faire que confirmer que des évidences déjà connues : est-il vraiment surprenant que les extraits les plus cités soient issues de la Bible ou des pièces de Shakespeare ? Une large part des résultats obtenus par l'extraction ou la visualisation de données sont, quant à eux, fortement dépendants de la nature de la collection et de l'interprétation du chercheur qui a conçu les algorithmes.

Les collections numériques peuvent-elles réellement refléter avec fidélité la société qui les a construites ? Quels sont les biais induits par les outils que nous utilisons pour les constituer et pour y accéder ? Quelle proportion de notre mémoire risque d'être perdue, voire l'est déjà, avec la volatilité du numérique ? Les réponses à ces questions ne peuvent émerger que de l'étude des collections, en utilisant des méthodes rigoureuses. Une part importante de ces investigations est encore à l'heure actuelle consacrée à l'étude de la nature des données et à l'évaluation scientifique des méthodes et des outils utilisés pour conduire les analyses, une tâche qui n'est possible qu'à travers une proche collaboration entre les bibliothécaires et les chercheurs. Ainsi, les algorithmes de fouille et autres outils numériques ne sont pas à considérer comme des objets de recherche en tant que tels, mais plutôt comme des moyens à mettre en œuvre, parmi d'autres, pour répondre aux problématiques scientifiques identifiées par les chercheurs.

2.3. Diversité des résultats de recherche

Indépendamment de ces questionnements épistémologiques, l'existence de données disponibles en masse et d'outils aptes à les traiter constitue une opportunité pour le développement de ces nouvelles recherches, telles qu'on les a vues émerger depuis plusieurs années à la BnF. Il est intéressant de noter qu'elles ont débuté à l'origine de manière dispersée, sans vue d'ensemble ni stratégie affirmée de l'établissement en faveur de la fouille de texte et de données. 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'améliorer la visibilité de telle ou telle collection... Mais en général, ces projets étaient toujours le fruit d'une rencontre entre une équipe de chercheurs et une équipe de bibliothécaires partageant un intérêt commun pour une collection ou un corpus.

Si l'on s'intéresse aux résultats obtenus, différents cas de figure se présentent. Certains projets débouchent sur l'étude exhaustive de corpus présentant un intérêt en soi, comme dans le cas du Labex OBVIL cité plus haut, des incunables du web ou encore du projet Europeana Newspapers (Moreux, 2016). Ces corpus identifiés comme sources de valeur scientifique, parfois pluridisciplinaire (dans le cas de la presse du XIXe siècle, les disciplines concernées vont de l'histoire à la littérature en passant par la sociologie, la généalogie, l'histoire des arts ou des sciences…), font l'objet d'analyses intensives qui visent à en extraire des données à des fins diverses, potentiellement réutilisables pour plusieurs autres projets. Le produit de ce type de recherche peut être une plateforme intermédiaire qui va elle-même servir d'outil pour investiguer d'autres questions scientifiques plus pointues.

Le deuxième type de résultat porte sur l'amélioration des outils informatiques : la profondeur chronologique et la couverture géographique ainsi que la diversité documentaire des corpus patrimoniaux en fait des objets particulièrement intéressants pour le développement d'algorithmes de recherche performants. En 2014, le laboratoire ETIS, un laboratoire d'informatique de l'Université de Cergy Pontoise, a proposé d'explorer la possibilité d'automatiser partiellement l'indexation des images en extrayant automatiquement des labels ou des mots-clefs à partir de leur analyse. Conduit sur un sous-ensemble de 4.000 images extraites de la banque d'images du Département de la reproduction, le projet a montré (Picard et al., 2015) qu'en comparaison avec les images génériques utilisées dans la plupart des benchmarks informatiques dans le domaine de la fouille d'images, ce corpus soulevait des problématiques spécifiques liées à un haut degré d'interprétation et à la connaissance spécifique qu'il était nécessaire de mobiliser pour extraire des connaissances à partir de telles images. C'est donc le défi présenté à l'outil et les erreurs de celui-ci qui sont intéressants afin d'identifier les points à améliorer. Ainsi, du point de vue du bibliothécaire, le projet ne débouche pas sur un projet opérationnel qui permettrait d'alléger les tâches de catalogage des images. Cependant, du point de vue du laboratoire d'informatique, l'échec lui-même est intéressant pour faire progresser les outils développés.

