Ouvrir les données de la Bibliothèque nationale de France pour la recherche

Arnaud Laborderie, Florence Tfibel

Ouvrir les données de la Bibliothèque nationale de France pour la recherche

Opening the National Library of France's data for research

Arnaud LABORDERIE, Chef de projet Gallica, chargé de l'exploitation des données pour la recherche, Bibliothèque nationale de France (BnF)

Florence TFIBEL, Adjointe à la Cheffe du service Diffusion des métadonnées, et responsable de l'équipe Services aux professionnels, BnF

Résumé

Métadonnées, documents numérisés, archives du web ou logs : appréhender ces données comme source et terrain pour la recherche pose des problématiques d'ordre à la fois méthodologique, technique et juridique. Quelle est la singularité des données de la Bibliothèque nationale de France (BnF) ? Quelles en sont les modalités d'accès et de réutilisation ? Quelles limites à leur ouverture et à leur exploitation ?

Mots-clés : ouverture des données, données patrimoniales, métadonnées, Gallica, archives du web

Abstract

Metadata, digitized documents, web archives, logs: using these data as a source or a field of research raises methodological, technical and legal issues. What are the specificities of the National Library of France (BnF)'s data? How can they be accessed and reused? What are the limitations to their openness and usability?

Key-words : open data, heritage data, metadata, Gallica, French web archives

Introduction

Les bibliothèques sont de grandes productrices de données. À la BnF, celles-ci représentent une masse considérable (plus de 6 Po), d'une grande diversité, tant par leur forme que par leur contenu : métadonnées des catalogues, documents numérisés de Gallica et Gallica intramuros, jeux vidéo, CD et DVD, ou documents nés numériques comme les archives du web, les logs[1] de connexion, etc.

L'augmentation exponentielle de ces données et leur reconnaissance comme source et terrain pour la recherche ont fait émerger des problématiques d'accès, d'exploitation et de réutilisation. À la BnF, la question de l'ouverture et de l'accès aux données est complexe car elle dépend du type de données et du cadre juridique dans lequel celles-ci s'inscrivent. Quelle est la singularité des données de la BnF ? Comment sont-elles produites, avec quel niveau de qualité et pour quels usages ? Quelles en sont les modalités d'accès, d'exploitation et de réutilisation ? Quelles limites à l'ouverture des données de la BnF et quelles exceptions permettent néanmoins aux chercheurs de s'emparer des données protégées ?

Des données riches et très diverses

Les données sont au cœur des activités de la BnF. Ses agents contribuent à en produire quotidiennement : des équipes de catalogage rédigent des notices, des experts dédiés les enrichissent, manuellement ou par calcul, des départements numérisent leurs collections, le Dépôt légal collecte le web français, etc. Des métadonnées bibliographiques sont ainsi créées et évoluent dans les catalogues décrivant des collections ; des documents numériques sont produits par la numérisation des ressources conservées dans les départements ou collectés nativement, et font l'objet de traitements permettant d'exploiter leur contenu (OCR[2], tables des matières…) ; des données d'usage de la Bibliothèque ou de Gallica par le public sont recueillies, parmi d'autres encore. Ces données, d'abord nécessaires au bon fonctionnement de l'établissement, sont également une source très riche pour les chercheurs, qui doivent néanmoins en comprendre toute la diversité pour pouvoir les exploiter.

Les données bibliographiques des catalogues

Les métadonnées bibliographiques de la BnF sont hétérogènes à plusieurs titres. Elles proviennent de sources diverses, depuis les métadonnées fournies dans le cadre du Dépôt

[1] Dans un contexte informatique, un *log* est un « journal » qui désigne la documentation automatiquement générée et horodatée des événements concernant un système particulier. Pratiquement tous les systèmes et logiciels produisent des « fichiers journaux ».

