Bibliothèque nationale de France
délégation à la Stratégie et à la Recherche

Télécom ParisTech
département Image, données, signal
département Sciences économiques et sociales

TeraLab
Institut Mines-Télécom, GENES

Florence d'Alché-Buc (Télécom ParisTech), Valérie Beaudouin (Télécom ParisTech)
Emmanuelle Bermès (BnF), Philippe Chevallier (BnF)
Aude Le Moullec-Rieux (BnF), Adrien Nouvellet (Télécom ParisTech)
Christophe Prieur (Télécom ParisTech), François Roueff (Télécom ParisTech)

Analyse des logs de Gallica et de Data BnF et modélisation des comportements

15 décembre 2017

Contexte et méthode

Gallica (http://gallica.bnf.fr) est l'une des plus importantes bibliothèques numériques librement accessibles sur le web. Elle donne accès à 4,3 millions de documents de types variés : imprimés (livres, presse, revues, etc.), documents iconographiques (estampes, photographies, etc.), documents audiovisuels, manuscrits, cartes et plans, etc. Gallica reçoit environ 1,5 million de visites par mois.

Dans le cadre du Bibli-Lab, partenariat de recherche entre la BnF et Télécom ParisTech, et avec le soutien du TeraLab, a été conduite une analyse inédite des logs de connexion aux serveurs de Gallica, en leur appliquant des méthodes d'apprentissage automatique (*machine learning*). L'objectif n'était pas de connaître les usagers ni leurs profils mais, en partant des traces d'usages que sont les logs[1], d'identifier des parcours-types. Durant 15 mois (avril 2016-juillet 2017), un chercheur en contrat postdoctoral, Adrien Nouvellet, encadré par quatre enseignants-chercheurs de Télécom ParisTech[2], a mis au point un algorithme de partitionnement de données (ou *clusterisation*) permettant de regrouper des sessions de Gallica présentant des similitudes dans l'enchaînement et la durée des actions[3]. Les logs analysés couvraient des durées variables, allant d'une semaine à un mois, avec vérification systématique de la stabilité des modèles obtenus.

L'intérêt de ces méthodes d'apprentissage est de tirer profit de ce qui fragilise au contraire les méthodes traditionnelles de connaissance des usages : la masse des connexions (45 000 visites par jour sur Gallica). Cette masse interroge en effet la représentativité des enquêtes en ligne – représentatives d'abord et avant tout des internautes les plus engagés, mais pas de l'ensemble des internautes.

Malgré la puissance des algorithmes, l'apprentissage automatique requiert cependant un nombre important de décisions qui nécessitent de disposer d'autres sources de connaissance sur les usages et les usagers. Pour cette raison,

[1] Fichiers qui contiennent toutes les requêtes reçues par les serveurs. Entre autres informations importantes pour la connaissance des usages, le log contient : l'adresse I.P. (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), la requête http qui, dans le cas de l'appel d'un document de Gallica, contient son identifiant pérenne ARK.

[2] Florence d'Alché-Buc et François Roueff du département Image, données, signal (IDS) ; Valérie Beaudouin et Christophe Prieur du département Sciences économiques et sociales (SES).

[3] Les cinq « actions » identifiées dans les logs sont : consultation de la page d'accueil, consultation des pages de médiation (présentation des collections et blog), utilisation du moteur de recherche interne, consultation d'un document dans l'interface de Gallica, téléchargement.

le choix méthodologique fort a été ici de faire dialoguer les modèles statistiques avec les résultats issus d'autres approches (observations ethnographiques, entretiens, etc.[4]). Ce dialogue a permis à la fois de : *a)* fixer les paramètres de départ (définition d'une session et des actions élémentaires qui la composent) ; *b)* contrôler les modèles obtenus, extrêmement sensibles aux artefacts techniques ; *c)* proposer des premières clés d'interprétation.

L'intérêt du travail réalisé sur les logs de Gallica a convaincu la BnF et Télécom ParisTech d'ajouter à cette recherche un deuxième volet de quatre mois (juillet-novembre 2017) dédié aux logs de Data BnF, mais aussi aux parcours des internautes entre Gallica, Data BnF et BnF catalogue général. Mis en ligne en 2011, Data BnF permet de rendre les données de la BnF plus visibles des internautes et plus utiles sur le web. Ce site s'inscrit dans une démarche d'ouverture des données et d'adoption des standards du web sémantique. Indexable par les moteurs de recherche, Data BnF regroupe dans des pages HTML dédiées à un auteur, une œuvre, un thème, une date ou un lieu, une sélection de références bibliographiques et de liens vers des documents numériques issus des différentes bases de la BnF (catalogue général, BnF archives et manuscrits, Gallica, etc.).