Pour la bibliothèque, ces cas d'usages permettent de rassembler l'ensemble des questionnements pertinents, des plus techniques et basiques (comment mettre en place un serveur permettant au chercheur de stocker et traiter ses données ?) aux plus stratégiques (la bibliothèque devrait-elle s'engager dans des projets qui peuvent déboucher sur des échecs plutôt que des outils réutilisables ?) ; des questions épistémologiques (que peut-on attendre de l'analyse quantitative d'un corpus qui a été numérisé ou collecté par la BnF ?) aux aspects organisationnels (quelles ressources la BnF peut-elle dégager pour aider les chercheurs à mettre en place leurs nouvelles méthodologies de travail ?) On voit ici apparaître l'ensemble des problématiques auxquelles la bibliothèque se trouvera confrontée à l'heure de développer une activité autour des humanités numériques.

3. Vers un laboratoire d'humanités numériques à la BnF ?

3.1. Une priorité stratégique

Sur la base de ces premières expériences, la BnF a fixé comme l'une de ses priorités stratégiques le fait d'offrir ce type de services à ses usagers. Le contrat de performance de l'établissement pour la période 2017-2021 indique ainsi parmi ses objectifs : « Offrir aux chercheurs, dans les emprises de la Bibliothèque, des outils de fouille et d'exploration de textes et de données sur des corpus numériques de la BnF. » La décision de lancer le projet Corpus a été motivée par la difficulté évidente de continuer à servir les équipes de recherche de manière satisfaisante si jamais la demande venait à augmenter.

Les usagers concernés sont très différents de ceux qui fréquentent habituellement les salles de lecture. Ils se présentent généralement en équipe, assemblant plusieurs types de compétences : des chercheurs émérites ou des doctorants, des ingénieurs de recherche, des experts en contenus numériques ou en méthodologie... et disposent souvent d'un financement obtenu dans le cadre d'un appel à projet. Cette particularité a conduit la BnF à les traiter comme des partenaires plutôt que comme des lecteurs : signature d'une convention, accueil dans les locaux du personnel, mise à disposition d'équipes dédiées et d'infrastructures informatiques... Ils ne sont donc pas considérés comme des usagers, en réalité, mais comme des partenaires professionnels, alors même que leur objectif premier est d'étudier les collections.

Par ailleurs, leurs compétences dans le domaine des données sont variables. Les chercheurs issus des humanités ont parfois besoin d'un accompagnement poussé sur les aspects technologiques, surtout quand ils n'ont pas d'ingénieur au sein de leur équipe. Dans des disciplines proches des sciences dures, ils souhaitent au contraire disposer de données les plus brutes possibles afin de tester leurs propres outils et infrastructures. Entre ces deux extrêmes, il existe une grande variété de situations pouvant conduire la bibliothèque à proposer des services diversifiés : numérisation à la demande, conseil sur les formats et standards mais aussi sur les collections et les données, organisation d'ateliers, mise à disposition d'outils, et bien d'autres, allant de la création des corpus à leur mise à disposition dans des infrastructures sécurisées.

Or, la BnF n'est en mesure de s'associer en tant que partenaire, d'apporter des ressources humaines dédiées et de construire des outils ad hoc que pour un nombre limité de projets chaque année. Développer les humanités numériques et d'une façon générale, tous les usages liés aux collections numériques devrait plutôt être perçu comme un levier de fidélisation, voire de reconquête d'un public de chercheurs qui avait pu se détourner de la fréquentation des salles de lecture en raison de la disponibilité à distance d'une partie croissante des collections.

L'idée de services dédiés aux usages innovants des collections numériques, dont la réutilisation des données ouvertes et la fouille de textes et de données, est largement répandue dans les bibliothèques de recherche à l'heure actuelle. Tandis qu'une initiative similaire à la British Library a pris la forme d'événements multiples et de projets réunis virtuellement (McGregor et al., 2016), la bibliothèque du Congrès aux Etats-Unis a publié un rapport encourageant la création d'un laboratoire (Gallinger, Chudnov, 2016). Le centre pour les humanités numériques de l'Université de Leiden est un autre exemple du type de services que les bibliothèques peuvent créer pour soutenir les chercheurs dans le développement de compétences pour l'analyse de corpus numériques (Oudenhoven, 2017). Une étude publiée par l'ADBU (ADBU, 2016) explore le potentiel du « TDM »[3] pour la recherche et le rôle que pourraient être appelées à jouer les bibliothèques. Enfin, un groupe « Digital Humanities » s'est créé au sein de la Ligue Européenne des Bibliothèques de Recherche (LIBER) qui en a fait un de ses axes stratégiques (https://libereurope.eu/strategy/digital-skills-services/digitalhumanities/).

