# Les données au service de la connaissance des usages en ligne : l'exemple de l'analyse des logs de Gallica

Article inédit, mis en ligne le 15 novembre 2018.

## Philippe Chevallier

*Philippe Chevallier est adjoint au responsable de la coordination de la recherche à la Bibliothèque nationale de France. Docteur en philosophie, il est l'un des fondateurs du « Bibli-Lab », partenariat de recherche entre la Bibliothèque nationale de France et Télécom ParisTech sur les usages du patrimoine numérique des bibliothèques. Il a collaboré avec Anne Monjaret à l'édition de l'ouvrage collectif dirigé par Mélanie Roustan, La recherche dans les institutions patrimoniales : sources matérielles et ressources numériques (Presses de l'Enssib, 2016).*

### Plan de l'article



## Résumé

Connaître les usages d'une bibliothèque numérique comme Gallica nécessite de renouveler les dispositifs d'enquête traditionnellement utilisés par les bibliothèques, en explorant de manière semi-automatisée les données des serveurs. Cette exploration recourt à des modèles mathématiques qui rendent plus difficile le dialogue entre les chercheurs et les professionnels des bibliothèques. Un projet de recherche conduit par la Bibliothèque nationale de France et Télécom ParisTech sur les logs de Gallica témoigne de la possibilité d'inscrire la fouille de données dans un dialogue où les chercheurs et les professionnels s'efforcent de s'éclairer mutuellement sur les décisions à prendre pour conduire une analyse pertinente. Il met également en lumière l'importance de croiser les méthodes scientifiques pour comprendre les usages en ligne, aucune de celles-ci ne pouvant prétendre se suffire à elle-même.

### Mots clés

Bibliothèque numérique ; Usage ; Web ; Patrimoine ; Fouille de données ; Apprentissage automatique.

## Title

Data serving the understanding of online uses : the example of Gallica's log analysis

### Abstract

Knowing the uses of a digital library like Gallica requires to renew the survey methods traditionally used by the library, by semi-automatized exploration of server data. This exploration requires mathematical models that make the dialogue between researchers and library professionals more difficult. A research project conducted by the National Library of France and Télécom ParisTech on the Gallica logs shows the possibility of inscribing the data mining in a dialogue where researchers and professionals try to enlighten each other on the decisions to be made to conduct a relevant analysis. It also highlights the importance of combining scientific methods to understand online uses, none of which can claim to be self-sufficient.

### Keywords

Digital Library ; Use ; Web ; Heritage ; Data Mining ; Machine Learning.

## Título

Datos que sirven para el conocimiento de los usos en línea : el ejemplo del análisis de registro de Gallica

### Resumen

Conocer los usos de una biblioteca digital como Gallica requiere renovar los dispositivos de encuesta tradicionalmente utilizados por la biblioteca, mediante la exploración semiautomatizada de los datos del servidor. Esta exploración utiliza modelos matemáticos que dificultan el diálogo entre investigadores y profesionales de la biblioteca. Un proyecto de investigación llevado a cabo por la Biblioteca Nacional de Francia y Télécom ParisTech en los registros de Gallica muestra la posibilidad de inscribir la minería de datos en un diálogo en el que investigadores y profesionales intentan informarse mutuamente sobre las decisiones que deben tomarse para conducir un análisis relevante. También destaca la importancia de combinar métodos científicos para comprender los usos en línea, ninguno de los cuales puede afirmar ser autosuficiente.

### Palabras clave

Biblioteca Digital ; Uso ; Web ; Patrimonio ; Minería de Datos ; Aprendizaje Automático.

## Introduction

Face au rythme accéléré des innovations sociotechniques et à l'éclatement des pratiques numériques, construire une vision globale des usages d'une bibliothèque numérique comme Gallica (gallica.bnf.fr) constitue un véritable défi qui nécessite de mobiliser de nouvelles méthodes, en particulier celles permettant d'analyser des données de masse. Ces méthodes requièrent des compétences nouvelles par rapport à celles mobilisées traditionnellement dans l'univers de la connaissance des pratiques culturelles en général et des bibliothèques en particulier. C'est là une mutation importante. Le domaine des « études de publics », qui s'est développé dans des établissements tels les bibliothèques et musées depuis les années 1980, au confluent de plusieurs disciplines (sociologie, ethnologie, marketing, *etc.*), était traditionnellement le lieu d'un dialogue relativement équi-

