La découvrabilité des collections numériques patrimoniales sous l'angle des usages de Gallica

Arnaud Laborderie, Irène Bastard

La découvrabilité des collections numériques patrimoniales sous l'angle des usages de Gallica

Discoverability of digital heritage collections from the perspective of Gallica uses

Arnaud LABORDERIE (1), Irène BASTARD (2)

(1) Bibliothèque nationale de France (BnF), Département de la Coopération
Université Paris-VIII, Laboratoire Paragraphe
arnaud.laborderie@bnf.fr

(2) Bibliothèque nationale de France (BnF), Délégation à la Stratégie et à la recherche
irene.bastard@bnf.fr

Résumé. Avec 10 millions de documents dans Gallica, la bibliothèque numérique de la BnF et de ses partenaires franchit une frontière en termes d'accès avec des collections très diverses dont une grande part est relativement méconnue et partiellement consultée. Dans cet océan de ressources, la « découvrabilité » des contenus devient une gageure et un enjeu stratégique pour l'institution qui s'interroge sur la stratégie à mettre en œuvre. Toute réflexion sur la découvrabilité passe par une analyse préalable des usages : qui découvre les collections numérisées et comment se fait cette découverte ? Nous revisitons ici les études de publics menées à la BnF au regard de cette question, des compétences mobilisées dans cette activité de recherche et de la mise en œuvre de parcours de sérendipité.

Mots-clés. Bibliothèque numérique, découvrabilité, pratiques numériques, études de publics

Abstract. With 10 million documents in Gallica, the digital library of the BnF and its partners crosses a border in terms of access, with very diverse collections, a large part of which is relatively unknown and partially consulted. In this large amount of resources, the "discoverability" of content, that is to say the possibility of making it visible on the web and identified by Internet users, becomes a strategic challenge for the institution which is asks about the strategy to be implemented. This matter of usage is a precondition for work on "discoverability": who discovers digitized collections and how is this find out made? Here, we revisit the public studies carried out at the BnF relating to this question, the skills mobilized in this research activity and the implementation of serendipity pathways.

Keywords. Digital Library, discoverability, digital practices, audience studies.

1 Introduction

Selon la définition donnée par le ministère de la Culture, « la découvrabilité d'un contenu dans l'environnement numérique se réfère à sa disponibilité en ligne et à sa capacité à être repéré parmi un vaste ensemble d'autres contenus, en particulier par une personne qui n'en faisait pas précisément la recherche »[1]. Forgée au Canada en 2016[2], cette notion s'est d'abord appliquée au secteur audiovisuel avec pour enjeu de favoriser la visibilité des contenus canadiens sur le web dans un contexte de surabondance informationnelle et de prédominance d'acteurs américains (Desjardins, 2016). En 2020, un rapport ministériel franco-québécois réaffirmait l'importance de cet enjeu avec la mise en place d'une stratégie commune en France et au Québec (Min., 2020). De surcroît, lors des confinements liés à la crise sanitaire, les pratiques culturelles ont entièrement basculé en ligne, le plus souvent au bénéfice des contenus anglophones, à travers des plateformes en situation de quasi-monopole dans la plupart des secteurs culturels comme la librairie avec Amazon, la VOD avec Netflix ou Amazon Prime Video, la musique avec Apple Music, YouTube ou Spotify, etc. Concernant aujourd'hui toutes les industries culturelles, la découvrabilité ne dépend pas seulement des contenus eux-mêmes, mais aussi et surtout des stratégies mises en place par les grandes plateformes numériques à travers leurs algorithmes de recommandation. Inconnu il y a quelques années, le concept de découvrabilité se trouve désormais au cœur des préoccupations en matière de développement culturel.

Pour la Bibliothèque nationale de France (BnF), la question de la découvrabilité s'applique plus particulièrement aux collections patrimoniales numérisées et disponibles dans Gallica. La bibliothèque numérique de la BnF et de ses partenaires, une des plus grandes au monde, comprenait 8,9 millions de documents[3] fin 2021 (en augmentation annuelle de 10%). La presse représente une très large part de la collection (76%), suivie par les livres (9%), les images (6%), les monnaies et objets (4%), les manuscrits (2%), les cartes (1%), les partitions (1%), les enregistrements sonores et vidéos (1%)[4]. Par-delà l'hétérogénéité des contenus, la numérisation vise principalement des documents du domaine public et donc relativement anciens, du moins pour les accès à distance[5]. S'il n'est pas possible de consulter dans Gallica la dernière édition critique des *Fleurs du Mal* de Baudelaire, on peut y trouver un ePub de l'édition originale de 1857, les épreuves d'imprimerie corrigée de la main de l'auteur[6] ou encore les journaux où se publiaient les poèmes au fil du temps ainsi qu'une exposition virtuelle[7]. L'intérêt de ces contenus ne sera pas le même que

[1] https://www.culture.gouv.fr/Thematiques/Europe-et-international/Publications/Decouvrabilite-en-ligne-des-contenus-culturels-francophones

[2] À l'occasion du Sommet de la découvrabilité organisé par le Conseil de la radiodiffusion et des télécommunications canadiennes (CRTC).

[3] Partenaires moissonnés et collections sous droit consultables intramuros comprises.

[4] Selon le Rapport annuel d'activité 2021 de la BnF : https://www.bnf.fr/fr/bnf-rapport-dactivite-2021

[5] La version de Gallica consultable dans les emprises de la BnF, Gallica Intramuros, présente des ressources sous droit.

[6] https://gallica.bnf.fr/ark:/12148/btv1b86108314

[7] Baudelaire, la modernité mélancolique : http://expositions.bnf.fr/baudelaire/

l'on passe le baccalauréat, que l'on soit passionné de Baudelaire ou que l'on réalise un travail de recherche sur le poète. Il n'est donc pas possible de préjuger la valeur d'une ressource par rapport à une autre en dehors du cadre d'usage : la pertinence d'un contenu est relative et dépend de son contexte d'utilisation.