Nous présentons ci-dessous les principaux résultats des deux volets de cette recherche. Le rapport complet de l'analyse des logs de Gallica, incluant une présentation de la méthodologie, est accessible dans HAL[5].

Volet 1 / Gallica : principaux résultats

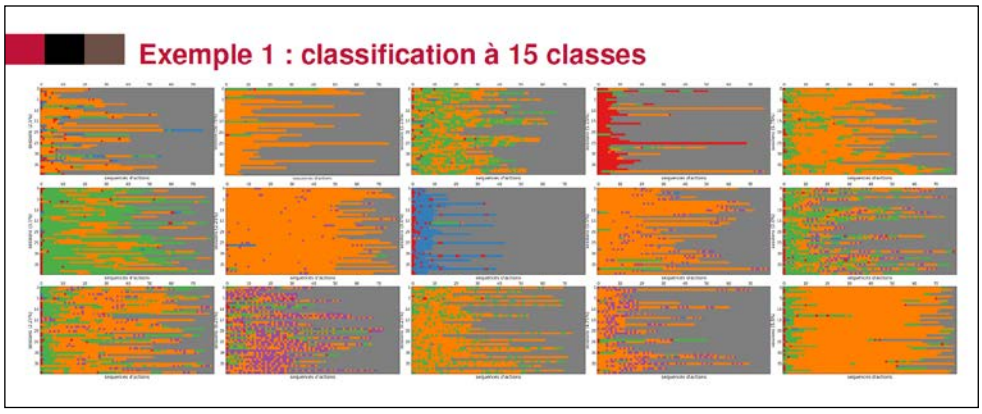
Importance des sessions très courtes dans l'audience de Gallica

Reflets de la vie du web, les usages furtifs de Gallica font la masse de l'audience : 50 % des visites font moins de 12 secondes ; 30 % ne font qu'une seule requête et seulement 8 % des sessions consultent plus de quatre documents uniques[6]. Par ailleurs, seule une session sur quatre a recours au moteur de recherche interne, ce qui a deux explications : *a)* une part importante des visites n'obéit pas à une logique d'exploration des fonds (mais à la seule consultation d'un document précis) ; *b)* le moteur de Google se substitue pour certains gallicanautes au moteur interne, non seulement au début, mais également à l'intérieur même d'une session, quand il s'agit de lancer une nouvelle recherche.

L'étude des sites référents (adresse de la page web à l'origine d'une requête) montre que la provenance web des gallicanautes a une influence sur la profondeur des sessions, mesurée ici par le nombre d'actions effectuées sur Gallica. Si Google, sans surprise, est le principal site référent quel que soit le nombre d'actions au sein d'une session, il n'en va pas de même de Facebook : le réseau social est le mieux représenté à l'origine des sessions faisant entre 2 et 4 actions (30 % des sessions). Contrairement aux idées reçues, les sessions en provenance de Facebook ne sont donc pas forcément des sessions de l'ordre du simple « clic ». Au-delà de quatre actions dans une session (40 % des sessions), apparaissent dans les sites référents, après Google, mais avant Wikipedia et Facebook, le domaine bnf.fr et un premier site thématique concernant la généalogie (Geneanet). Est ici vérifiée l'importance des sites thématiques – qui drainent une part importante de chercheurs-amateurs[7] –, à l'origine de consultations « profondes » de Gallica.

Des parcours « hors-normes »

Afin de disposer de modèles suffisamment riches, une méthode de formation de « *clusters* » (regroupement de sessions similaires) a été appliquée aux sessions de plus de cinq actions, soit seulement 35 % des sessions de Gallica.



[4] Cf. *infra*, pour les références aux autres études conduites en amont ou en parallèle à cette recherche.

[5] 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>.

[6] Une session courte ne doit pas pour autant être vue comme un échec : elle peut être le fait d'un internaute ayant aussitôt trouvé ce qu'il cherchait en cliquant sur un lien (hypertexte). 79 % des sessions mono-tâches sur Gallica consistent en un simple « appel » de document.

