Gallica, mine d'or et source de culture

Par **Arnaud BEAUFORT**
Directeur des services et des réseaux et Directeur général adjoint,
Bibliothèque nationale de France (BnF)

Introduction

Mise en ligne en 1997, Gallica, la bibliothèque numérique de la BnF et de ses partenaires, est passée au milieu des années 2000 d'une bibliothèque de l'honnête homme rassemblant les œuvres les plus significatives depuis l'Antiquité à une vaste collection universelle, encyclopédique et multiforme (livres mais aussi estampes, photographies, partitions, vidéos, marionnettes, médailles, etc.), promise à une croissance continue : elle franchissait le cap du million de documents en 2010, en comptait 5 millions en 2019 et 6 millions début 2020. Cette masse de documents est indexée par les grands moteurs de recherche et rencontre les intérêts tant des chercheurs que du grand public. Une étude d'usages approfondie menée en 2016 révèle le poids des recherches personnelles aux côtés des recherches plus académiques. Le site reçoit quelque 50 000 visites quotidiennes (ce chiffre passe à 70 000 en période de confinement). 50 % de ses contenus ont été consultés au moins une fois en 2019 (soit près de 3 millions de documents).

Les contenus librement accessibles de cette bibliothèque forment, en eux-mêmes, une collection organisée dont les contours sont sans cesse interrogés par l'évolution des modalités de diffusion, d'exploration et d'appropriation au sein d'un Web dans lequel les accès sont sans doute moins intuitifs, moins simples et limpides qu'il n'y paraît : la recherche d'information et la capacité à identifier des sources pertinentes relèvent davantage de compétences réelles que de l'intuition, et les choix faits en matière de transmission sont aussi importants que les contenus transmis eux-mêmes.

Son enrichissement est le fruit de nombreuses années de numérisation du patrimoine documentaire de la BnF auquel se sont adjoints d'importants gisements documentaires provenant d'autres bibliothèques : Gallica est une bibliothèque numérique collective qui entraîne avec elle à ce jour plus de 400 institutions.

Dans ce contexte, Gallica se définit davantage comme un dispositif évolutif multidimensionnel : le site gallica.bnf.fr est complété par des applications (pour Androïd et iOS), par une version intra-muros enrichie par des documents sous droits, par des marques blanches… La BnF développe ce dispositif selon trois pistes : l'amélioration constante du référencement, un positionnement en tant qu'acteur de confiance, et la capitalisation au profit d'un patrimoine numérique du XXI^{e} siècle en cours de constitution.

Référencer six millions de documents et rendre possible la trouvaille

Gallica, un Web dans le Web, ou la multiplication des portes d'entrée

Tout comme le Web en général, Gallica requiert des contenus disponibles et directement accessibles, des serveurs capables de gérer des connexions massives, un moteur apte à traiter des requêtes complexes en temps réel, et des interfaces familières et continues.

Parler aux moteurs

Moins d'un utilisateur sur cinq passe par la page d'accueil de la bibliothèque numérique, et 39 % des visites proviennent d'un moteur de recherche : Gallica est un réservoir dont le centre est potentiellement partout, et la circonférence, nulle part… Si la longue traîne, définie par un principe de rareté, est naturellement référencée par ces moteurs, il en va autrement des contenus les plus communs, les plus étudiés, les plus sujets aux homonymies. Excluant tout achat de liens sponsorisés, la BnF mène depuis dix ans une entreprise de traduction et de structuration de ses données dans des formats connus des moteurs, adoptant les principes du Web sémantique : cette traduction est au fondement de data.bnf.fr (2012), qui surplombe la variété des contenus numériques de la Bibliothèque, et dont 84 % de la fréquentation vient aujourd'hui des moteurs. Cette colonne vertébrale agit comme un pivot vers les autres sites de la BnF, elle hisse les contenus patrimoniaux des institutions aux premiers rangs dans les résultats des requêtes.

L'enjeu des API

Chacun des atomes de Gallica peut être trouvé non seulement sous une forme classique (inséré dans une page du site) mais également directement *via* une ligne de code (URL). De ce point de vue, le développement de l'API IIIF a grandement favorisé l'interopérabilité des contenus et leur dissémination : elle permet en effet d'appeler et de manipuler des contenus iconographiques [(1)] depuis un site pour les diffuser sur un autre sans obliger l'internaute à changer d'interface. Parmi les utilisateurs de Gallica qui ont choisi cette solution, on peut citer par exemple un site de généalogie équine [(2)], les Archives polaires françaises [(3)], ou encore un site commercial de vente de coques personnalisables pour smartphones [(4)].