Ouverte en 1997, Gallica, en tant que service et collection, a beaucoup évolué, tant sur le plan documentaire et technique que des usages. Ce fut d'abord une bibliothèque savante, avec la constitution d'une collection de référence à destination de « l'honnête homme », autrement dit l'érudit. À partir de 2007, Gallica change de dimension avec une numérisation de masse à vocation encyclopédique et pluridisciplinaire. Dès 2015, Gallica développe une stratégie d'éditorialisation de ses collections et met en œuvre une médiation numérique tournée vers la conquête et la fidélisation de nouveaux publics, notamment via les réseaux sociaux (Bertrand et Degrange, 2021). En même temps, d'autres usages émergent, des pratiques scientifiques dans le domaine des humanités numériques avec les projets de recherche qui travaillent la collection en tant que données : les chercheurs constituent des corpus numériques pour faire de la fouille de textes et d'images, des analyses statistiques, etc. La BnF doit ainsi penser la stratégie de développement de sa bibliothèque numérique en fonction d'usages multiples et en perpétuelle construction.

À quoi correspond l'usage des ressources patrimoniales ? Qui consulte ces matériaux et dans quels buts ? Les usagers sont-ils en quête de trouvabilité (trouver ce qu'ils cherchent) ou de découvrabilité (découvrir ce qu'ils ne cherchent pas) ? L'objectif de cet article est de revisiter et d'articuler les résultats des diverses études conduites ces dix dernières années sur Gallica et les usages des ressources documentaires pour interroger la découvrabilité des contenus selon différents prismes d'observation.

2 Les documents consultés : une (très) longue traîne

Une première vision macroscopique de la consultation des contenus nous est fournie par l'analyse statistique des données de XiTi[8]. Pour mesurer l'audience de ses sites et documents, la BnF utilise cet outil qui identifie le nombre d'affichage d'une page ou d'un enchaînement de pages au cours d'une session. Gallica connaît une large audience[9], avec plus de 18,5 millions visites en 2021 et des pics à 70 000 visites par jour. On estime à 84% la part des visites de Gallica passant par la consultation d'un ou plusieurs documents, c'est-à-dire utilisant le visualiseur au moins une fois. Un document est caractérisé par un identifiant pérenne appelé ARK (*Archival Resource Key*)[10]. On peut alors établir l'audience de chaque ARK et l'audience globale rapportée à la collection dans

[8] XiTi est un outil de web analytics et de mesure d'audience qui permet de connaître le nombre d'ARK consultés, le nombre de visites comprenant la consultation d'un ou plusieurs ARK, le nombre de consultations comptabilisé par les ARK, et donc d'établir des moyennes de consultations par visite et par ARK : https://www.xiti.com/

[9] Par rapport à la fréquentation *in situ* de 521 914 visites sur tous les sites de la BnF en 2021 (en baisse de moitié à cause de la crise sanitaire).

[10] L'ARK (*Archival Resource Key*) est un format d'identifiant créé en 2001 par la California Digital Library (CDL) qui a vocation à identifier des ressources de tous types – physiques (échantillons destinés à une expérience scientifique, produits éditoriaux, etc.), numériques (livres numérisés, notices de catalogue, etc.) ou même immatériels (concepts). En savoir plus : https://www.bnf.fr/fr/lidentifiant-ark-archival-resource-key

son ensemble. Ainsi, en 2021, 3,1 millions d'ARK différents ont été consultés, ce qui représentant 48% des ARK disponibles[11]. Autrement dit, 52% de la collection n'est pas consultée et donc potentiellement inédite et « découvrable » pour les internautes. Que dit ce premier résultat ?

2.1 Des livres, mais pas seulement

Tout d'abord, cette part de documents non visités peut paraître importante et interroger les stratégies de numérisation : faut-il encore numériser plus si déjà plus de la moitié des ressources sont noyées dans l'océan du web ? Cette question se discute si l'on regarde la consultation *in situ*. Environ 500 000 documents ont été communiqués à des lecteurs dans les salles de recherche de la BnF en 2021. Cette consultation physique concerne une part bien moindre de l'ensemble de la collection de la BnF. La bibliothèque numérique démultiplie donc en ce sens la consultation des ressources de l'établissement.

Ensuite, si l'on regarde le type de documents des ARK consultés par rapport aux ARK proposés, les livres et la presse figurent loin devant les images et autres documents en volume[12]. En proportion, la presse ne représente que 20% des consultations alors qu'elle constitue 76% de la collection numérique, et 58% des exemplaires numérisés n'ont pas été consultés au cours de l'année 2021. En revanche, le petit volume de livres (9% de la collection) totalise 43% des consultations de Gallica, et ce sont 79% des livres consultables qui ont été consultés. Certains types de documents sont très consultés, voire dans leur quasi-intégralité : 97% des vidéos, 91% des cartes ou 82% des partitions consultables étaient effectivement consultées en 2021. Ces proportions résultent bien sûr de la politique de numérisation, puisque la mise en ligne des titres de presse ne répond pas seulement à des logiques d'usage mais aussi à un impératif de conservation des journaux qui se dégradent fortement au cours du temps. La diversité des types de documents consultés est bien plus large en ligne que *in situ* : 10% des communications *in situ* portent sur des documents antérieurs à 1950[13], 80% des documents communiqués en salle de recherche du site François-Mitterrand sont des monographies et environ 15% des périodiques. Parmi ceux-ci, les demandes *in situ* concernent principalement les périodiques publiés au cours de la seconde moitié du XX^{e} siècle : soit les consultations antérieures se font sur Gallica, soit les sujets de recherche traités en salle ne sont pas les mêmes que ceux traités en ligne.

2.2 Des contenus propres à chaque recherche

Le constat qui émerge de ces données souligne que la consultation des ressources sur Gallica suit une (très) longue traîne. L'ensemble des gallicanautes[14] consultent ou lisent des contenus très variés plutôt qu'ils ne se

[11] Les statistiques XiTi ne portent que sur 6,5 millions d'ARK disponibles en ligne sur Gallica, excluant les documents sous droits (consultables intramuros) et les documents moissonnés (consultables sur le site des partenaires).