libré entre les professionnels de ces établissements et les experts en charge de mener l'enquête, qu'ils soient chercheurs académiques ou consultants. Si des difficultés se présentaient, elles relevaient rarement de la compétence scientifique des interlocuteurs, les méthodes mobilisées, telles que l'enquête par questionnaire, l'observation ou l'entretien, étant facilement appropriables par tous. Cela ne signifiait bien entendu pas que tous pouvaient mener l'enquête, mais que tous pouvaient participer à la rédaction d'un questionnaire, comprendre l'enjeu des questions posées aux publics et se représenter le type de traitements réalisés sur les données collectées. Une telle participation au processus d'enquête est rendue plus difficile dans le cadre de la fouille de données (*data mining*), indispensable désormais pour la connaissance des usages en ligne. Un projet de recherche inédit sur les logs de connexion à Gallica, conduit dans le cadre d'un partenariat entre la Bibliothèque nationale de France (BnF) et Télécom ParisTech, témoigne pourtant de la possibilité d'inscrire la fouille de données dans un dialogue où chercheurs et professionnels s'efforcent de s'éclairer mutuellement sur les décisions à prendre pour conduire une analyse pertinente. Nous poserons dans un premier temps le cadre institutionnel et méthodologique dans lequel s'inscrit la connaissance des usages de Gallica, avant de présenter les apports de l'apprentissage automatique à cette connaissance et la manière dont ils peuvent être intégrés à la réflexion qu'une bibliothèque comme la BnF porte sur son offre en ligne.

## Un contexte institutionnel singulier

La BnF a une tradition ancienne d'enquêtes auprès des usagers de sa bibliothèque numérique Gallica, l'une des plus grandes librement accessibles sur le web, avec plus de quatre millions de documents patrimoniaux numérisés et près de quinze millions de visites par an. Il n'y a là rien d'original par rapport à tout établissement culturel qui développe des services en direction d'un public qu'il doit s'efforcer de connaître. Depuis les années 1980, « connaître ses publics » est devenu, comme le rappelle Olivier Donnat, une « *figure rhétorique obligée pour la plupart des responsables culturels* » (Donnat, 2016, p. 6). Mais si l'objet est en apparence simple, la manière dont il sera connu l'est beaucoup moins. Se pose la question non seulement des méthodes convoquées, mais aussi et surtout des compétences mobilisées et de leur place dans l'organisation : qui a l'initiative de commander une telle étude, à quelle fin et avec quelles formes d'implication dans sa réalisation ? Le champ d'activité que l'on nomme sans beaucoup de précision « étude de publics », d'une importance pourtant stratégique pour les établissements culturels, est rarement interrogé en ces termes, ceux de ses conditions de réalisation. Dans le cas des bibliothèques, le précieux guide *Mener l'enquête. Guide des études en bibliothèque* (Evans, 2011) laisse par exemple ouverte la question des acteurs des études, ceux-ci variant considérablement d'une institution à l'autre, ne serait-ce que pour des questions de moyens. Ce guide s'adresse en effet à « des non spécialistes qui travaillent dans des bibliothèques [...]. Que ces professionnels soient amenés à réaliser eux-mêmes un projet d'étude, en se transformant en enquêteurs pour l'occasion, ou que leur rôle se limite à accompagner ce projet en en confiant la réalisation à un tiers extérieur (prestataire spécialisé, stagiaire ou autre) » (Evans, 2011, p. 9). Deux cas de figure sont donc envisagés pour la réalisation des études : dans le premier cas, le bibliothécaire peut « à l'occasion » se faire lui-même enquêteur, en s'appropriant sans trop de difficulté quelques principes et outils élémentaires de la sociologie ; dans le second cas, la compétence est déléguée à un tiers – ce qui recouvre en fait deux situations très différentes selon qu'il existe ou non, au sein de l'organisation concernée, une instance en charge du pilotage des « études de publics ». Si celle-ci existe, il convient alors de réfléchir à ses compétences, ses missions et sa place dans l'organigramme : des éléments structurels qui ont des conséquences sur la manière de conduire des études.

Dans le cas de la BnF, les « études de publics » sont validées dans leur principe et accompagnées tout au long de leur réalisation – de leur instruction à la diffusion des résultats –, par une instance en charge de définir chaque année un programme d'études prospectives et d'évaluation sur les activités de la Bibliothèque : la délégation à la Stratégie et à la recherche, directement rattachée à la direction générale. Ce programme d'étude, validé en comité de direction, est instruit à partir des besoins exprimés par l'ensemble des services et des propositions de ladite délégation. Dans cette situation propre à la BnF, l'instance en charge de la connaissance des publics est donc distincte des instances opérationnelles qui ont besoin de cette connaissance pour faire évoluer leurs services, fidéliser ou conquérir de nouveaux publics. Une telle distinction ne se retrouve pas toujours dans les organisations qui préfèrent désormais lier la démarche de connaissance des publics à celle de leur conquête. Citons, à titre d'exemple, la Réunion des musées nationaux–Grand Palais (RMN-GP) qui a créé une « cellule études et marketing transversale » (Babault, Lévy-Fayolle, 2016, p. 61) réunissant les deux fonctions. Cette fusion organisationnelle peut avoir deux conséquences sur la connaissance des publics : soit celle-ci se limite aux publics les plus influents et les plus engagés, sur lesquels le marketing de l'offre cherche alors à s'appuyer ; soit elle est d'abord l'occasion de créer du lien avec les publics, collecter des données personnelles, communiquer sur de nouvelles offres, *etc.* Les études de publics connaissent ainsi depuis plusieurs décennies une double évolution qui se traduit tout d'abord par le passage d'une logique de la représentativité (connaître de la manière la plus rigoureuse possible qui sont les publics d'un service ou d'un produit) à une logique de l'influence (qui sont, au sein des publics, ceux qui ont le plus d'influence) ; et le passage ensuite de la connaissance de ses publics à la volonté de les faire participer à l'élaboration de l'offre (Beaudouin, Denis, 2014, p. 27).