[2] L'OCR (*Optical Character Recognition)* est une technologie de reconnaissances de textes imprimés à partir d'images numérisées. Disponible dans *Gallica* depuis 2005, l'OCR constitue aujourd'hui une chaîne d'entrée interne à la BnF, qui est engagée notamment dans des programmes de rétroconversion.

Légal éditeurs, jusqu'aux notices rédigées manuellement et enrichies par des experts, en passant par des données obtenues par calcul ou par chargement[3], ce qui implique des niveaux de qualité et de validation différents : une notice rédigée document en main par un spécialiste n'a pas la même fiabilité qu'une donnée récupérée automatiquement par flux, ou issue d'une rétroconversion.
Les données bibliographiques sont également diverses dans leur structuration et leur format : Intermarc pour le Catalogue général, XML-EAD pour le catalogue BnF Archives et manuscrits, Dublin Core[4] pour la bibliothèque numérique Gallica, RDF[5] pour le portail data.bnf.fr, WARC pour les archives du web… Le choix de ces formats est lié à la nature des collections qui sont décrites[6] ainsi qu'aux usages que l'on souhaite en faire. Le même document, décrit en Intermarc dans le Catalogue général, sera ainsi signalé de façon beaucoup plus sommaire et moins structurée, mais bien plus interopérable et adaptée à une bibliothèque numérique, en Dublin Core dans Gallica, ou plus adaptée à l'exposition web dans data.bnf.fr grâce au format RDF. Les usagers peuvent ainsi accéder à différentes qualités et structurations de données selon leurs usages et besoins.

Pour les chercheurs, qui exploitent ces métadonnées en elles-mêmes ou autour de collections numériques, cela soulève notamment des questions d'appropriation et de conversion des formats pour l'intégration des données dans leurs bases de recherche, d'alignement de leurs vocabulaires avec les référentiels de la BnF, etc.

Les données numérisées de Gallica

Avec plus de 10 millions de documents, Gallica est l'une des plus grandes bibliothèques numériques au monde, très riche et d'une large diversité documentaire : monographies, périodiques, images, manuscrits, cartes, objets, partitions, son, vidéo… Les documents numérisés sont issus des collections patrimoniales de la BnF mais également de bibliothèques partenaires dans le cadre de la coopération nationale. Le processus de numérisation produit différents types de données : des données brutes (images au format TIFF ou JPG, son et vidéo

[3] Opération permettant d'intégrer des données en masse dans les catalogues, par exemple dans le cadre de partenariats ou de rétroconversions.
[4] Le format Dublin Core, est un format de description simplifié (15 champs) et un langage du web sémantique (format obligatoire dans le cadre du protocole OAI-PMH) qui permet d'exprimer les données dans un modèle RDF. Le Dublin Core est utilisé par Gallica et RDF par DataBnF.
[5] Langage de base du web sémantique, RDF (*Resource Description Framework*) est un modèle de données destiné à décrire les ressources web et leurs métadonnées sous forme de triplet (sujet, prédicat, objet).
[6] Ainsi XML-EAD est un format plus adapté à la description des fonds d'archives qu'Intermarc, format de description bibliographique à la granularité très fine et couvrant une gamme très large de types de documents.

au format MPEG), des données dérivées (OCR, tables des matières) et des métadonnées (METS[7], manifest[8]).
La collection numérique s'est constituée sur une trentaine d'années au cours desquelles les technologies se sont considérablement améliorées (niveau de résolution, OCR, normalisation) d'où une certaine hétérogénéité de qualité des données qui peut être sensible pour les usagers. Les besoins ont aussi évolué, notamment en matière d'OCR : ce qui importait hier était de pouvoir interroger la collection par mots-clés et de rechercher en plein-texte dans les documents. Si ce besoin perdure aujourd'hui, les pratiques de fouille de texte et de données (TDM[9]) et le traitement automatique du langage (TAL) ont relevé le niveau d'exigence des taux de reconnaissance de caractères ; dans le même temps, une chaîne d'entrée a été mise en place en interne pour produire en masse la part de la collection qui n'est pas encore accessible en mode texte[10]. Si la quantité prévaut dans la logique de la numérisation de masse et reste nécessaire pour interroger largement la collection, la qualité des données est primordiale pour les chercheurs en humanités numériques. Prise entre ces deux besoins, la BnF tâche de trouver un point d'équilibre pour fournir des données les plus qualitatives aux usagers.