S'appuyer sur des relais

Au-delà des progrès technologiques, ces usages relèvent globalement d'une politique de relais qui inclut aussi une présence active sur les réseaux sociaux et l'appui sur une communauté – les Gallicanautes – dont la configuration et l'implication évoquent les communautés des débuts du Web. Ils interagissent avec la BnF, partagent leurs trouvailles.

Les liens vers Gallica sur le site Wikipédia représentent également une source croissante de visites : les visites de Gallica en provenance de ce site ont progressé de 30 % entre 2018 et 2019.

Enfin, les marques blanches, qui permettent aux institutions de profiter de l'infrastructure de Gallica pour placer leurs collections numériques sur le chemin de leurs publics tout en enrichissant la collection collective de nouveaux documents, s'avèrent un véhicule de plus en plus plébiscité. De cette mutualisation de l'infrastructure de Gallica résulte un dispositif de coopération vertueux, qui répond aux enjeux de sobriété numérique et d'intelligence informationnelle. Dix marques blanches sont en ligne, dix sont en préparation [(5)].

Outiller pour donner à voir l'inédit

La communauté professionnelle mobilisée pour augmenter la diffusion des contenus sur le Web se rassemble également autour de l'enjeu de leur exploration. Elle s'appuie pour cela d'une part sur un dialogue constant avec les chercheurs, d'autre part sur les perspectives technologiques, en particulier sur l'évolution du moteur de recherche de Gallica (Cloud view, de Dassault systèmes).

(1) Voir https://iiif.io/
(2) https://www.pedigreequery.com/
(3) https://www.archives-polaires.fr/
(4) https://cover.boutique
(5) https://www.bnf.fr/fr/cooperation-autour-de-gallica#bnf-gallica-en-marque-blanche

Si les outils de fouille au cœur de la totalité des images de Gallica sont encore en cours de développement – ils progressent activement à la faveur de l'intelligence artificielle, et des expérimentations sont déjà disponibles sur certains fonds – les outils de fouille de texte permettent d'accompagner le travail d'exploration et de dépouillement des collections (c'est notamment le rôle du rapport de recherche [6]) et d'interroger les documents selon des modalités inédites (ainsi en est-il de la recherche par proximité [7] ou de l'analyse de vastes corpus [8]).

Gallica contribue à la création de nouveaux métiers, mais elle révolutionne aussi les activités et métiers déjà existants par la richesse des contenus auxquels elle donne accès et le gain de temps qu'elle permet : les doctorants, les chercheurs, les journalistes, les auteurs, ou les dessinateurs s'en font régulièrement l'écho, tels Pierre Lemaître, Daniel Schneidermann, Alain Rey, Benoît Peeters, Maylis de Kerangal…

Cependant, la multiplication des portes d'entrée, des sites et des outils, pas plus que le renouvellement du dialogue avec les utilisateurs, ne suffisent à une navigation fluide et tranquille au sein de ces contenus : il faut y placer des repères.

Gallica, source de culture et acteur de confiance sur le Web

Editorialiser, conseiller

Comment trouver tout de suite la bonne édition des *Pensées* de Pascal ? Où figurent les premiers textes de référence sur l'intelligence artificielle ou bien sur les problématiques climatiques ? Gallica n'apporte pas tant l'information en tant que telle que la source de l'information. Elle seconde l'esprit qui explore. Elle se décline à travers des dispositifs visant certains publics en particulier : les étudiants du BTP (site *Passerelle(s)* [9]), le jeune public (Gallicadabra [10]), etc. Elle met en valeur le travail de lecture et d'analyse des contenus réalisé par des Gallicanautes… La capillarité entre data.bnf.fr, les marques blanches et les réseaux sociaux fait naître un niveau affiné de conseils (« Gallica vous conseille » en tête des listes de résultats du site), de sélections, de billets de blog, etc., à travers lesquels la bibliothèque joue son rôle de référence, et qui complètent les traitements algorithmiques par de l'intelligence humaine.