[7] Beaudouin V., Pehlivan Z. (2016), « Cartographie de la Grande Guerre sur le Web », Rapport final de la phase 2 du projet « Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre », en ligne : <https://hal.archives-ouvertes.fr/hal-01425600/document>.

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. Les pages de présentation des collections, quant à elles, apparaissent dans un unique cluster, 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.

L'intégration du facteur temps a permis d'offrir une vue plus juste des sessions au-delà des seules successions d'action : elle rapproche des sessions au préalable éclatées mais qui s'avèrent avoir la même « silhouette » temporelle. À titre d'exemple, le temps de consultation d'une même « vue » d'un document sur Gallica peut varier de 0 à 52 minutes, selon les conventions choisies pour l'analyse[8] : il convenait donc de ne pas assimiler des niveaux d'engagement aussi hétérogènes.

Avec le nouveau modèle, le cluster le plus important (42 %) rassemble des sessions qui font en moyenne 7 minutes et où alternent actions de consultation (3 minutes en moyenne) et moteur de recherche (2 minutes en moyenne), avec présence aléatoire et plus brève des autres actions – modèle donc plus proche de celui des concepteurs du site. D'autres clusters méritent attention, car ils permettent de quantifier des comportements à la fois simples et typiques : ainsi, dans 28 % des sessions domine l'activité de téléchargement, souvent associée à de la consultation ; 13 % des sessions alternent exclusivement actions de consultation et moteur de recherche, avec une part plus longue consacrée aux premières ; 3,5 % sont au contraire constitués de longues séquences de recherche (11 minutes) qui s'achèvent sur une brève consultation. Enfin, si le passage par la page d'accueil lors d'une session n'excède normalement pas les 30 secondes, 7 % des sessions y passent 4 minutes en moyenne (pas forcément en continu), au sein d'une succession variée d'actions, incluant même, de temps à autre, la consultation de pages de présentation des collections. Ce dernier résultat montre qu'il existe, même s'il est peu visible dans la masse des connexions, un public de la page d'accueil qui sait en tirer profit pour des consultations attentives et variées. Il importe en effet que les clusters les plus importants ne dissimulent pas des comportements non négligeables au plan quantitatif et intéressants pour le développement des publics et des usages.

Si l'on revient maintenant à l'ensemble des sessions, quel que soit leur nombre d'actions, on se rend compte que 4 % d'entre elles ne dépassent pas la page d'accueil de Gallica (soit 13 % des sessions à une action), n'entrent donc pas dans le site.

Une faible diversité des types de documents consultés au sein d'une session

Pour mesurer la diversité des usages documentaires, les logs ont été enrichis par les métadonnées des documents présentes dans l'entrepôt OAI[9]. Corrigeant sur ce point les réponses au questionnaire mis en ligne en 2016[10], on se rend compte que c'est la presse qui est le type de document le plus consulté, devant les monographies, puis les images, même si, rapporté au nombre de titres de presse disponibles dans Gallica, le ratio de la consultation est le plus faible.

C'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 largement monotypes. C'est le cas de 45 % des sessions à plus de 5 documents, avec prédominance bien entendu des sessions mono-fascicules de presse et mono-monographies. Les sessions analysées, 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[11]. Un défi pour l'interface de Gallica sera de favoriser une logique de rebond d'un type à l'autre (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.

Les actions les plus fréquentes au sein d'une session varient fortement en fonction des documents requêtés : lors d'une session dédiée exclusivement aux fascicules de presse, le gallicanaute passe la plus grande partie de son temps à consulter (*i.e.* à faire des requêtes contenant un ARK) et peu à télécharger, contrairement aux sessions dédiées aux manuscrits où le téléchargement occupe la majeure partie du temps de l'usager[12]. Ce type d'analyse peut aider à

---

[8] La vidéo-ethnographie conduite en parallèle à cette analyse a vérifié l'existence de très longues consultations d'une simple vue (cf. Rollet N., Beaudouin V., Garron I. (2017), « Vidéo-ethnographie des usages de Gallica », Rapport final de la phase 2 du projet « Mettre en ligne le patrimoine », en ligne : <https://hal.archives-ouvertes.fr/hal-01709210>), ce qui a conduit à revoir la définition d'une session : une session sur Gallica se termine lorsque le temps entre deux requêtes excède 60 minutes (là où l'état de l'art conseille, pour les autres services web, 10 minutes). Rappelons cependant que les logs ne disent rien de l'activité de l'usager hors-Gallica.