C'est dans ce contexte international favorable que la BnF a lancé le projet Corpus. Afin de définir ses propres services aux chercheurs, celle-ci devait cependant prendre en compte ses spécificités, notamment la richesse considérable de ses collections mais aussi le fait que les documents issus du

3 TDM : Text and Data Mining

dépôt légal ne sont accessibles que dans les emprises physiques de l'établissement. C'est la raison pour laquelle, au-delà même d'un nouveau service, un espace physique est envisagé.

3.2 L'étude des besoins

Afin de mieux caractériser les besoins des chercheurs, une étude a été conduite d'août à décembre 2017 (Moiraghi, 2018). Basée sur 30 entretiens conduits avec des chercheurs et des experts, aussi bien internes qu'externes à la BnF, l'étude s'est conclue sur l'organisation d'un atelier collaboratif utilisant la méthode des « personas » (Cooper, 1995) pour concevoir les futurs services à mettre en œuvre.

Le rapport met en lumière les contraintes techniques, humaines, d'organisation et juridiques qui pèsent sur les projets. Il fait également apparaître les motivations principales, pour les chercheurs, à venir travailler à la bibliothèque : ce n'est plus l'attractivité des collections qui motive leur venue dans les emprises de l'établissement, car celles-ci étant numériques, il est généralement considéré qu'elles devraient être accessibles à distance, en dépit des contraintes juridiques souvent méconnues des chercheurs. En revanche, la possibilité d'avoir un accès immédiat aux experts de la bibliothèque est perçue comme la principale valeur ajoutée d'un tel espace. Expertise sur les fonds, conseil juridique, soutien informatique et autour de l'analyse des données, orientation pourraient être apportés par des membres du personnel BnF (conservateurs, juristes, ingénieurs, informaticiens, experts des formats des données) mais impliqueraient également la présence permanente de personnel spécialisé dans la science des données et ayant une formation en sciences humaines et sociales.

En outre, la pénurie de locaux dans les universités parisiennes peut conduire les chercheurs interrogés à imaginer un usage étendu de la bibliothèque, pour y travailler, y donner des cours, voire y installer une partie des équipes de recherche de façon permanente sur des durées allant de quelques mois à plusieurs années. Il existe également une attente de pouvoir valoriser dans cet espace les réalisations de l'équipe de recherche, que ce soit par des événements, des démonstrateurs, des présentations… le prestige de la BnF étant de nature à rejaillir sur les projets présentés dans ses espaces.

Tel que défini dans le rapport, le futur espace consacré dans les emprises de la Bibliothèque doit être facile d'accès, convivial et capable d'évoluer au rythme de l'innovation et du progrès technologique. Il devrait comprendre idéalement à terme des bureaux pour les agents de la BnF, mais aussi pour les chercheurs en résidence et les enseignants, un lieu susceptible d'accueillir des événements, un espace de restauration et de détente, des salles pour le travail en groupe et les formations, et enfin un espace de présentation et d'exposition des résultats de la recherche et de partage d'expérience.

Si ces différents éléments correspondent en réalité à ce que l'on peut naturellement attendre des services d'une bibliothèque comme la BnF, le fait de les concentrer dans un lieu dédié aux humanités numériques devrait contribuer à améliorer leur visibilité et par là, celle des collections numériques elles-mêmes. Le département de l'orientation et de la recherche bibliographique a été pressenti pour accueillir cette activité dans sa salle de lecture, renforçant ainsi sa compétence de point d'entrée dans les collections encyclopédiques de la BnF d'une part, et de manipulation experte des outils d'accès aux collections d'autre part.

3.3. Le pendant virtuel du futur laboratoire

L'idée d'un espace physique dans la bibliothèque ne doit cependant pas oblitérer la dimension numérique du projet. La politique de dissémination des données déjà en œuvre à la BnF implique

des API comme IIIF[4], disponible pour Gallica depuis 2016, d'autres protocoles pour les métadonnées (Z39.50, OAI-PMH, SRU…) et des jeux de données comme ceux de Data.bnf.fr ou encore ceux qui ont été générés et retraités dans le cadre de projets comme celui du Labex OBVIL ou Europeana Newspapers, mentionnés plus haut. Certaines des demandes des chercheurs peuvent être déjà largement couvertes par l'accès à ces services. Depuis 2017, un site web dédié, le site API et données (api.bnf.fr) propose un recensement complet, assorti d'une documentation exhaustive et d'exemples, de toutes ces ressources techniques. Si cette initiative ne cible pas spécifiquement les chercheurs, elle constitue un moyen simple et unifié pour ceux-ci d'accéder aux ressources et à la documentation qui leur permet d'alimenter de nouveaux angles de travail sur les collections (Langlais, 2017). Le travail d'extraction et de prétraitement des corpus réalisé conjointement par la BnF et les équipes de recherche dans le cadre de partenariats offre ainsi de nouvelles opportunités pour le développement des humanités numériques. La dissémination de ces données avec une politique claire d'autorisation de réutilisation constitue un apport notable au développement de la science ouverte.