[12] 7 545 043 visites de Gallica comprennent la consultation d'au moins un livre ; 3 537 217 visites comprennent la consultation d'au moins un périodique ; 1 245 055 visites comprennent la consultation d'au moins une image.

[13] Soit environ 50 000 communications sur l'année à comparer aux 50 000 visites moyennes par jour sur Gallica.

[14] Personne utilisant le site Gallica et qui participe à la diffusion des documents du site.

concentrent sur quelques titres vus par tous. Les documents les plus consultés de la bibliothèque numérique ne concentrent au mieux que 30 000 visites, soit moins de 1% du trafic, et il s'agit souvent de ressources mises en avant par une médiation comme un billet de blog ou une publication sur les réseaux sociaux.

	ARK	titre	auteur	type	Visites
1	bpt6k57440691	Liste officielle ... des prisonniers de guerre français : d'après les renseignements fournis par l'autorité militaire allemande : nom, date et lieu de naissance, unité / Centre national d'information sur les prisonniers de guerre	Centre national d'information sur les prisonniers de guerre. Auteur du texte	NUMÉRO	32 927
2	bpt6k54897701	[La grotte d'Isturitz / [signé Pierre Loti)]	Loti, Pierre (1850-1923). Auteur du texte	LIVRE	24 077
3	bpt6k851196s	La véritable magie noire ou Le secret des secrets. Manuscrit trouvé à Jerusalem, dans le sépulcre de Salomon. Contenant : 1.° Quarante-cinq talismans avec leurs gravures, ainsi que la manière de s'en servir, et leurs merveilleuses propriétés ; 2.° T	-	LIVRE	23 520
4	bpt6k6456365s	Résumé alphabétique des marques de porcelaines de toutes les fabriques européennes / par Ch. de Grollier et A. Popoff	Grollier, Charles de. Auteur du texte	LIVRE	17 173
5	bpt6k932314w	Le Manuscrit d'Arezzo, écrits inédits de saint Hilaire et pèlerinage d'une dame gauloise du IVe siècle aux Lieux Saints / par dom Fernand Cabrol, ...	Cabrol, Fernand (1855-1937). Auteur du texte	LIVRE	13 708
6	btv1b52500984v	Horae ad usum Romanum, dites Grandes Heures d'Anne de Bretagne	Bourdichon, Jean (1457 ?-1521). Enlumineur	manuscrit	11 280
7	btv1b72002410	Carte de l'Algérie divisée par tribus / par MM. E. Carette et Auguste Warnier ; Membres de la Commission Scientifique de l'Algérie	Carette, Ernest (1808-1890). Fonction indéterminée	Carte	10 994
8	bpt6k8502798	Cours complet de sténographie pratique s'apprenant sans maître et permettant de suivre la parole (système abréviatif greffé sur l'alphabet Duployé) / par F. Canton, ...	Canton, Firmin (1861-1936). Auteur du texte	LIVRE	10 870
9	bpt6k5454984c	Les 120 journées de Sodome, ou L'école du libertinage / par le marquis de Sade ; publié pour la première fois d'après le manuscrit original, avec des annotations scientifiques, par le Dr. Eugène Dühren	Sade, Donatien Alphonse François de (1740-1814). Auteur du texte	LIVRE	10 682
10	bpt6k50614b	[Dictionnaire universel, contenant généralement tous les mots françois tant vieux que modernes, et les termes de toutes les sciences et des arts... ([Reprod.]) / par feu Messire Antoine Furetière,...]	Furetière, Antoine (1619-1688). Auteur du texte	LIVRE	9 171
11	bpt6k1200508p	Grand dictionnaire de la philosophie / sous la dir. de Michel Blay	-	LIVRE	8 328
12	bpt6k5744130h	Liste officielle ... des prisonniers de guerre français : d'après les renseignements fournis par l'autorité militaire allemande : nom, date et lieu de naissance, unité / Centre national d'information sur les prisonniers de guerre	Centre national d'information sur les prisonniers de guerre. Auteur du texte	NUMÉRO	8 270
13	bpt6k28110w	Topographie militaire simplifiée : méthode nouvelle pour apprendre en peu de jours, sans le secours de la géométrie, à lever le terrain et à en figurer le relief, suivie de tous les renseignements nécessaires à l'exécution du dessin et à la rédaction	Roux, Louis (18..-19..? ; militaire). Auteur du texte	LIVRE	8 196
14	btv1b52000858n	Marco Polo, Livre des merveilles ; Odoric de Pordenone, Itinerarium de mirabilibus orientalium Tartarorum, traduit en français par Jean le Long ; Guillaume de Boldensele,Liber de quibusdam ultramarinis partibus et praecipue de Terra sancta, traduit e	Polo, Marco (1254-1324). Auteur du texte	manuscrit	7 816
15	btv1b53095291n	Carte générale de la France. Tableau d'assemblage, Nouvelle carte Qui Comprend les principaux Triangles qui servent à la description Géométrique de la France levée par ordre du Roy / Par Mess.rs Maraldi et Cassini de Thury, de l'Académie Royale des S	Maraldi, Giovanni Domenico (1709-1788). Cartographe	Carte	7 745

Illustration : les 15 ARK de Gallica les plus consultés en 2021

On est donc bien loin des mécanismes de concentration de l'attention que l'on observe dans les industries culturelles, où quelques titres phares attirent le succès et laissent dans l'ombre des milliers de titres. Chris Anderson proposait dès 2004 qu'internet servirait effectivement à entretenir la consommation des contenus de niche, en les rendant accessibles pour tout passionné ou curieux. Ce témoignage, collecté dans l'Observatoire des publics de la BnF[15], dit bien cette ouverture à des usagers et l'intérêt d'une démocratisation de l'accès aux ressources patrimoniales par le numérique :

> *« J'utilise les services de la BnF et de Gallica (livres, manuscrits, etc.) depuis le début (…), c'est pour moi la meilleure façon de se cultiver. Permet à tous de découvrir des archives que seuls quelques chercheurs pouvaient regarder avant ! c'est la démocratisation du savoir, vivement que toutes les archives anciennes et manuscrites soient numérisées ! bravo et merci » (Verbatim de l'Observatoire des publics 2020).*