Les archives du web et les données d'usage

Les archives de l'Internet français, que la BnF collecte puis 2006 et dont les fonds remontent à 1996, sont un autre réservoir de données majeur. Ces collections constituées grâce à des campagnes de collectes larges et ciblées du domaine français représentent maintenant 1,8 pétaoctet de données pour 48 milliards d'URL[11].
Le web présente un cas singulier de données pour la plupart librement accessibles en ligne mais qui, lorsqu'elles sont collectées par la BnF dans le cadre du Dépôt légal, deviennent des données patrimoniales avec des restrictions d'accès. En effet, les ressources en libre accès sur le web n'en sont pas pour autant libres de droits. Certains sites contiennent des documents (audiovisuels, multimédias, etc.) qui peuvent représenter de forts enjeux économiques.

Un autre type de données susceptibles d'intéresser les chercheurs mais qui restent d'accès réservé, sont les traces d'usages que représentent les milliards de logs de connexion aux serveurs de Gallica. Ces logs enregistrent l'activité des utilisateurs et permettent de suivre leurs actions. Initialement conservés par la BnF uniquement à des fins de sécurité et

[7] METS *(Metadata Encoding and Transmission Standard*) est un standard permettant d'encoder les métadonnées descriptives, administratives et de structure, spécifiques aux objets numériques.
[8] Manifest est un fichier XML décrivant la structure d'un document numérique ou le contenu d'un paquet, d'une application logicielle, etc.
[9] Pour l'anglais *Text Data Mining*.
[10] Aujourd'hui 77% des périodiques et 70% des monographies intégrées dans Gallica sont accessibles en mode texte.
[11] cf. V. Tybin, « Les collections du Dépôt légal du Web de la BnF au cœur des réseaux de coopération internationale pour la recherche », *Culture et Recherche* n°143, p. 126-127.

d'évaluation de la qualité de service, ces données ont été anonymisées[12] pour des raisons juridiques et éthiques afin d'être mises à la disposition des chercheurs.

Ouvrir les données : une question juridique et technique

Dès 2014, la BnF s'est engagée dans une démarche d'ouverture de ses données dans une perspective de science ouverte. Elle en encourage les appropriations les plus variées, en développant une offre globale de diffusion des données[13]. Cette ouverture des données pose cependant des questions à la fois juridiques et techniques autour de l'accès et des usages des chercheurs.

Une offre globale de diffusion des données : services et outils

Pour permettre de constituer des corpus et d'exploiter les données qu'elle produit, collecte et conserve, la BnF a mis en place de la documentation et des outils permettant de les rechercher, les extraire et les réutiliser.
Les catalogues, la bibliothèque numérique Gallica et les autres plateformes de la BnF proposent depuis leurs interfaces de consultation des formulaires de recherche et des options de récupération. Le Catalogue général par exemple permet un export au format CSV des notices[14], mais aussi dans des formats techniques plus utilisés dans le monde des bibliothèques (ISO27.09) ; les usagers de Gallica peuvent récupérer les données à l'unité, au fil de leur consultation dans Gallica grâce aux fonctionnalités du visualiseur (téléchargement, partage, etc.) et des plugins[15] associés (Zotero, IIIF) ; data.bnf.fr[16] expose sur le web selon un modèle entité-relations les données produites par l'établissement et propose un accès centré sur les œuvres, auteurs, thèmes, dates ou encore lieux, et un SPARQL endpoint[17] permet d'interroger la base et de récupérer les données.