Cette double évolution des études peut cependant réduire le temps consacré à l'activité de connaissance en elle-même, privilégiant le résultat rapide sur la fiabilité du chemin qui y conduit, et limitant le champ d'exploration des usages à quelques cibles ou signaux forts. Pour éviter ces écueils, la BnF a choisi depuis 2013 d'inscrire son activité de connaissance des publics en ligne dans le temps long d'un partenariat de recherche avec Télécom ParisTech, grande école spécialisée dans les technologies de l'information et des télécommunications. Ce temps long est nécessaire à l'apprivoisement mutuel de cultures professionnelles différentes et à l'élaboration de problématiques communes où les questions des chercheurs et les besoins d'un établissement comme la BnF espèrent trouver un terrain d'entente, sans décision déjà induite, dans un esprit de concertation et de respect mutuel. Cette rencontre n'est jamais acquise – le présent projet sur l'analyse des logs en témoigne – mais prend toujours la forme d'une négociation, portée par une conviction : le respect de l'autonomie de la recherche n'est pas un obstacle mais la condition de production de résultats fiables et utiles. Ce partenariat a donné naissance au Bibli-Lab, le « Laboratoire d'étude des usages du patrimoine numérique des bibliothèques », qui abrite des projets élaborés, financés et pilotés conjointement par la BnF et Télécom ParisTech, pour une durée qui varie entre six mois et trois ans, laissant une place importante à l'expérimentation.

## La connaissance des usages de Gallica : un problème de méthode

Comme la plupart des études internationales sur les bibliothèques numériques, les études sur Gallica se sont inscrites traditionnellement dans le paradigme de la sociologie des « usages », au sens où il s'agit de mesurer un écart entre un usage prescrit par un dispositif et son déplacement par l'appropriation d'un « usager » considéré comme un acteur relativement autonome (Jahjah, 2017 ; George, 2012).

Créée fin 1997 avec 20 000 documents accessibles, Gallica a fait l'objet dès 2003 d'un premier projet de recherche mené en partenariat avec France Télécom R&D. Croisant les méthodes de l'enquête en ligne, de l'analyse de trafic sur panel et de l'entretien qualitatif, ce projet mettait en avant une approche « *centrée utilisateur complète, qui reste rarement mise en œuvre dans les études d'usage d'envergure sur le Web* » (Assadi *et al.*, 2003, p. 2). Force est cependant de constater que ce sont les méthodes traditionnelles d'enquêtes qui ont dominé par la suite l'approche des usages de Gallica, avec cinq études aux objets divers conduites entre 2007 et 2011 : entretiens semi-directifs, individuels ou collectifs, et questionnaires en ligne. La tendance était alors, consciemment ou non, d'étudier les usagers de la bibliothèque numérique comme ceux de la bibliothèque physique : on entrerait dans Gallica comme on pousse la porte du site François-Mitterrand ou Richelieu. S'y retrouveraient par conséquent les mêmes finalités (lire, se documenter) et les mêmes chaines d'actions (chercher un document dans le catalogue, le consulter), dans une temporalité à peine resserrée (45 % des répondants à une enquête en ligne en 2011 déclaraient passer plus d'une demi-heure sur Gallica (GMV, 2011, p. 21)). Cette tendance à recourir aux méthodes traditionnelles était alors largement partagée si l'on en croit la recension réalisée par Carol Tenopir de quelques deux cent études sur les usages des bibliothèques numériques dans le monde anglo-saxon (Tenopir, 2003).

Si elles demeurent des sources d'information irremplaçables, ces méthodes traditionnelles se heurtent cependant au caractère à la fois majoritairement furtif (proche du « braconnage » : je regarde, je prends, je m'en vais) et entrelacé (multi-support, multi-activité) des usages d'une interface comme celle de Gallica, comme le vérifient à la fois l'analyse de l'audience de l'interface et l'observation de l'activité réelle (Rollet *et al.*, 2017). Quelques chiffres soulignent la nécessité de changer de référentiel par rapport aux usages d'une bibliothèque physique : 50 % des visites de Gallica font moins de 12 secondes ; 30 % ne font qu'une seule requête et seulement 8 % des sessions consultent plus de quatre documents uniques. Ces difficultés, bien connues de la sociologie du web (Beuscart *et al.*, 2016), limitent la portée des informations collectées *via* les méthodes d'enquête traditionnelles – qualitatives (entretiens) ou quantitative (questionnaires en ligne). Les usagers ont de plus en plus de mal à raconter, en entretien, le détail de leurs pratiques. Et leur propension à répondre à des enquêtes en ligne décroît à mesure qu'augmentent la rapidité et l'habitude avec lesquelles ils font des « choses » sur le web.