Pourtant, au niveau des sessions et des individus, la découvrabilité pose des questions de modalités et de parcours. Au cours d'une visite, sont consultés en moyenne 1 à 3 documents ARK, ce qui peut sembler faible. Les témoignages disent en effet la difficulté à se repérer dans la masse de ressources :

> *« Je trouve vraiment extraordinaire de pouvoir disposer d'autant de ressources et ceci dans bien des domaines en ligne ! Je "m'égare" souvent sur le site en découvrant autant de connaissances qui me permettent de satisfaire mon insatiable curiosité. » (Ibid.).*

[15] Cette enquête, réalisée en ligne du 16 octobre au 30 novembre 2020, a collecté 5 198 réponses (cf. Bastard, 2022)

Comme si, en ligne, chacun reproduisait une exploration similaire à celle pratiquée sur les rayonnages des bibliothèques, c'est-à-dire saisir le livre d'à côté de celui visé. Mais quels artefacts matérialisent l'étagère en ligne et comment la navigation rend possible la découvrabilité des contenus ? Avant d'étudier ce que sont les parcours de navigation, revenons sur les individus qui cliquent et manient la souris derrière leur écran.

3 La curiosité des usagers comme socle de découvrabilité

Pour savoir qui sont ces visiteurs et mieux connaître leurs usages, il nous faut changer de points d'observation, ne pas partir des statistiques globales mais passer du côté des usagers. La nouvelle échelle d'analyse observe donc les milliers de répondants aux sondages en ligne et les quelques dizaines de chercheurs qui se sont prêtés à des enquêtes ethnographiques.

La BnF conduit régulièrement des enquêtes par questionnaire auprès de ses publics, en particulier en 2011 et 2016 (TMO, 2017) auprès des gallicanautes spécifiquement, et en 2020 auprès de l'ensemble des publics de l'établissement (Bastard, 2022). Les répondants, de l'ordre de 3 000 à 7 000 individus, créent un effet de loupe sur un public fidèle, habitué de Gallica, qui connaît bien l'interface et a pris le temps de répondre à de nombreuses questions car il souhaite participer à l'amélioration du service. Ce n'est donc pas forcément un public élogieux qui répond en ligne mais bien plus un public engagé et attaché à l'institution. Ces dispositifs d'enquête peinent à recruter des usagers occasionnels ou des publics touchés par des intermédiaires, comme les réseaux sociaux. Les jeunes sont d'ailleurs peu présents et 41% des gallicanautes dans l'Observatoire 2020 ont plus de 65 ans.

3.1 Qui sont les usagers des ressources de la BnF ? Profils types et rapports aux collections

Cette enquête a affiné notre connaissance des usagers des collections de la BnF en construisant des profils idéaux-typiques qui témoignent de l'hétérogénéité des publics de l'établissement. Quatre catégories se trouvent ainsi mises en sens selon le cadre dans lequel les individus utilisent les documents. Cette approche permet de faire saillir ce que Gallica transforme dans la consultation de ressources.

Un tiers des répondants s'inscrivent dans une démarche académique : ce sont des étudiants, doctorants, professionnels de la recherche en activité ou en retraite. Parmi eux, les plus jeunes utilisent moins les collections, probablement du fait de besoins documentaires encore circonscrits à leurs cours. Les chercheurs qui travaillent sur des documents patrimoniaux disent consulter, à distance, des ressources numérisées pour préparer une visite ou approfondir une observation, et, en salle, des manuscrits et originaux pour retrouver la matérialité de l'objet (Roustan, 2014). Ils utilisent plus souvent le Catalogue général que d'autres ; la recherche dans Gallica et celle dans le Catalogue pouvant alors répondre à des logiques ou des chantiers différents. Ces usagers sont plus attentifs aux évolutions techniques permises par les humanités numériques. Ils éprouvent néanmoins un besoin de médiation et d'interaction *:*

> *« (…) J'ai bénéficié aussi de la présence des bibliothécaires sur les réseaux sociaux. Ils ont partagé avec moi des documents que je*

n'aurais jamais trouvés moi-même, et ils m'ont permis de découvrir les coulisses des expositions relatives à mon sujet de recherche » (Verbatim de l'Observatoire 2020).

Un deuxième tiers de répondants fait émerger la figure des amateurs[16] de sciences et savoirs. Il s'agit de retraités (souvent) ou actifs déclarant avoir un centre d'intérêt, un hobby, une passion pour l'histoire, la littérature ou la généalogie, qui les conduit à la bibliothèque. Ces profils sont plus âgés, plus distants et bien souvent moins diplômés que les profils académiques. Ils sont très nombreux à n'utiliser que Gallica, naviguant dans les documents plutôt que dans les inventaires ou bibliographies, aussi du fait de la barrière physique et symbolique de l'accès aux salles de lecture. Leur consultation des ressources en ligne est tout aussi intense que celle des chercheurs, voire plus si les personnes profitent du temps libre de la retraite. La familiarité numérique reste pour certains un frein à la compréhension des outils, tout comme la compréhension de la construction des collections et la répartition entre les différentes institutions.

Le dernier tiers de répondants identifiés se divise en deux, entre les professionnels des bibliothèques, de l'art, de l'enseignement, usagers consultant le plus l'ensemble des outils autour de Gallica, et les visiteurs qui fréquentent l'établissement pour les activités culturelles, les espaces libres, etc.

Soulignons la part égale établie par l'Observatoire par entre les « académiques » et les « amateurs », ce qui renforce une dimension forte des usages de Gallica : les deux tiers des répondants sont engagés dans une démarche de construction des connaissances, bien plus que dans une démarche récréative. Néanmoins, dès la première enquête de 2011, les répondants montraient des usages mixtes, avec des objectifs moins sérieux, par exemple préparer un déguisement, jouer avec les enfants ou trouver la typographie d'un menu de mariage. Cette diversification des motifs de consultation est rendue possible par une maîtrise progressive de l'interface et des outils. La capacité à explorer des contenus s'acquiert au fil de la pratique plus que par des formations. Pour comprendre ces modes d'apprentissage, il faut encore descendre d'un niveau et suivre des individus dans les méandres de leur recherche.