[12] L'anonymisation porte sur l'adresse IP, remplacée par une clé hashée. Sur option de traitement, la ville de l'adresse IP peut être également remplacée par une clé hashée, mais le pays de l'adresse IP reste mentionné.
[13] Cf. l'offre sur le site bnf.fr : https://www.bnf.fr/fr/reutiliser-les-donnees-de-la-bnf
[14] CSV désigne un format de fichiers dont le rôle est de présenter des données séparées par des virgules. Il s'agit d'une manière simplifiée d'afficher des données afin de les rendre transmissibles d'un programme à un autre. https://catalogue.bnf.fr/aide/content/exporter-en-csv
[15] Les plugins sont de petits programmes complémentaires qui ajoutent des fonctions aux applications Web et programmes de bureau : cf. S. Bertrand, G. Chenard, S. Pillorget, C. Prunet (coordinatrice), R. Robineau, « Découverte et interopérabilité sans frontières des images patrimoniales », *Culture et Recherche* 143, p. 111-117.
[16] https://data.bnf.fr/semanticweb
[17] Format de requête permettant d'interroger le langage de base du Web sémantique (RDF) : https://api.bnf.fr/sparql-endpoint-de-databnffr

Ces données peuvent être exposées sous d'autres formes afin d'être récupérées en nombre et exploitées par des machines, y compris hors de la communauté des professionnels des bibliothèques, grâce à des API[18]. Il est ainsi possible de requêter et de récupérer les métadonnées du Catalogue général ou de Gallica par flux (*SRU - Search and Retrieve via URL*), ponctuellement (Z39.50[19]) ou sous forme de lots de données (OAI-PMH[20]). Pour rendre cette offre plus visible, le portail *API et jeux de données*[21] recense et documente les API qui permettent d'interagir avec ses données. On y trouve également une documentation détaillant le contenu des données et les possibles utilisations de chacune de ces API, ainsi qu'un certain nombre des jeux de données (*dumps*[22] de data.bnf.fr, lots de notices de la Bibliographie nationale…), dont certains ont été constitués dans le cadre de programmes de recherche.
Des extractions complètes des bases sont également disponibles depuis plusieurs plateformes de données ouvertes telles https://www.data.gouv.fr/fr/ (données publiques françaises) et https://data.culture.gouv.fr (données ouvertes du ministère de la Culture).

Des métadonnées accessibles librement ou à certaines conditions

La grande majorité des données de la BnF sont accessibles et exploitables librement. L'intégralité des métadonnées du Catalogue général, qui contient plus de 15 millions de notices bibliographiques et plus de 5 millions de notices d'autorité, est sous licence Etalab (à l'exception des notices de la Classification décimale Dewey). Les données du Catalogue collectif de France (CCfr)[23] sont également sous Licence Ouverte, donc réutilisables librement et gratuitement sous réserve d'en mentionner la source.
La collection numérique de Gallica comprend des documents de toutes époques, de l'Antiquité à nos jours, librement et gratuitement accessibles en ligne et exploitables pour ceux entrés dans le domaine public : seule leur utilisation commerciale est soumise à redevance.

Des contraintes juridiques peuvent néanmoins peser sur les usages de certains ensembles de données spécifiques, comme les logs de connexion de Gallica, les documents sous droits ou

[18] Les API sont des interfaces de programmation applicatives permettant à des systèmes informatiques de communiquer entre eux et d'échanger des données de manière standardisée.
[19] Protocole informatique utilisé par les bibliothèques pour interroger simultanément plusieurs catalogues.
[20] *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) est un protocole informatique permettant d'échanger des métadonnées en interrogeant des entrepôts OAI, qui est largement répandu dans les institutions patrimoniales et notamment les bibliothèques.
[21] https://api.bnf.fr
[22] En informatique, un *dump* de base de données est un fichier qui contient une structure et un contenu de base de données.
[23] Le Catalogue collectif de France (CCfr) est un outil de recherche bibliographique et documentaire, permettant d'interroger simultanément une dizaine de catalogues (SUDOC, base patrimoine, Catalogue général des manuscrits, Calames, etc.) : https://ccfr.bnf.fr/