Conclusion

La BnF doit trouver sa place dans l'écosystème de la recherche : outre le fait de remplir sa mission de fournisseur de données, il lui revient de développer formation et accompagnement autour de son savoir-faire, de l'histoire de ses données et des collections. Elle pourrait également orienter les utilisateurs vers des réseaux académiques de pairs et des formations autour des outils et des techniques, auxquelles elle ne doit pas se substituer. Elle a enfin un rôle à jouer pour valoriser les outils existants et les résultats de la recherche, en lien avec d'autres institutions comme les bibliothèques des établissements de l'enseignement supérieur et de la recherche, les laboratoires et instituts du CNRS, les grandes infrastructures de recherche... C'est tout un tissu d'acteurs divers, dont les rôles seraient à articuler, qui se construit aujourd'hui autour des nouvelles opportunités ouvertes par les humanités numériques. Dans un univers aussi expérimental et indéfini, il est raisonnable de construire un service orienté usager mais également d'imaginer, stimuler et guider de nouveaux usages.

4 IIIF : International Image Interoperability Framework ; http://iiif.io

Références :

(ADBU, 2016) Rob Johnson, Olga Fernholz, Mattia Fosci, *Text and data mining in higher education and public research*. Rapport commandé par l'ADBU, décembre 2016. En ligne : http://adbu.fr/competplug/uploads/2017/10/TDM-in-Public-Research-Revised-15-Dec-16-1.pdf

(Bermès, 2017) Emmanuelle Bermès. « Text, data and link-mining in digital libraries: looking for the heritage gold. » Présenté à : *IFLA Satellite Meeting - Digital Humanities – Opportunities and Risks: Connecting Libraries and Research,* Aug 2017, Berlin, Germany. <hal-01643293> En ligne : https://hal.inria.fr/hal-01643293

(Baudouin, Pehlivan, 2017) Valérie Beaudouin, Zeynep Pehlivan. 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". [Rapport de recherche] Bibliothèque nationale de France; Bibliothèque de documentation internationale contemporaine; Télécom ParisTech. 2017. <hal-01425600>

(Chambers, 2016) Sally Chambers, "It's not about the catalogue, it's about the data - Catalogue 2.0: The future of the library catalogue", février 2017. En ligne : https://biblio.ugent.be/publication/8511250/file/8511251.pdf

(Cooper, 1995) Alan Cooper, *About face: the essentials of user interface design*, Foster City, 1995

(Gallinger, Chudnov, 2016) Michelle Gallinger and Daniel Chudnov, *Library of Congress Lab : Library of Congress Digital Scholars Lab Pilot Project*. Report dated Dec. 2016, http://digitalpreservation.gov/meetings/dcs16/DChudnov-MGallinger_LCLabReport.pdf?loclr=blogsig

(Game, 2006) Game, Valérie et Illien, Gildas. « Le Dépôt légal d'Internet à la Bibliothèque nationale de France ». Bulletin des bibliothèques de France (BBF), 2006, n° 3, p. 82-85. En ligne : <http://bbf.enssib.fr/consulter/bbf-2006-03-0082-013>. ISSN 1292-8399.

(Heller, 2006) Heller, Danièle. « Le Dépôt légal ou comment aimer le papier d'un amour fou ! ». Bulletin des bibliothèques de France (BBF), 2006, n° 4, p. 5-9. Disponible en ligne : <http://bbf.enssib.fr/consulter/bbf-2006-04-0005-001>. ISSN 1292-8399.

(Illien, 2008) Illien, Gildas. « Le Dépôt légal de l'internet en pratique : ». Bulletin des bibliothèques de France (BBF), 2008, n° 6, p. 20-27. Disponible en ligne : <http://bbf.enssib.fr/consulter/bbf-2008-06-0020-004>. ISSN 1292-8399.