Ces lacunes des enquêtes traditionnelles – qui ne les invalident pas pour autant – avaient été relevées au lancement du Bibli-Lab par Valérie Beaudouin et Jérôme Denis, dans un travail sur les enjeux théoriques et méthodologiques de la connaissance des usages de Gallica (Beaudouin et Denis, 2014). Ces chercheurs préconisaient non pas une substitution mais une « articulation des outils et des formats d'investigation » (*Ibid.*, p. 26). Cette articulation touchait en fait de près à l'organisation même de la production d'information sur les publics en ligne au sein de la BnF. Ces publics sont en effet rendus présents à travers des formes d'engagement variées, qui se trouvent gérées ou analysées par différences services de la BnF : 1) connaissance sociologique des publics par des chargés d'étude, 2) suivi de l'audience par des informaticiens, 3) veille sur les réseaux sociaux par des *community manager*s, 4) relation-client par des responsables-produits. La conclusion était que la BnF n'avait pas tant un déficit d'informations sur ses publics en ligne – elle en avait même beaucoup plus que ce que chaque service isolé imaginait – , qu'un déficit d'articulation raisonnée entre ces informations. Rapprocher la connaissance sociologique des publics avec le suivi d'audience invitait en particulier à prendre en compte, aux côtés de la parole d'un nombre forcément limité d'usagers, l'ensemble des connexions à Gallica (15 millions de visites par an). Cela voulait dire par la même occasion décentrer l'analyse et les notions traditionnellement manipulées : passer de la notion anthropologiquement

et sociologiquement connotée d'» usager » à celle d'» usage ». Autrement dit, identifier la cohérence interne des usages, sans présumer de celle de l'usager supposé rationnel et régulier dans ses engagements – un même usager pouvant avoir plusieurs usages, à des moments différents et pour des motivations différentes (Beaudouin et Denis, 2014, p. 27).

La voie était ainsi ouverte pour une analyse inédite des logs de connexion à Gallica, en leur appliquant des méthodes de fouille de données, incluant de l'apprentissage automatique (*machine learning*). En effet, si le recours aux logs pour améliorer l'expérience-usager des bibliothèques numériques n'est pas une idée nouvelle, les études internationales que nous avons pu consulter en amont se limitent à des traitements statistiques simples, descriptifs, sans visée prédictive (*cf.* Ceccarelli *et al.*, 2011, et les travaux qu'ils citent dans ce domaine). La perspective neuve était ici de prendre l'ensemble des connexions à un service pour y repérer des similitudes dans les enchaînements d'événements : quand on fait a, quelle est la probabilité de faire *b*?

## Des informations inédites, complémentaires des autres approches

Les logs de connexion à un site web sont des fichiers qui contiennent toutes les requêtes reçues par les serveurs hébergeant le site. Dans le cas de Gallica, ils n'étaient initialement conservés par la BnF qu'à des fins de sécurité et d'évaluation de la qualité de service. Un certain nombre d'ajustements techniques ont donc dû être opérés pour la présente recherche : conservation de l'agent-utilisateur (*user-agent*) et du site afférent (*referer*) dans les logs, résolution d'un problème d'incomplétude des logs, anonymisation des données pour des raisons juridiques et éthiques. Après ces ajustements acceptés et réalisés par le département des systèmes d'information de la BnF – impliqué dès le début du projet –, les chercheurs de Télécom ParisTech ont pu disposer de fichiers contenant des informations importantes pour la connaissance des usages de Gallica : l'adresse I.P. (qui fait office d'identifiant unique d'une connexion, anonymisé pour le présent projet), la date et l'heure (à la seconde près) de la requête, la provenance de l'utilisateur (site référent), ou encore la requête http qui, dans le cas de l'appel d'un document de Gallica, contient son identifiant pérenne ARK. L'objectif, en traitant ces données de masse, n'était pas de connaître les usagers et leurs profils, accessibles seulement à travers des enquêtes déclaratives, mais d'identifier des types de navigation. De cette manière, la proposition de Valérie Beaudouin et Jérôme Denis (2014) de décentrer l'analyse de l'usager à l'usage, était bien respectée.

Durant quinze mois (avril 2016-juillet 2017), un chercheur en contrat postdoctoral, Adrien Nouvellet, encadré par quatre enseignants-chercheurs de Télécom ParisTech (Valérie Beaudouin, Florence d'Alché-Buc, Christophe Prieur et François Roueff) a eu en charge les missions suivantes : 1) nettoyer les logs de connexion (filtrage des robots internet) ; 2) structurer les logs de connexion en leur appliquant une notion de session (succession de requêtes d'une même adresse I.P. dont l'écart temporel n'excède pas 60 minutes) ; 3) définir les quatre actions-types utiles à l'analyse des sessions obtenues (consultation de la page d'accueil ; utilisation du moteur de recherche interne ; consultation d'un document dans l'interface ; téléchargement) ; 4) analyser les parcours d'usage en mettant au point un algorithme de partitionnement de données (ou *clusterisation*) qui permette de regrouper des sessions présentant des similitudes dans l'enchaînement et la durée des actions ; 5) analyser les types de documents consultés dans les sessions et les types d'action qui leur sont liés.