les archives du web. Rappelons que la BnF est soumise à deux cadres législatifs : le *Code du patrimoine* (CP) et le *Code de la propriété intellectuelle* (CPI), qui encadrent à la fois la collecte, l'accès et la réutilisation des données avec un certain nombre d'exceptions. Les documents sous droits de Gallica sont ainsi soumis au CPI et ne sont donc accessibles que dans l'enceinte de la BnF (certains d'entre eux étant malgré tout partiellement consultables en ligne sous forme d'extraits). Sans entrer dans le détail, il faut aussi rappeler que le Dépôt légal (CP art. L.131-132), dans le cadre duquel sont constituées les archives du web, crée des restrictions d'usage puisque les documents ainsi collectés ne sont consultables que dans les salles de recherche de la BnF[24].

Pour accompagner les usages et encadrer les pratiques en matière de fouille de texte et de données (TDM), deux nouvelles exceptions ont par ailleurs été introduites dans le CPI[25], dont une concerne les fouilles réalisées à des fins de recherche (CPI art. L.122-5-3-II). Celles-ci sont désormais possibles sans accord préalable des ayants droit et sans que ceux-ci puissent s'y opposer, permettant également aux institutions de recherche de conserver les corpus, sans pouvoir les diffuser, mais rendant possible la reproductibilité des traitements et la vérification des résultats. Cette grande avancée pose néanmoins des difficultés pour les chercheurs en termes de citabilité, notamment pour les archives du web : si les permaliens permettent de citer les URL des archives, les chercheurs ne peuvent pas publier librement leurs sources au-delà du droit de courte citation (CPI art. L122-5).

Développer les usages et collaborer avec les chercheurs sur les données

Afin de faciliter et d'encourager l'usage des données de la BnF pour la recherche, la BnF s'investit dans divers projets dont le but est de composer avec les contraintes techniques et juridiques existantes, et s'attache à accompagner au mieux les chercheurs. Accessibles sur demande sous forme de jeux de données anonymisées, les logs de connexion aux serveurs de Gallica par exemple ont pu faire l'objet de plusieurs travaux[26], dont l'objectif n'était pas de connaître les usagers ni leurs profils mais, en partant des traces d'usages que sont les logs, de mieux connaître les sessions de consultation et d'identifier des parcours de recherche sur Gallica.