3.2 Focus sur les chercheurs

Les pratiques des chercheurs font figure de référence pour comprendre comment se travaillent un rapport aux sources, dans la mesure où les amateurs se réfèrent souvent explicitement à ces pratiques (quitte à s'en distancier). On va donc chercher à comprendre les questions que se pose chaque individu sur l'intérêt d'un document et les dilemmes à résoudre, ainsi que l'apprentissage d'une bibliothèque numérique comme Gallica. En effet, l'intérêt de chaque

[16] Les « amateurs » constituent une figure travaillée depuis plus de 20 ans en sociologie de la culture, grâce aux travaux d'Olivier Donnat sur les pratiques amateurs (1996) et ceux de Patrice Flichy sur les communautés d'amateur constituées en ligne (2011). Ainsi, la désignation n'est pas péjorative mais souligne la noblesse d'une démarche désintéressée et conduite par goût et passion, avec souvent un niveau d'exigence très fort. Les « pro-ams » (en musique, en littérature, en histoire) s'appliquent ainsi certaines règles de pratiques des professionnels alors même que leur carrière n'en dépend pas.

document ne lui est pas intrinsèque : un entrefilet dans une publication semi-professionnelle du début du siècle peut n'avoir aucun intérêt pour un usager et constituer le clou des recherches de son voisin. À chaque clic, les internautes doivent estimer le temps à consacrer sur un document ou poursuivre la navigation. À partir de quels éléments se fait ce jugement ? Quels sont les artefacts qui y participent, en lieu et place des bibliothécaires disponibles en salle pour reformuler une question ?

Une méthode originale mise en œuvre par Valérie Beaudouin et Nicolas Rollet consiste à filmer une session de navigation d'un gallicanaute, avant de revenir en entretien avec lui sur son parcours pour expliciter les liens suivis et le cheminement de la pensée (Beaudouin *et al.*, 2017). Leurs travaux ont mis en évidence que les chercheurs pratiquent deux modes de recherche combinés. D'une part, des recherches ciblées, très précises, qui peuvent être profondes et se présentent plutôt en silos, souvent monotypes et disciplinaires : c'est « creuser un filon » (*Ibid.*). D'autre part, des recherches élargies, plus exploratoires, qui visent à parcourir un ensemble documentaire dans un domaine, à se constituer un corpus sur un sujet. De telles lectures dans un champ plus large sont plus superficielles. Ces deux modes de recherche ne s'opposent pas mais, au contraire, se combinent et peuvent se pratiquer simultanément. La sérendipité, la dérive exploratoire et le hasard font partie intégrante de la recherche documentaire et c'est un point fort de Gallica (par rapport aux salles de lectures physiques) que de pouvoir rebondir d'un document à l'autre. La vidéo-ethnographie a montré l'existence de très longues consultations d'une simple vue, portant jusqu'à 60 minutes le temps d'une session.

Gallica s'inscrit dans l'écosystème du chercheur, à la fois son espace de travail et ses autres outils, numériques ou non. Ses usages dépendent du contexte de production et des finalités de recherche : un cours, un article, un séminaire, une thèse, une référence à vérifier, etc. Contrairement à ce qu'on pourrait croire, les usages avancés de Gallica sont peu maîtrisés par les chercheurs. Ce n'est pas l'âge mais le temps passé sur Gallica qui détermine l'expertise. Beaudouin souligne ainsi que le rapport à Gallica est marqué par une double dimension de l'expertise liée, d'une part, à la recherche documentaire (le Catalogue comme point d'entrée, la constitution de corpus et la recherche systématique) et, d'autre part, à la maîtrise du numérique, à la capacité à explorer l'interface (utilisation experte du moteur de recherche, formulation pertinente puis affinage des requêtes et le fait d'avoir des routines de recherche).

Par rapport à la découvrabilité des contenus, on peut noter que les chercheurs parlent de « rebond » : il conviendrait alors de penser la découvrabilité comme un réseau entre des documents liés plutôt qu'uniquement par l'émergence spontanée d'un document comme dans une liste de résultats ou autre.

Ces pratiques individuelles nous éclairent sur les modalités d'appropriation de Gallica et la découvrabilité des contenus. Les usagers oscillent entre deux besoins contraires : la *pertinence*, lorsqu'il convient de trouver la ressource qui les intéressent, et la *sérendipité*, lorsqu'il s'agit d'élargir leur recherche dans une navigation plus exploratoire qui, potentiellement, leur permettra de découvrir un document rare, méconnu, peut-être capital. La découvrabilité interroge ainsi le « bruit » dans les résultats, qui est très relatif et dépend du mode de recherche adopté. Dès lors, peut-on concilier pertinence et sérendipité ? C'est tout l'enjeu des travaux qui, à partir des traces d'usage que sont les logs de connexion de

Gallica, visent à modéliser les parcours de navigation avec, en perspective, l'amélioration du moteur de recherche et le développement d'un algorithme de recommandation. Que nous apprennent les logs de Gallica ?

4 Traces d'usages et parcours de navigation

Les logs de connexion à un site web enregistrent toutes les requêtes et les clics des utilisateurs, ce qui permet de suivre leur activité en ligne. Ils restituent également le contexte de consultation (date, heure, temps, type d'appareil, etc.) et s'inscrivent dans des sessions, complétant ainsi les données d'audience (de type XiTi ou Google Analytics) par des données de navigation.

Initialement conservés par la BnF à des fins de sécurité et d'évaluation de la qualité de service, les logs de connexion aux serveurs de Gallica ont été anonymisés[17] pour des raisons juridiques et éthiques afin d'être mis à la disposition des chercheurs (Chevallier, 2018). La BnF a en effet résolument choisi de ne pas demander d'authentification pour accéder à ses contenus ni de profiler ses usagers à partir de leurs traces d'usage : les logs sont donc décorrélés des profils. Pour appréhender des milliards d'actions, les chercheurs doivent modéliser les données en appliquant des algorithmes de *clustering*[18].