[9] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) est un protocole informatique développé par l'Open Archives Initiative afin d'échanger des métadonnées.

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

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

[12] Ce résultat est éclairé par une analyse qualitative conduite en parallèle : les personnes interviewées disent consulter la presse dans Gallica car les outils de lecture y sont très bien adaptés (en particulier les fonctions de zoom), tandis qu'ils téléchargent les monographies qui peuvent tout aussi bien être lues hors ligne (cf. Rollet N., Beaudouin V., Garron I. (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

prioriser les développements de fonctionnalités par type de documents. Des outils d'exploration des collections de presse, afin d'en faciliter le dépouillement (par exemple : la recherche par entités nommées), s'avèrent ainsi nécessaires.

Les réseaux sociaux : l'impact des modalités de publication sur Facebook sur l'audience directe

L'impact sur l'audience de Gallica des actions de médiation sur les réseaux sociaux est avéré et, comme nous l'avons vu, Facebook est bien représenté dans les sites référents. Les analyses l'ont vérifié sur un événement précis : le pic de consultation des ouvrages de Voltaire à l'occasion de l'anniversaire de sa mort en 2016 a pu être mis en regard du succès d'une publication sur la page Facebook de Gallica le même jour (1 055 likes, 568 partages). Par ailleurs, si le nombre de publications par semaine sur la page est resté relativement stable depuis sa création, le nombre des réactions par publication a significativement augmenté.

Une étude sur le type de lien vers Gallica présent dans les publications a 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.

Volet 2 / Data BnF et le catalogue général

Les habitudes horaires de consultation de Data BnF sont singulières : l'interface connaît des pics de consultation en journée (à 11h et 15-16h), alors que le pic pour Gallica se situe en soirée (entre 20h et 22h). On peut quantifier par ailleurs les visites approfondies de Data BnF : 800 sessions par jour font plus de 4 actions (5 % du total) et 250 sessions (1,5 %) consultent plus de 4 pages d'entité. « Auteur » est de loin l'entité la plus consultée, mais « Thèmes », « Spectacles » et surtout « Œuvres » (la deuxième la plus consultée) ont une proportion de consultation plus importante que leur présence dans Data BnF.

Sur un mois, 12 % des deux millions d'auteurs ont été consultés – chiffre important : la fonction de répartition des consultations des auteurs consultés au moins une fois sur un mois permet de formuler l'hypothèse, qu'il faudrait néanmoins vérifier sur une durée plus longues, que la plupart des auteurs devraient avoir été consultés après un laps de temps suffisant et qu'aucun ne domine véritablement les autres. Il n'y aurait donc pas de « zone noire » de Data BnF, dont le référencement de l'ensemble des pages sur le web semble bien fonctionner.

Comme souhaité initialement par le département des Métadonnées, ont été ensuite analysés les parcours entre Gallica, Data BnF et BnF catalogue général, afin de mieux comprendre comment les internautes accèdent aujourd'hui aux documents et métadonnées de la BnF sur le web : comment ils entrent dans les applications et circulent – ou non – de l'une à l'autre. Le poids considérable des consultations de Gallica *seul* a pour conséquence que seuls 4 % des sessions agrégées des trois sites sont multi-sites. Parmi ces sessions multi-sites, Data BnF est de loin le site le plus fréquemment présent (87 %), devant Gallica (69 %), preuve qu'il joue parfaitement son rôle de « pivot » entre les principaux services documentaires et bibliographiques de la BnF.

Data BnF est une porte d'entrée vers les documents de Gallica avant de l'être vers les données de BnF catalogue général : un rebond sur deux se fait entre Data BnF et Gallica, mais un rebond sur trois seulement entre Data BnF et le catalogue. L'accès direct aux documents est donc la première motivation de ceux qui, après Data BnF, poursuivent leur recherche dans le domaine bnf.fr. Ces rebonds se font parfois de manière inattendue depuis Gallica vers Data BnF, ou plus rarement, depuis BnF catalogue général, probablement grâce aux onglets de navigation du navigateur puisqu'il n'existe pas de liens directs depuis les documents de Gallica ou de BnF catalogue général vers Data BnF.