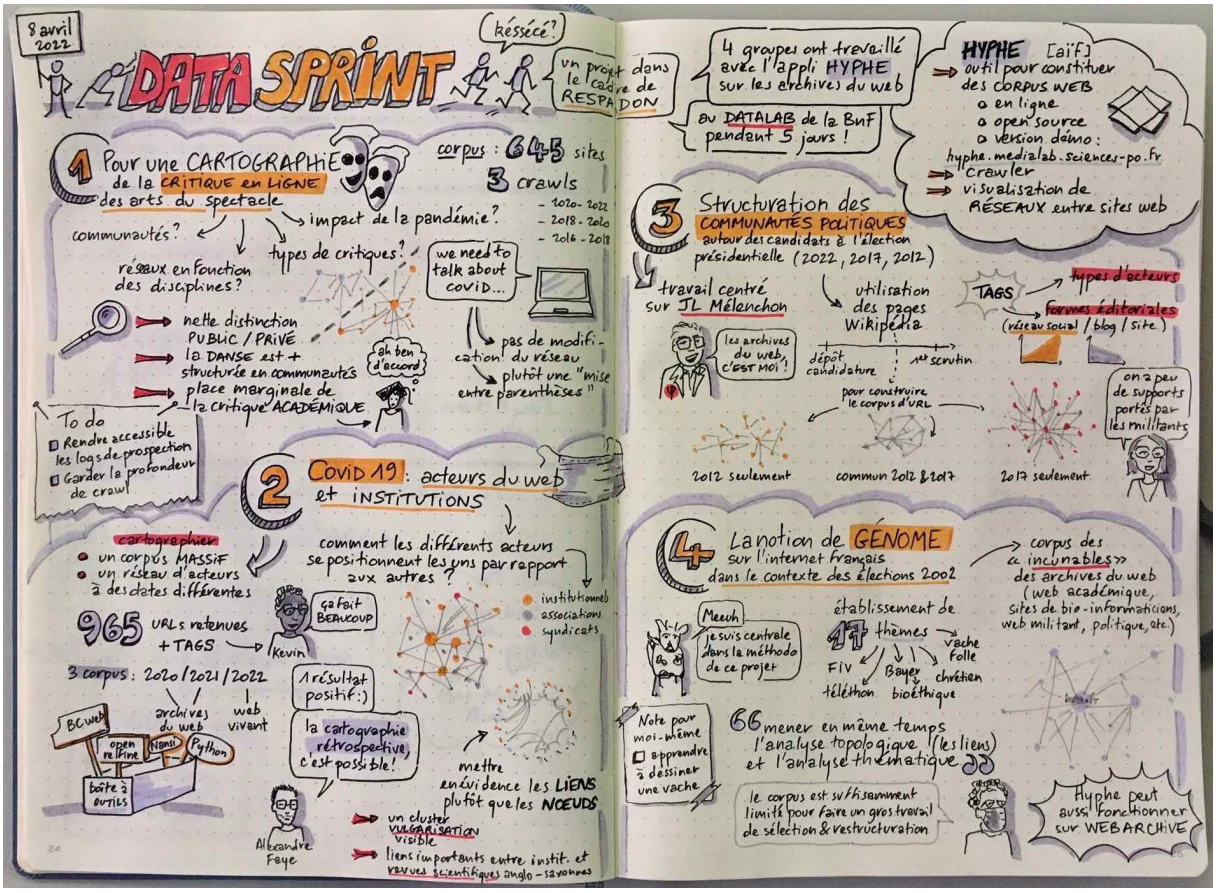
Depuis une dizaine d'années, Gallica collabore à des projets de recherche et de R&D dans le domaine des humanités numériques. La coopération avec les chercheurs permet de co-

[24] Les archives du Web sont également consultables en région dans 21 bibliothèques du Dépôt légal imprimeur (BDLI) : https://www.bnf.fr/fr/annuaire-des-poles-associes-de-depot-legal-imprimeur

[25] Ordonnance du 24 novembre 2021 transposant la directive européenne du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique.

[26] Nouvellet et al., Télécom ParisTech, 2017. Dumas-Primbault et al., EPFL, 2021. Trabelsi, La Rochelle Université, 2022. Falissard, SCAI, 2023

construire des outils qui s'inscrivent dans l'écosystème de Gallica, comme *Gallicagram*[27], outils et services qui, potentiellement, peuvent devenir des fonctionnalités de Gallica au bénéfice de tous les usagers, au-delà de la communauté scientifique. C'est le cas de Gallica images[28], dont l'objectif est d'identifier les images nichées au cœur des pages numérisées (photos de presse, illustrations, schémas, enluminures, etc.) et de proposer de nouvelles modalités de recherche visuelle. Ce projet a été préfiguré par une preuve de concept, GallicaPix[29], conduite avec l'Université de La Rochelle, expérimentant la recherche d'image par mot-clé et similarité de forme, et le projet Gallica Snoop[30], mené en partenariat avec l'INA et l'INRIA, éprouvant la recherche d'images par image exemple.