Un premier jeu de logs d'usage de Gallica[19] a été constitué en 2016-2017[20]. Anonymisés, ils ont fait l'objet de plusieurs travaux d'équipes de recherche (Nouvellet *et al.*, Télécom ParisTech, 2017 ; Dumas-Primbault *et al.*, EPFL, 2021 ; Trabelsi, La Rochelle Université, 2022). Comment ces approches informent-elles sur la découvrabilité des ressources en regard des déclarations dans les enquêtes ?

4.1 Restituer la complexité des parcours de navigation

Adrien Nouvellet (Nouvellet *et al.*, 2017) a conduit une première analyse des logs de Gallica à l'aide des méthodes d'apprentissage automatique (*machine learning*) avec pour objectif d'identifier des parcours-types. Face aux 2,8 milliards d'actions enregistrées sur l'interface de Gallica en un peu plus d'un an, comment saisir et comprendre la complexité des parcours de navigation ? Pour répondre à cette question, le chercheur a mis au point un algorithme de classification (ou *clusterisation*) permettant de regrouper des sessions de Gallica présentant des similitudes dans l'enchaînement des actions.

L'analyse des logs a montré que la moitié des sessions sont très courtes, de moins de 12 secondes, ce qui signifie que les usagers venaient voir un document mais ne restaient pas sur le site de Gallica. Les consultations restent

[17] L'anonymisation porte sur l'adresse IP, remplacée par une clé hashée, selon la méthode de hachage SHA-2.

[18] Le partitionnement de données (ou *data clustering* en anglais) est une méthode en analyse des données qui vise à diviser un ensemble de données en différents « paquets » homogènes, partageant des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité informatique) que l'on définit en introduisant des mesures et classes de distance entre objets.

[19] Durée de collecte : 15 mois, de janvier 2016 à avril 2017, donnant lieu à 2 844 553 550 enregistrements et 1 126 787 556 requêtes Gallica.

[20] Dans le cadre du Bibli-Lab, partenariat de recherche entre la BnF et Télécom ParisTech. Voir le rapport de recherche : https://hal.archives-ouvertes.fr/hal-01709264

largement monotypes[21], avec une logique en « silos », à l'image de l'organisation des collections et des pratiques de recherche encore cloisonnées (Chevallier, 2018). Seules 3 % des sessions à plus de 5 documents explorent presque l'ensemble des types de documents.

Les *clusters* vérifient la très grande diversité des logiques de parcours dans Gallica : 53 % des sessions ne passent pas par la page d'accueil, ne téléchargent pas les documents et n'utilisent pas le moteur de recherche de Gallica (mais Google ou le Catalogue général). La modélisation révèle des parcours singuliers, très variés, avec des enchaînements d'actions qui s'affranchissent du parcours utilisateur fonctionnel pensé par les concepteurs (page d'accueil > moteur de recherche > consultation du document > téléchargement).

4.2 Modéliser des parcours de lecture en mêlant approches quantitatives et qualitatives

Ces travaux ont été poursuivis par Simon Dumas-Primbault qui a combiné une approche quantitative d'analyse de logs avec une approche qualitative, faite d'entretiens semi-directifs et de mises en situation de recherche (Dumas-Primbault *et al.*, 2021). Son objectif était de modéliser les parcours de recherche en identifiant un certain nombre de *pattern* pour les confronter ensuite aux usagers. Dans son travail, le chercheur a procédé à l'extraction de parcours de consultation, à leur modélisation par chaînes de Markov, puis à l'identification de cheminements types par analyse topologique des données. La modélisation de « parcours de lecture » comme une série d'au moins trois documents non redondants, consultés à moins de 60 mn d'intervalle, aboutissait, pour lui, à des objets d'étude pertinents, regroupés en six *clusters*.

Dumas-Primbault a voulu qualifier ces parcours en s'appuyant sur les métadonnées et les descripteurs des documents[22]. Ces parcours sont-ils conditionnés comme dans l'espace physique de la bibliothèque ? S'inscrivent-ils dans des logiques disciplinaires ou, au contraire, attestent-ils d'une certaine « curiosité » dans l'exploration des sources et documents de recherche ? Confirmant les résultats d'autres travaux, Dumas-Primbault observe une singularité de parcours qu'il parvient à qualifier et catégoriser : des parcours assez courts, convoquant peu de disciplines, des parcours longs à l'intérieur d'une même discipline, des parcours en étoile, attestant de recherches exploratoires, mais également des stratégies de navigation sur différentes temporalités, et enfin des ruptures dans les parcours, attestant peut-être de « trouvailles » dans la collection.

L'approche qualitative visait à objectiver la morphologie des parcours par des métaphores. Les entretiens ont confirmé la pluralité des régimes de recherche (phases de recherche, vitesse de recherche, divers pics de pratiques, etc.) et la diversité des modes de recherche (recherche dirigée / exploratoire), avec des pratiques hybrides et mixtes.

Par leur complémentarité, ces deux approches quantitatives et qualitatives permettent de documenter les stratégies de navigation développées

[21] C'est le cas de 45 % des sessions à plus de cinq documents, avec une prédominance des sessions ne consultant que des fascicules de presse ou des monographies.

[22] D'après la classification Dewey, cadre de classement de Gallica pour les livres et les périodiques.

par les chercheurs, témoignant de leur volonté de trouver « des embranchements originaux » à l'ère du numérique (*Ibid.*). Pour la BnF, malgré l'aporie de l'indexation dans Gallica et les biais afférents, les travaux de Simon Dumas-Primbault ouvrent une piste intéressante pour la recommandation personnalisée par discipline, thème ou sujet.

4.3 Focus sur un échantillonnage de logs : recherche précise et exploratoire à part égale

On change d'échelle avec la recherche doctorale de Marwa Trabelsi qui, afin d'étudier les comportements des utilisateurs sur la base de leurs traces, a produit un échantillonnage représentatif (Trabelsi, 2022). Sa démarche fut d'extraire des modèles grâce au *processus mining* (fouille de processus)[23]. Comme les autres chercheurs, Trabelsi s'est trouvée confrontée à un jeu de logs comprenant 2,8 milliards d'actions enregistrées pour 1,1 milliard de requêtes sur Gallica. Et comme eux, elle dut, dans un premier temps, regrouper les traces avant de procéder à la modélisation, autrement dit passer des traces brutes non structurées à des traces modélisées sous forme de session.