À un niveau général, cette analyse des logs a révélé une diversité de parcours que les entretiens et enquêtes en ligne – qui ne captent que les usagers les plus engagés –, ont

tendance à fortement réduire : poids des sessions très courtes dans l'audience globale, forte fluctuation temporelle des thèmes les plus consultés, variation du nombre d'actions effectuées dans Gallica en fonction de la provenance web de l'internaute, *etc.* Ce niveau simplement descriptif se révèle cependant assez vite pauvre en informations, similaires à ce que fournissent les requêtes pré-codées des outils de mesure d'audience. Aller au-delà des indicateurs de trafic, c'est être en mesure de passer d'un objectif général de connaissance, à la formulation souvent vague (mieux connaître les usages de l'offre, savoir ce qu'il conviendrait d'améliorer) et à la formulation de véritables hypothèses de recherche, qui touchent à la finalité même de l'offre et à ce que l'institution met en œuvre pour l'atteindre : quelles postures et actions sont visées par les développements réalisés ? Ce moment réflexif pour l'institution est forcément critique, tant l'usage présumé d'une offre ou d'un produit peut être incertain, contradictoire, résultat de compromis et de logiques hétérogènes. Dans cette difficulté à formuler ce que l'on veut savoir précisément, l'objet technique peut donner l'impression de s'être développé tout seul, sans « script » explicitement formulé. La difficulté du dialogue avec les chercheurs est ici redoublée par l'opacité pour un observateur extérieur des modèles mathématiques utilisés pour explorer les données de masse : le non expert a des difficultés à se représenter le type de résultats susceptibles d'être produits tant qu'il n'en a rien « vu » et peine donc à formuler une demande dont l'expert a malgré tout besoin pour avancer, organiser ses données et montrer « quelque chose » qui relance le dialogue.

« Mais que voulez-vous savoir, précisément ? » fut la question la plus fréquemment posée au démarrage de ce projet de recherche, poussant la BnF à formuler deux hypothèses de recherche, qu'il convenait de mettre à l'épreuve des faits :

1) Hypothèse de la diversité documentaire : une bibliothèque numérique comme Gallica favorise l'exploration des fonds numérisés dans toute leur diversité documentaire (presse, livres, manuscrits, estampes, photographies, *etc.*) et leur profondeur historique, contrairement à ce qui est constaté dans les salles de lecture où domine la consultation de documents très récents, dans une logique monodisciplinaire (Pardé, 2015) ;

2) Hypothèse de l'impact de la médiation : les actions de médiation (création de pages de médiation présentant des collections particulières, éditiorialisation de la page d'accueil et usage des réseaux sociaux) favorisent la découverte et l'exploration des fonds de Gallica.

Ces deux hypothèse ont conduit à enrichir les logs par les métadonnées descriptives des documents présentes dans l'entrepôt OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting : protocole de moissonnage des données). Ces métadonnées incluent, entre autres, la date d'édition et le type de document. Il a également été décidé de distinguer une action-type « consultation d'une page de médiation » (présentation des collections et blogue), au sein des actions identifiées comme pertinentes pour l'analyse – pour mémoire, les quatre initialement définies étaient : consultation de la page d'accueil ; utilisation du moteur de recherche interne ; consultation d'un document dans l'interface de Gallica ; téléchargement.

Ces deux hypothèses sont loin d'avoir été confirmées par les analyses, qui ont apporté à l'institution des informations précises et inédites. Nous nous contentons ici de résumer les principaux résultats, l'intégralité du rapport étant librement accessible (Nouvellet *et al.*, 2017).

## Hypothèse de la diversité documentaire : une faible diversité des types de documents consultés au sein d'une session

Ce résultat est une surprise : malgré les facilités d'exploration qu'offrent les interfaces du web et la sérendipité si souvent mise en avant, les consultations de Gallica restent large-

ment monotypes. 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. Ces sessions à plus de cinq documents, pourtant plus longues que la moyenne, reproduisent une logique de consultation en « silos », à l'image de l'organisation des collections et des pratiques de recherche encore cloisonnées, comme l'avait vérifié l'étude en 2012 des demandes de documents en Rez-de-jardin (Pardé, 2015). Un défi pour l'interface de Gallica sera de favoriser une logique de rebond d'un type à l'autre (par exemple : d'un manuscrit d'Apollinaire à l'écoute de sa voix). Seules 3 % des sessions à plus de 5 documents explorent presque l'ensemble des types de documents.

### Hypothèse de l'impact de la médiation : une médiation statique peu efficace

Si la conception d'un site web induit toujours une présomption d'usage « normal » (par exemple : page d'accueil > moteur interne > consultation de document), les clusters vérifient la très grande diversité des logiques de parcours dans Gallica. Dans le premier modèle de clusters obtenus, ne prenant en compte que la succession des actions, 53 % des sessions correspondent à des séquences de pure consultation de documents qui ne passent pas par la page d'accueil, ne téléchargent pas et n'utilisent pas le moteur de recherche. Il convient donc pour Gallica de ne pas concevoir la page d'accueil comme la porte d'entrée principale, mais de faire au contraire de toute page, une porte d'entrée dans le site avec des propositions de parcours. En l'état actuel, les pages de présentation des collections apparaissent dans un cluster unique, de faible amplitude (2,5 %), vérifiant que ces pages ne sont pas sur la route de la plupart des *gallicanautes* ; leur consultation obéit à un comportement distinct de tous les autres observés. En revanche, l'impact sur l'audience de Gallica des actions de médiation sur les réseaux sociaux est avéré et Facebook est bien représenté dans les sites référents. Une étude sur le type de lien vers Gallica présent dans les publications a d'ailleurs montré que celui-ci avait des conséquences sur le nombre de « clics » : ainsi, un lien actif dans l'image engendre 25 fois plus de visites sur Gallica qu'un lien actif dans le texte (avec indication de l'URL). Ce résultat a incité l'équipe en charge de la page Facebook à modifier ses modalités de publication.