Sketchnote du *Datasprint* sur les archives du web, organisé au BnF DataLab le 8 avril 2022, en partenariat avec Sciences Po Paris. © Mélanie Leroy-Terquem / BnF

[27] Outil de lexicométrie permettant de visualiser l'évolution de l'usage des mots au cours du temps en fouillant les corpus de presse et de livres : https://shiny.ens-paris-saclay.fr/app/gallicagram

[28] Le projet « Gallica Images », porté par la BnF en partenariat avec la Bibliothèque nationale et universitaire de Strasbourg (BNU) et l'Institut national d'histoire de l'art (INHA), a reçu une subvention d'1,3 million d'euros dans le cadre de France Relance.

[29] https://gallicapix.bnf.fr/

[30] https://www.bnf.fr/sites/default/files/2022-05/Poster_Gallica_Snoop.pdf

Dans le cas des archives du web, le projet ResPaDon (2020-2023)[31] s'est employé à créer un réseau de partenaires (parmi lesquels des bibliothèques universitaires), afin de développer les usages scientifiques des collectes et de proposer un accès distant sécurisé dans des « datacapsules » permettant la consultation et l'exploitation des données, appréhendées comme source primaire et données de recherche. Un prototype est actuellement expérimenté à l'Université de Lille.

Ouvert en 2021, le BnF DataLab s'inscrit dans ce *continuum* d'outils, de services et de projets. Ceux issus l'Appel à projets 2021-2022[32] sont exemplaires des problématiques qui se posent aux chercheurs travaillant sur les données de la BnF : des enjeux liés la correction de l'OCR, à l'extraction des entités nommées et à la modélisation des sujets, afin de proposer des nouveaux modes de visualisation et d'exploration des corpus, comme c'est le cas du projet AGODA ; des enjeux liés à l'indexation et à la création automatisée de corpus numériques sur des notions relativement récentes, comme l'environnement, grâce à des techniques de traitement automatique du langage, avec le projet Gallica Env ; et enfin, des enjeux liés à la transcription automatique l'écriture manuscrite (HTR[33]) et la mise en place d'une chaîne de traitement adossée à des modèles, avec le projet Gallicorpora. À l'issue de ces projets, les équipes de recherche produisent des données enrichies (correction, annotation, modèles, etc.). La science ouverte prescrit de partager ces données et de les rendre accessibles selon les principes FAIR[34]. Pour la BnF, réintroduire ces enrichissements dans ses propres données reste encore un défi.

Conclusion

Qu'elles soient issues de la numérisation ou nativement numériques, toutes les données proposées par la BnF s'inscrivent dans le *continuum* de ses collections. Dans sa volonté d'ouvrir ses données aux chercheurs, l'établissement est confronté d'une part à la diversité et à la complexité technique qui parfois conditionne leur exploitation, et d'autre part à des contraintes juridiques, en particulier liées à sa mission de Dépôt légal. La BnF mobilise différents niveaux d'expertise pour répondre aux besoins des chercheurs et résoudre aux mieux les difficultés scientifiques et techniques qu'ils rencontrent, en proposant des services en ligne et un accueil sur place, en salle de recherche.

[31] ResPaDon (RÉSeau de PArtenaires pour l'analyse et l'exploration de DOnnées Numériques) est financé par le GIS Collex-Persée, porté par l'Université de Lille et la BnF, en partenariat avec Sciences Po Paris et le Campus Condorcet : https://respadon.hypotheses.org/

[32] En savoir plus : https://www.bnf.fr/fr/les-projets-de-recherche-bnf-datalab

[33] L'HTR (*Handwritten Text Recognition*) est une technologie de reconnaissance de l'écriture manuscrite à partir d'images numérisées.

[34] « Faciles à trouver », « Accessibles », « Interopérables » et « Réutilisables ».

En ce qui concerne les contraintes juridiques, l'exception TDM est une avancée majeure dans la prise en compte des usages de recherche. Dans le cadre de la rédaction du décret d'application de la loi Darcos du 30 décembre 2021 qui instaure le Dépôt légal des documents numériques, le ministère de la Culture porte une attention particulière aux modalités d'accès aux données, avec le souci de l'équilibre des territoires, ouvrant sans doute de nouvelles perspectives aux chercheurs.