Le regroupement des sessions multi-sites en « *clusters* » vérifie que :

- Data BnF n'est pas seulement une porte d'entrée vers les autres services : pour plus d'une session multi-sites sur quatre, des sessions longues sur Data BnF précèdent des consultations conséquentes de Gallica (13 %) ou de BnF catalogue général (5 %) ;

- 10 % des sessions multi-sites tirent pleinement profit des trois services et peuvent être qualifiées, par convention, d'« expertes » ;

- Les sessions les plus « attendues » - une brève consultation de Data BnF qui bascule sur une longue consultation de BnF catalogue général (idée au départ du « pivot documentaire ») – ne représentent en revanche que 3,5 % des sessions multi-sites.

Enfin, comme pour Gallica, les sessions de Data BnF ne consultent en majorité (54 %) qu'un unique type d'entité, le plus souvent des « Auteurs ». Les sessions bi-types sont le plus souvent des sessions consultant durant un temps plus ou moins long un unique type, avant de basculer brièvement sur le second, mais sans véritable alternance de types différents au sein d'une session (des alternances « Auteurs » / « Œuvres » ne sont visibles que dans 1 % des sessions).

---

phase 1 du projet « Mettre en ligne le patrimoine », en ligne : <https://hal.archives-ouvertes.fr/hal-01709238>). L'enquête de TMO régions en 2016 a par ailleurs pointé l'augmentation de la consultation attentive dans les pratiques déclarées des gallicanautes (TMO régions, 2017).

Pistes de travail

Mesurer des évolutions d'usage au fil des évolutions de l'interface

Pour Gallica, l'ensemble des traitements a été réalisé sur des logs antérieurs au 15 juin 2016, soit avant la refonte de la page d'accueil. Le renouvellement de telles analyses permettrait de mesurer très directement l'impact des évolutions de l'interface sur les comportements.

Explorer des segments de publics importants pour le développement de l'interface

L'analyse des logs permet d'identifier et de quantifier des usages qui dénotent un fort niveau d'engagement (temps passé, nombre d'actions effectuées, diversité des types de documents consultés, etc.). Même s'ils sont minoritaires au regard de la masse, ceux-ci sont loin d'être négligeables en valeur absolue et reflètent des usages riches de l'interface qu'il serait intéressant de mieux comprendre. Par exemple : alors que la majorité des sessions de Gallica sont monotypes, 3 % des sessions à plus de 5 documents vont d'un type à l'autre : comment mieux comprendre ces parcours ? Quels actions ou sites référents caractérisent-ils ? Qu'est-ce qui favorise actuellement cette exploration de la diversité ?

Disposer d'une classification sémantique plus fine des documents

Pour Gallica : au-delà des types de document et des autres informations de l'OAI, utilisés jusqu'ici séparément (auteur, thème, etc., qui ne sont pas toujours pertinents ou renseignés pour tous les documents), il serait important de pouvoir analyser les thématiques de recherche et leurs évolutions au gré des parcours en disposant d'une représentation du contenu des documents consultés (par exemple, via un vecteur de nombres). Une première tentative de classification a été réalisée à partir des mots de la notice OAI ; une autre méthode a été également évoquée (« word embedding » appris à partir du corpus d'articles de Wikipédia) mais n'a pas pu être mise en œuvre faute de temps.

Pour Data BnF : afin de comprendre plus finement ce qui construit l'audience de Data BnF et vérifier ou non l'effet « longue traîne » de sa consultation, la notoriété des auteurs ou des œuvres pourraient être évaluées de manière à la fois conventionnelle et automatisée par le recours à des informations internes (nombre de documents « à propos de cet auteur ») ou externes (Wikipédia) au site.

Intégrer la notion d'utilisateur

Au-delà de la notion de « session », il conviendrait d'analyser des logiques d'utilisateurs en ayant recours à un cookie de domaine, à durée de vie limitée (par exemple 1 mois). En effet, un même utilisateur peut avoir des usages variés de nos interfaces, alterner des sessions brèves et des sessions longues, etc. Cette possibilité fera l'objet d'une instruction début 2018.

→ Ces travaux ultérieurs devront être accompagnés d'une réflexion sur les différentes manières de visualiser ces résultats (*datavisualisation*).