## Des calculs en contexte de grande incertitude

La durée importante de cette recherche (15 mois) s'explique en partie par le temps nécessaire à tous les acteurs impliqués pour parvenir à se comprendre et assimiler le type de calcul susceptible d'être fait avec les données afin d'orienter la recherche. En effet, les opérations propres à la fouille de données, tel l'apprentissage automatique, ne se racontent pas facilement, contrairement à celles de l'ethnographie ou de la sociologie. En ce sens, le « *big data* » rend encore plus aigu le problème traditionnel du rapport entre recherche et institution, expertise et usage de cette expertise : il ne faudrait pas seulement «ouvrir les données», selon un leitmotiv contemporain qui se veut démocratique, il faudrait « ouvrir les modèles », au moins comprendre ce à quoi ils sont sensibles. Par exemple : l'ajout, dans une deuxième partie du travail, du facteur temps dans les clusters de sessions a été décisif pour l'interprétation finale, permettant de rapprocher des sessions au préalable éclatées mais qui présentent la même « silhouette » temporelle.

Sans doute, les difficultés éprouvées dans ce dialogue entre *data scientists* et professionnels des bibliothèques ne sont-elles que la résurgence d'un problème qui se pose depuis longtemps autour des sciences de la nature – où l'expert et le citoyen, le savant et le politique ont vu leurs positions respectives bouleversées par la spécialisation des savoirs. Mais ce

problème se posait traditionnellement sur le plan de l'éthique, ou encore du bon usage de la science (Moreau, 2008), alors qu'il se pose ici résolument au plan épistémologique, qui est celui de la manière dont la science procède, construit et valide des hypothèses. La grande richesse statistique des « *big data* » est en effet extrêmement sensible aux artefacts techniques et pauvres en explications : ce paradoxe est inhérent aux nouvelles méthodes sociologiques basées sur les traces (Beuscart *et al.*, 2016). Ce contexte de grande incertitude sur l'activité réellement décrite par les données est sans doute une clé du problème : la manipulation des données de masse a besoin de recourir à des informations tierces, détenues par d'autres acteurs, et le *data scientist* ne saurait travailler seul.

Conscient de ce besoin d'informations supplémentaires, le présent projet a reposé sur un dispositif original de gouvernance où compétences et méthodes ont systématiquement été croisées. Aux compétences initiales dans le domaine de la fouille de donnée (maîtrise des méthodes permettant d'explorer des données de masse et d'en extraire des informations susceptibles de décisions) se sont ajoutées des compétences de trois types : 1) informatiques (connaissance du format des données d'origine et de leurs lacunes éventuelles : manière dont elles ont été collectées et conservées) ; 2) bibliothéconomiques (connaissance des objets concernés par les traces d'usage : la «collection numérique » et sa place aujourd'hui dans les pratiques savantes) ; 3) celles enfin issues d'autres disciplines scientifiques analysant l'activité humaine, en partant du principe que les nouvelles approches ne se substituent pas aux plus traditionnelles (Beaudouin et Denis, 2014), et que les représentants des sciences dites de la « complexité » et ceux des sciences sociales doivent collaborer (Lazega et Prieur, 2014).

L'analyse quantitative a ainsi été enrichie dès le départ de données issues d'autres enquêtes et observations qualitatives, en particulier une série d'entretiens exploratoires (Beaudouin *et al.*, 2016) et une vidéo-ethnographie de *gallicanautes* (Rollet *et al.*, 2017). Cet enrichissement s'est révélé nécessaire tout d'abord en amont, pour construire les premières définitions nécessaires à la modélisation : les notions de « session » et d'» action ». La vidéo-ethnographie avait en effet vérifié l'existence de très longues consultations d'une simple vue, ce qui a conduit à revoir la définition d'une session par rapport à l'état de l'art : si celui-ci considère qu'une session sur un site web se termine lorsque le temps entre deux requêtes excède 30 minutes, il a été décidé de porter ce temps pour Gallica à 60 minutes. Autre exemple notable d'articulation entre les approches qualitatives et quantitatives : l'usage du moteur de recherche de Google pour chercher dans Gallica, y compris à l'intérieur d'une même session, avait été repéré au préalable dans les entretiens exploratoires (Beaudouin *et al.*, 2016, p. 20) ; il a donc été décidé, en cours de recherche, d'ajouter aux cinq actions caractéristiques de l'usage de Gallica (*cf. supra*) une sixième : « Je fais une recherche *via* Google » – ce qui souligne au passage le caractère nécessairement évolutif des définitions, et donc des modèles. En aval, ce croisement des méthodes et des informations a été tout aussi décisif pour donner du sens aux résultats et ne pas attribuer à des acteurs ce qui ne relève que du dispositif technique.