(Johnson, 2016) Rob Johnson et al., *Text and data mining in higher education and public research*. Report commissioned by the Association des Directeurs & personnels de direction des Bibliothèques Universitaires et de la Documentation (ABDU), December 2016. http://adbu.fr/competplug/uploads/2016/12/TDM-in-Public-Research-Revised-15-Dec-16.pdf

(Koltay, 2017) Tibor Koltay, "Data literacy for researchers and data librarians". *Journal of Librarianship and Information Science*, 2017, Vol. 49(1) 3–14. http://journals.sagepub.com/doi/abs/10.1177/0961000615616450

(Langlais, 2017) Pierre-Carl Langlais, “Les bibliothèques numériques sont-elles représentatives ?” in *Sciences Communes*, avril 2017 https://scoms.hypotheses.org/799

(Martin, 2017) Frédéric Martin, “Les archives de l’internet comme axe de coopération nationale”. *Webcorpora*, 19/09/2017 https://webcorpora.hypotheses.org/394

(McGregor et al., 2016) McGregor, N., Ridge, M., Wisdom, S., Alencar-Brayner, A. (2016). “The Digital Scholarship Training Programme at British Library: Concluding Report & Future Developments”. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 623-625.

(Moiraghi, 2018) Eleonora Moiraghi. Le projet Corpus et ses publics potentiels. : Une étude prospective sur les besoins et les attentes des futurs usagers.. [Rapport de recherche] Bibliothèque nationale de France. 2018. 〈hal-01739730〉

(Moiraghi et al., 2018) Eleonora Moiraghi, “Explorer des corpus d’images. L’IA au service du patrimoine” in Carnet de recherche à la Bibliothèque nationale de France, 16/04/2108 https://bnf.hypotheses.org/2809

(Moretti, 2013) Franco Moretti, *Distant Reading*, London/New-York, Verso, 2013

(Moreux, 2016) “Approches innovantes pour la presse ancienne numérisée : fouille et visualisation de données” in Carnet de recherche à la bibliothèque nationale de France, 3 décembre 2016 https://bnf.hypotheses.org/208 which is a summary of “Data Mining Heritage Newspapers”, Document Analysis Systems 2016, Santorin; IFLA News Media Satellite Conference, Hambourg, 2016 and “Innovative Approaches for Heritage Newspapers: Data Mining, Data Visualization, Semantic Enrichment”, IFLA News Media satellite conference, Lexington, 2016.

-> Jean-Philippe Moreux. Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment : Facilitating Access for various Profiles of Users. *IFLA News Media Section, Lexington, August 2016, At Lexington, USA*, Aug 2016, Lexington, United States. <http://uknowledge.uky.edu/ifla-news-media/>. <hal-01389455>

(Nouvellet, 2017) Adrien Nouvellet et al., « Modélisation des comportements à partir de l’analyse des logs de Gallica », Journée d’étude « Quels usages aujourd'hui des bibliothèques numériques ? Enseignement et perspectives à partir de Gallica », BnF, Paris, 3 mai 2017, http://www.bnf.fr/fr/professionnels/anx_journees_pros_videos/a.video_170503_05_table_ronde_4.html.

(Oudenhoven, 2017) Martine Oudenhoven, “On the role of a university library in the TDM landscape” in *FutureTDM*, June 2017 http://www.futuretdm.eu/blog/role-university-library-tdm-landscape/

(Pajou, 2016) Pajou, Jean-Charles. « L’Observatoire du dépôt légal ». Bulletin des bibliothèques de France (BBF), 2016, n° 9, p. 134-144. Disponible en ligne : <http://bbf.enssib.fr/consulter/bbf-2016-09-0134-002>. ISSN 1292-8399.

(Picard et al., 2015) David Picard, Philippe-Henri Gosselin, Marie-Claude Gaspard. “Challenges in Content-Based Image Indexing of Cultural Heritage Collections.” *IEEE Signal Processing Magazine*, Institute of Electrical and Electronics Engineers, 2015, 32 (4), pp.95 - 102

(Roe et al., 2016) Roe, G, Gladstone, C, Morrissey, R et al 2016, 'Digging into ECCO: Identifying Commonplaces and other Forms of Text Reuse at Scale', *Digital Humanities DH2016*, ed. Maciej Eder, Jan Rybicki, The Alliance of Digital Humanities Organizations, Krakow, Poland, pp. 336-339. Abstract: http://dh2016.adho.org/abstracts/343

(Stirling, 2017) Peter Stirling, "Le dépôt légal de l'internet dans le projet Corpus." in *Webcorpora*, 24/05/2017 https://webcorpora.hypotheses.org/111

(Observatoire, 2017) Observatoire du dépôt légal : reflet de l'édition contemporaine. Paris, Bibliothèque nationale de France : 2017 (à paraître). En ligne : http://www.bnf.fr/fr/professionnels/depot_legal_definition/s.depot_legal_observatoire.html?first_Art=non