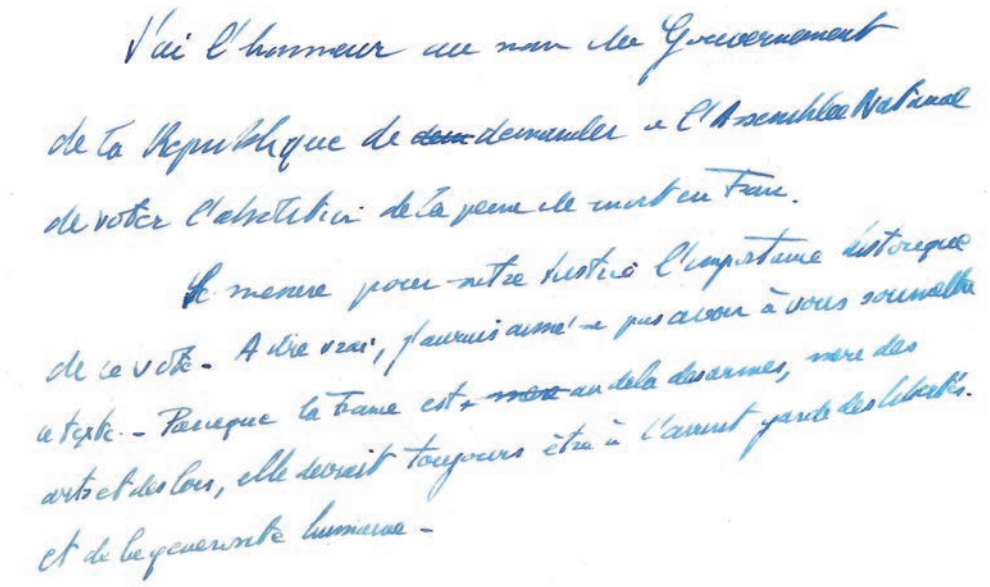
Ce travail de médiation s'intéresse en particulier aux trésors (du manuscrit du discours prononcé en 1981 par Robert Badinter contre la peine de mort [11] aux anciens numéros de la présente revue [12]…), aux essentiels (les éditions de référence des classiques de la littérature, du droit, de la politique…), et aux documents plus ou moins anciens qui entrent en résonance avec l'actualité.

Ethique du moteur interne et du site tout entier

En tant que source, en tant que service public, la galaxie Gallica respecte des principes de sécurité, de neutralité, d'ouverture, de légalité, et de stabilité.

Les contenus consultés lors de sessions précédentes, les parcours des internautes, ne sont pas utilisés pour modifier l'ordre des résultats dans la liste, ni pour enfermer un utilisateur dans un environnement déterminé par ses préférences. Le travail d'un chercheur qui, après une patiente recherche, aurait rapproché plusieurs titres, n'est pas divulgué *via* des suggestions de consultation aux autres internautes.

(6) https://c.bnf.fr/G97
(7) https://c.bnf.fr/Haa
(8) Voir en particulier Pierre-Carl LANGLAIS, *Reconstituer les genres romanesque sur Gallica : essai de classification automatisée de 1500 romans* (1815-1850), https://scoms.hypotheses.org/986
(9) http://passerelles.bnf.fr/
(10) https://c.bnf.fr/Hav
(11) https://c.bnf.fr/Hap
(12) Par exemple, cette série : https://c.bnf.fr/Has



Robert Badinter, manuscrit autographe du discours sur l'abolition de la peine de mort prononcé à l'Assemblée nationale le 17 septembre 1981 (https://c.bnf.fr/IaD)
© Bibliothèque nationale de France

Impliquée dans les débats nationaux et internationaux sur les questions juridiques, la BnF est parvenue en outre à une position d'équilibre en matière d'accès à ses ressources, *via* une formule articulant les enjeux de l'*open data*, ceux de l'*open science*, et les règles de l'univers marchand. Elle opère une distinction entre les données – entièrement libres – et les contenus numérisés, soumis à différents régimes. La politique de l'*open data* mise en œuvre dès 2014 s'inscrit dans une politique de l'État visant à faire émerger de nouveaux usages citoyens, voire de nouvelles opportunités économiques. En ce qui concerne la réutilisation des contenus numérisés librement accessibles en ligne, la BnF promeut la gratuité dans l'univers gratuit, éducatif et académique, et soumet à des redevances les acteurs de l'univers marchand.

Enfin, quiconque cite un document de Gallica, peut être sûr non seulement que le lien restera valable (la BnF va même jusqu'à proposer son propre outil pour raccourcir les URL, c.bnf.fr, qui garantit la pérennité des liens courts) mais que le contenu cité restera le même. Ainsi, l'étude d'usages de 2016 a fait apparaître une très forte progression des pratiques de consultation attentive en ligne [13] entre 2011 et 2016, qui rend inutile le téléchargement de sauvegarde.