La condition pour qu'un tel dialogue entre compétences et méthodes soit fructueux est l'existence d'une compréhension partagée des analyses conduites. Si les arcanes des modèles n'ont pas vocation à être compris en profondeur par toutes les parties prenantes, il est indispensable de prévoir des dispositifs de visualisation des résultats et de retour aux données collectées, qui deviennent une ressource pour le travail collectif de validation et d'interprétation des résultats. Aux traitements statistiques et à leurs lignes de code, il faut être en mesure de faire correspondre des « images », l'imagination étant, comme le rappelle Kant, une faculté constitutive de la connaissance qui permet l'articulation du perçu et du conçu.

## Conclusion

L'application aux logs de Gallica des méthodes de fouille de données permet de compléter les méthodes traditionnelles en prenant en compte la très forte hétérogénéité des parcours, incluant en particulier les usages furtifs, faiblement motivés ou engagés, difficilement accessibles par les questionnaires en ligne ou les entretiens. Au-delà de ce que fournissent les statistiques de consultation descriptives, les méthodes d'apprentissage automatique permettent d'identifier dans la masse des connexions des régularités dans les chaînes d'actions et d'isoler des sessions-types dont on peut mesurer le poids au sein de l'ensemble des sessions. Par rapport à l'approche classique orientée « usagers », une telle recherche fait l'économie, au moins dans un premier temps, de « présupposés de cohérence, de rationalité et de constance chez les [usagers] », évitant de trop vite projeter sur les usages l'unité de «personnes», aux caractéristiques sociologiques et aux motivations précises (Beaudouin et Denis, p. 29). Elle permet ainsi de poser des constats macroscopiques solides sur ces usages, qui viennent dans certains cas contredire ceux imaginés par ses concepteurs (par exemple : la diversité des recherches documentaires) et même ceux déclarés par les usagers eux-mêmes (durée de leurs sessions, manière dont ils rationalisent à posteriori leurs stratégies de recherche, *etc.*).

À l'issue de cette recherche, le souhait pour l'institution de pérenniser certains traitements, afin de pouvoir les rejouer ponctuellement – dans la perspective en particulier de mesurer l'évolution des usages suite à une évolution majeure de l'interface –, a été exprimé et demeure à l'étude, ce qui pose aussitôt la question des compétences en science des données dont un établissement comme la BnF doit se doter. Le souhait de disposer d'outils fournissant des résultats simples et rapides pour la prise de décision ne doit cependant pas occulter l'importance d'inscrire la fouille de données dans une démarche de recherche où les éléments structurant chaque étape, en particulier les hypothèses que l'on formule et les définitions temporaires que l'on se donne, doivent être discutés et peuvent être à tout moment modifiés en fonction d'autres observations. L'importance de ces autres observations, afin de multiplier les points de vue sur le même objet et lui redonner le cas échéant son épaisseur sociale (Beuscart, 2017), est un appel à l'interdisciplinarité : la fouille de données n'a ici de sens que si elle est au service d'hypothèses qui sont construites ailleurs, en particulier *via* les enquêtes qualitatives auprès des usagers, conduites par les sociologues ou les ethnologues, mais aussi *via* les échanges avec les professionnels des bibliothèques qui sont aussi un point de contact avec les usagers et un lieu d'expertise sur les collections numériques. Il convient enfin de rappeler que certaines hypothèses ne peuvent être testées dans les modèles statistiques, comme par exemple la distinction entre des parcours de recherches ciblées et une exploration libre, très prégnante pour les usagers (Beaudouin *et al.*, 2016 ; Auray, 2017). C'est donc bien une complémentarité des approches qui doit être défendue.

## Références bibliographiques

Assadi Houssem, Beauvisage, Thomas ; Lupovici, Catherine ; Cloarec, Thierry (2003), « Users and Uses of Online Digital Libraries in France », p. 1-12, in Koch, Traugott ; Sølvberg Ingebord T. (dir.), *Research and Advanced Technology for Digital Libraries,* ECDL 2003, Lecture Notes in Computer Science, vol. 2769, Springer, Berlin, Heidelberg : Springer.

Auray, Nicolas (2017), *L'Alerte ou l'enquête : Une sociologie pragmatique du numérique,* coll. « Sciences sociales », Paris : Presses des Mines.

Babault, Gaëlle ; Lévy-Fayolle, Florence (2016), « Les usages en ligne autour des expositions du Grand Palais », *Culture et Recherche*, n° 134, p. 58-61.

Beaudouin, Valérie et Denis, Jérôme (2014), « Observer et évaluer les usages de Gallica. Réflexion épistémologique et stratégique », Rapport de recherche, BnF, Telecom ParisTech, [en ligne], Consulté le 16 février 2018, https://halshs.archives-ouvertes.fr/halshs-01078530/document.