Trabelsi a procédé par regroupement des données (*clustering*) afin de transformer les logs en un ensemble d'activités (*event logs*). Pour cela, elle a établi une typologie de requêtes en supprimant celles non pertinentes (appels de *bot*, etc.) et en tachant de normaliser chaque requête par la définition de neuf activités. Trabelsi a encodé les traces en sous-séquences fréquentes afin d'en réduire le nombre, puis converti les traces en secteurs pour faire du *clustering*. Elle a ainsi organisé les requêtes en session et obtenu 1,7 million de traces. Celles-ci peuvent être très longues, jusqu'à 5 000 actions. Dans son modèle, Trabelsi s'est limité à 1 000 actions en 1h maximum afin de comprendre les opérations de feuilletage des documents (la plupart des traces comprennent moins de 1 000 événements). Elle a ainsi produit un échantillon de 20 000 traces d'une longueur maximale de 1 000 actions. Après le nettoyage des traces qui ne contiennent pas des sous-séquences fréquentes, Trabelsi obtient ainsi deux modèles[24] : recherches précises (dite *lookup*), qui représentent 29,4% des traces modélisées ; recherches exploratoires, qui représentent 33,6% des traces modélisées.

Trabelsi vient ainsi confirmer les deux modes de recherche observés par les études qualitatives en les estimant à part quasi-égales. Les 36,9% de traces ne contenant pas de sous-séquences fréquentes, non pertinentes pour la modélisation, nous intéressent néanmoins. Comment caractériser ces traces-là ? Est-ce le taux de rebond[25] ? Derrière ces traces, il y a certainement des *bots*, mais il y a aussi des lecteurs inattendus qui paraissent ne pas rechercher spécifiquement de contenus sur Gallica, mais tombent sur un document par hasard, peut-être au détour d'une liste de résultats sur Google, ou via les pages

[23] Recherche d'information et humanités numériques : une approche et des outils pour l'historien. https://www.theses.fr/233616950 et https://tel.archives-ouvertes.fr/tel-02009843

[24] Sur l'échantillon de 20 000 traces, nettoyé des 7 385 traces ne contenant pas des sous-séquences fréquentes, le premier modèle (recherche précise, *lookup*) contient 5887 traces et le deuxième modèle (recherche exploratoire) 6728 traces.

[25] Pourcentage de visiteurs qui accèdent à une page, puis quittent le site sans ne cliquer nulle part ni accéder à une autre page du même site.

de médiation de Gallica. Proposer des contenus à ces usagers fugaces est un enjeu fort de la recommandation.

5 Conclusion

La recherche documentaire n'est qu'une des modalités d'accès à l'information. La navigation de lien en lien en est une autre, passant par les contenus produits collectivement comme Wikipedia ou par une forme de signalement que seraient les publications des amis dans les réseaux sociaux. Sur les plateformes numériques de type Amazon ou Netflix, les consommateurs recourent de plus en plus souvent à la recommandation comme mode de découvrabilité des contenus.

Contrairement aux moteurs de recherche qui interrogent les données du web ou des catalogues à partir d'une requête, les algorithmes de recommandation poussent les contenus vers les consommateurs. S'ils reposent principalement sur l'analyse des traces, ils peuvent utiliser différentes approches, automatisées ou non (comme les listes établies par des experts). Ce sont des outils de concentration de l'attention et de navigation de proches en proches, potentiellement optimisés par les métadonnées et les techniques de référencement. On parle alors de « découvrabilité programmée », c'est-à-dire prédéterminée par les plateformes, à travers leur logique éditoriale, leur système de recommandation personnalisée ainsi que par les stratégies commerciales liées à leurs modèles d'affaires (Tchéhouali, 2020).

Si la découvrabilité peut s'appuyer sur ces algorithmes aux applications multiples (recherche d'information, marketing individualisé, etc.), il ne faudrait néanmoins ne pas réduire cette problématique à une question technique. Au contraire, il nous semble important de replacer la recommandation dans une relation entre usagers. Les études montrent combien la recommandation entre pairs, telle que mise en œuvre par de simples publications relationnelles sur les réseaux sociaux, prévaut sur la recommandation personnalisée. Le rôle des prescripteurs et la curation humaine (listes publiées par des amateurs ou par des experts, comme des pistes de lecture, des bibliographies, etc.) constituent aussi des références privilégiées, notamment pour les biens singuliers (Karpik, 2007).

Une des pistes pour Gallica serait de croiser recommandation de contenu par curation humaine et recommandation personnalisée par algorithme. Pour cela, la BnF peut s'appuyer sur différents types de données : non seulement les traces d'usages (logs), mais aussi les métadonnées descriptives des documents (permettant de recommander des contenus par similarité de notice), auxquelles peuvent s'ajouter des données d'enrichissements fournies par les bibliothécaires eux-mêmes ou générées par les utilisateurs. Ainsi s'agit-il d'appréhender la découvrabilité par recommandation sous trois aspects principaux :

- recommandation éditoriale, c'est-à-dire « humaine », par l'éditorialisation des collections et la médiation numérique ;
- recommandation sociale, notamment par les réseaux sociaux et les pratiques de partage, *like*, republication, etc.
- recommandation algorithmique, soit à partir des traces d'usage, soit « de contenu à contenu » (par indexation et similarité de contenu).

Au terme de notre propos, il convient de réinscrire la problématique de découvrabilité des contenus dans ces nouvelles frontières numériques que

dessinent, pour le public comme pour la recherche, la surabondance informationnelle et les données massives (*big data*) couplées à l'économie des algorithmes et de l'attention. Si les modalités de recommandation personnalisée par algorithme sont bien connues de la sphère commerciale (Amazon, Netflix etc.), elles restent encore inexplorées par les bibliothèques qui n'en sont pas moins confrontées à cette même problématique, tout en restant loin des mécanismes de concentration de l'attention.