Gallica, un terreau pour le patrimoine numérique du XXIe siècle

Le terrain exploré par la BnF dans le cadre des évolutions de Gallica concerne d'une certaine manière le patrimoine à venir : en tant que source, Gallica conduit à de nouvelles créations, et les dispositifs mis en place permettent d'étendre au-delà du Web les fonctionnalités et les usages qu'il a lui-même suscités.

L'ouverture des possibles et la co-construction

Outre son mérite de déplacer l'évaluation des résultats vers un champ davantage qualitatif que quantitatif, la focalisation sur les usages à laquelle nous invite le dossier de ce numéro permet de

(13) 66 % des répondants disent le faire « souvent » ou « à chaque fois », contre 31 % en 2011. https://multimedia-ext.bnf.fr/pdf/mettre_en_ligne_patrimoine_enquete.pdf

montrer combien une approche globale gagne à considérer non seulement les usages individuels, mais aussi ce qu'ils peuvent avoir de réticulaire : il s'agit moins d'une somme d'usages particuliers juxtaposés que d'interactions constantes et inspirantes. Le besoin de personnalisation – que la BnF assume à travers des dispositifs comme « Gallica vous conseille », le rapport de recherche, la numérisation à la demande, Adoptez un livre, etc. – ne s'envisage pas sans une dimension collaborative, sans un transport collectif.

L'univers du Web, qui associe étroitement la lecture et l'écriture, favorise cette conception des usages. A la BnF, par exemple, le hackathon de 2016 a donné naissance à Gallicarte, à présent implanté dans Gallica : un mois après le lancement de cette fonctionnalité de géolocalisation, 5000 points supplémentaires en bénéficiaient grâce à l'implication des internautes (14). L'usage peut donc aussi consister en de l'annotation, de la correction d'OCR, de la transcription de manuscrits… Comme l'expliquent Henri Verdier et Nicolas Colin, dans *L'âge de la multitude*, « il y aura presque toujours plus d'intelligence, plus de données, plus d'imagination et de créativité à l'extérieur qu'à l'intérieur d'une organisation » (15).

Pour cette raison, la BnF doit offrir, maintenir et encourager une multiplicité d'usages possibles, en commençant par les plus basiques : imprimer, télécharger, etc., sans négliger, en regard, les usages physiques. « Souvent qualifiés d'usagers "distants", les Gallicanautes nous rappellent que les offres physiques et numériques se nourrissent l'une de l'autre : 38 % disent avoir déjà fréquenté les espaces physiques de la BnF » en 2016 (16). La force de la bibliothèque est de ne pas dissocier ces formes de matérialité, et elle est sans doute l'un des lieux les plus indiqués pour garantir leur coexistence.

L'exploration du dépôt légal numérique

Enfin, la puissance de l'outil Gallica profite aux documents numériques protégés par la propriété intellectuelle et à la recherche documentaire en général : *Gallica intra-muros* donne aujourd'hui accès à un million de documents supplémentaires. À la faveur du dépôt légal numérique, elle sera demain bien plus importante que Gallica, et les chercheurs pourront aussi faire de la fouille sur place, dans le DataLab de la BnF. Le cercle commencé avec Gallica est vertueux : les outils seront d'autant plus puissants que la BnF recevra ce dépôt légal protégé. C'est un enjeu de supériorité informationnelle.

Conclusion

Si la dimension révolutionnaire de la bibliothèque numérique est à nuancer – elle est le reflet de pratiques transdisciplinaires et documentaires qui lui préexistent – la visibilité spécifique du patrimoine sur le Web et l'extension de ses usages recouvrent un triple enjeu: l'entraînement des intelligences artificielles sur des documents non contemporains (en particulier sur le patrimoine iconographique), l'intégration massive de documents francophones dans les corpus numériques, et l'éclairage des débats d'actualité.

Pour honorer ces enjeux, la BnF s'appuie à la fois sur l'engagement professionnel de ses personnels, sur une attention fine à ses publics, sur la technologie – de la fouille perfectionnée à l'interrogation des contenus en langage naturel – et sur la coopération, afin que le patrimoine en ligne soit véritablement partagé.

(14) https://gallica.bnf.fr/blog/21032018/gallicarte-arrive-dans-gallica?mode=desktop
(15) Armand Colin, 2015, p. 58
(16) https://gallica.bnf.fr/blog/10052017/resultats-de-lenquete-2016-aupres-des-usagers-de-gallica?mode=desktop