Beaudouin, Valérie ; Garron, Isabelle ; Rollet, Nicolas (2016), « Je pars d'un sujet, je rebondis sur un autre : pratiques et usages des publics de Gallica », étude qualitative exploratoire, Rapport final de la phase 1 du projet « Mettre en ligne le patrimoine : transformation des usages, évolutions des savoirs ?, Bibliothèque nationale de France, labex Obvil, Télécom ParisTech, [en ligne], Consulté le 19 février 2018, https://hal.archives-ouvertes.fr/hal-01709238/document.

Beuscart, Jean-Samuel ; Dagiral Eric ; Parasie, Sylvain (2016), *Sociologie d'internet*, coll. « Cursus », Malakoff : Armand Colin.

Beuscart, Jean-Samuel (2017) « Des données du Web pour faire de la sociologie... du Web ? », in Menger, Pierre-Michel ; Paye, Simon (coord.), *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*, nouvelle édition, Paris : Collège de France, [en ligne], Consulté le 10 mai 2018, http://books.openedition.org/cdf/4987, DOI : 10.4000/books.cdf.4987.

Ceccarelli, Diego ; Gordea, Sergiu ; Lucchese, Claudio ; Nardini, Franco Maria ; Tolomei, Gabriele (2011), « Improving Europeana search experience using query logs », p. 384-395, in Gradmann, Stefan ; Borri, Francesca ; Meghini, Carlo ; Schuldt, Heiko (coord.), *Research and Advanced Technology for Digital Libraries*, International Conference on Theory and Practice of Digital Libraries, Berlin Heidelberg : Springer-Verlag.

Donnat, Olivier (2016), « La question du public, d'un siècle à l'autre », *Culture et Recherche*, n° 134, p. 6-8.

Evans, Christophe (coord.) (2011), *Mener l'enquête. Guide des études de publics en bibliothèque*, coll. « La boîte à outils », Villeurbanne : Presses de l'Enssib.

George, Éric (2012), «L'étude des usages des TIC au prisme de la recherche critique en communication », p. 25-63, in Vidal, Geneviève (coord.), *La sociologie des usages. Continuités et transformations*, coll. « Environnement et services numériques d'information », Cachan : Lavoisier.

GMV (2011), *Évaluation de l'usage et de la satisfaction de la bibliothèque numérique Gallica et perspectives d'évolution,* Rapport détaillé, Bibliothèque nationale de France.

Jahjah, Marc (2017), « État de l'art théorique, méthodologique et critique sur les usages et les pratiques », Rapport, Phase 1 du projet « Mettre en ligne le patrimoine : transformation des usages, évolutions des savoirs ?, Bibliothèque nationale de France, labex Obvil, Télécom ParisTech, [en ligne], Consulté le 19 février 2018, http://www.enssib.fr/bibliotheque-numerique/documents/67532-etat-de-l-art-theorique-methodologique-et-critique-sur-les-usages-et-les-pratiques.pdf.

Lazega, Emmanuel ; Prieur, Christophe (2014), « Sociologie néostructurale, disciplines sociales et systèmes complexes », *Revue Sciences/Lettres*, t. 2, [en ligne], Consulté le 17 février 2018, http://rsl.revues.org/455, DOI : 10.4000/rsl.455.

Moreau, Didier (2008), « Le Savant et la pédagogue : despotisme ou démocratie ? », p. 347-363, in Mustière, Philippe ; Fabre, Michel (coord.), *Jules Verne, le partage du savoir, Actes du colloque international,* École centrale, Nantes : Coiffard.

Nouvellet, Adrien ; Beaudouin, Valérie ; D'Alché-Buc, Florence ; Prieur, Christophe ; Roueff, François (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], Consulté le 28 février 2018, https://hal.archives-ouvertes.fr/hal-01709264.

Pardé, Thierry (2015), « Les usages documentaires dans une bibliothèque de Recherche », *Bulletin des bibliothèques de France (BBF)*, n° 5, p. 112-119, [en ligne], Consulté le 17 février 2018 : http://bbf.enssib.fr/consulter/bbf-2015-05-0112-002.

Rollet, Nicolas ; Beaudouin, Valérie ; Garron, Isabelle (2017), « Vidéo-ethnographie des usages de Gallica », Rapport final de la phase 2 du projet « Mettre en ligne le patrimoine : transformation des usages, évolutions des savoirs ?, Bibliothèque nationale de France, labex Obvil, Télécom ParisTech, [en ligne], Consulté le 19 février 2018, https://hal.archives-ouvertes.fr/hal-01709210.

Tenopir, Carole (2003), « Use and Users of Electronic Library Resources An Overview and Analysis of Recent Research Studies », *Council on Library and Information Resources, Washington DC,* [en ligne], Consulté le 10 mai 2018, https://clir.org/wp-content/uploads/sites/6/pub120.pdf.

TMO régions (2017), « Enquête auprès des usagers de la bibliothèque numérique Gallica », Rapport d'enquête, Bibliothèque nationale de France, [en ligne], Consulté le 17 février 2018, http://www.bnf.fr/documents/mettre_en_ligne_patrimoine_enquete.pdf