Pour une bibliothèque numérique comme Gallica, la recommandation doit répondre à un besoin (faire découvrir les collections, mettre les documents en réseau, ne pas produire de bulles informationnelles, permettre à l'usager des trouvailles, lui proposer des chemins de traverse, etc.) et s'articuler à des pratiques informationnelles, celles de l'usager. Elle doit aussi répondre à une éthique, une déontologie de bibliothèque selon laquelle il n'est pas envisageable de profiler les utilisateurs pour faire de la recommandation personnalisée. C'est donc sous d'autres angles et leviers qu'il convient d'aborder cette question. Ces leviers sont liés à des cas d'usage réels, des expérimentations et des travaux de recherche sur lesquels la BnF s'appuie pour mettre en place une stratégie originale de découvrabilité de ses collections numérisées.

Bibliographie

Bastard, I. (2022). Les publics de la BnF. Synthèse de l'Observatoire 2020. Bibliothèque nationale de France, délégation à la Stratégie et à la Recherche. Référence : BnF-ADM-2022-008278-01.

Beaudouin, V., et Denis, J. (2014). *Observer et évaluer les usages de Gallica. Réflexion épistémologique et stratégique*. [Rapport de recherche]. BnF, Telecom Paris Tech. Disponible à https://shs.hal.science/halshs-01078530

Bertrand, S., Degrange, I. (2021). Gallica sur les réseaux sociaux numériques ou la réappropriation d'une mémoire collective. In *Balisages*, Villeurbanne, ENSSIB. Disponible à https://publications-prairial.fr/balisages/index.php?id=450

Bibliothèque nationale de France, rapport d'activité 2021. Disponible à https://www.bnf.fr/fr/bnf-rapport-dactivite-2021

Bibliothèque nationale de France, Tableau de bord de la communication en physique et en ligne 2021.

Bullich, V. et Guignard, T. (2014). Les plateformes de contenus numériques : une nouvelle intermédiation ? in Jeanpierre, Laurent et Roueff, Olivier (dir.), *La culture et ses intermédiaires : Dans les arts, le numérique et les industries créatives*. Paris, Éditions des archives contemporaines, 2014, pp. 201-210

Cardon, D. (2015). À *quoi rêvent les algorithmes : Nos vies à l'heure des big data*. Paris, Éditions du Seuil.

Chevallier, P. (2018). Les données au service de la connaissance des usages en ligne : l'exemple de l'analyse des logs de Gallica. In *Les Enjeux de l'information et de la communication*, 19(2), 57-67.

Chevallier, P. (2020). Usages et usagers d'une bibliothèque numérique : L'exemple de Gallica. In *L'édition en sciences humaines et sociales : Enjeux et défis*. Éditions de l'École des hautes études en sciences sociales.

Citton, Y. (2014). *L'économie de l'attention. Nouvel horizon du capitalisme ?* Paris, La Découverte.

Desjardins, D. (2016). *Découvrabilité. Volet 1 : Vers un cadre de référence commun*. Rapport financé par le Fonds des médias du Canada, l'Office national du film du Canada et Téléfilm Canada. Disponible à https://cmf-fmc.ca/fr/futur-et-medias/rapports-de-recherche/decouvrabilite-vers-un-cadre-de-reference-commun/

Dumas-Primbault, S., Baudry, J. Bert, J.-F., Kaabachi, B. (2021). Des embranchements originaux. Parcours de lecture et recherche exploratoire sur

Gallica. École polytechnique fédérale de Lausanne EPFL. Disponible à https://infoscience.epfl.ch/record/288902

Karpik, L. (2007). *L'économie des singularités*. Paris, Gallimard.

Min. (2020). *Mission franco-québécoise sur la découvrabilité en ligne des contenus culturel francophones*. Rapport du ministère de la Culture et des Communications du Québec, et ministère de la Culture de France, 2020. Disponible à https://www.culture.gouv.fr/Thematiques/Europe-et-international/Decouvrabilite-en-ligne-des-contenus-culturels-francophones

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*. [Rapport de recherche]. Télécom ParisTech, Bibliothèque nationale de France. En ligne : https://hal.archives-ouvertes.fr/hal-01709264

TMO (2017). Enquête auprès des usagers de la bibliothèque numérique Gallica, avril 2017, TMO Régions. Disponible à https://www.bnf.fr/sites/default/files/2020-12/mettre_en_ligne_patrimoine_enquete.pdf

Rioux, M. (2022). La découvrabilité va-t-elle devenir essentielle pour se frayer un chemin dans notre monde hyperconnecté ? in *Nectart*, 15, 22-30. Disponible à https://www.cairn.info/revue--2022-2-page-22.htm.

Rollet, N., Beaudouin, V., Garron, I. (2017). *Vidéo-ethnographie des usages de Gallica: Une exploration au plus près de l'activité*. Document numérique, 20, 97-114. Disponible à https://www.cairn.info/revue-document-numerique-2017-2-page-97.htm

Roustan, M. (2014). Pour un accès renouvelé aux collections : Une ethnographie de la BnF-site Richelieu et de ses publics. [Rapport de recherche] Bibliothèque nationale de France. 2013. Disponible à https://hal.science/hal-01405341

Suire, C. (2018). *Recherche d'information et humanités numériques : une approche et des outils pour l'historien*. Thèse de doctorat en informatique, La Rochelle Université. Disponible à https://www.theses.fr/233616950

Tchéhouali, D., Agbobli, C. (2020). État des lieux de la découvrabilité et de l'accès aux contenus culturels francophones sur Internet. In *Mission franco-québécoise sur la découvrabilité en ligne des contenus culturel francophones*. Rapport du ministère de la Culture et des Communications du Québec, et ministère de la Culture de France.

Trabelsi, M. (2022). *Modélisation des processus utilisateurs à partir des traces d'exécution, application aux systèmes d'information faiblement structurés*. Thèse de doctorat en informatique, La Rochelle Université.