Bibliothèque publique d'information
Centre Pompidou

Analyse des logs de consultation d'Internet en accès libre à la Bpi : qu'apporte le Big Data ?

Premières observations : 2016-2017

Muriel Amar
Avec la collaboration de Dana Diminescu et Quentin Lobbé

2017

Service Études et Recherche – Bpi

Analyse des logs de consultation d'Internet en accès libre à la Bpi : qu'apporte le Big Data ?

Si l'accès à Internet en bibliothèque s'est banalisé ces quinze dernières années, il reste souvent bridé, accessible sous conditions ou avec limitation[1]. À la faveur d'un renouvellement de son offre, la Bibliothèque publique d'information (Bpi) a cherché à connaître les usages développés à partir de son parc d'une centaine de postes connectés au web. Avec plus de 20 000 réservations mensuelles et 2 millions de logs de connexion par jour, la Bpi entrait, sans le savoir, dans le Big Data. Retour d'expérience d'un partenariat de recherche avec l'équipe de Dana Diminescu et Quentin Lobbé à Télécom ParisTech.

Sommaire

Internet public à la Bpi (1995-2017) : déjà une longue histoire

L'internet public fait son entrée à la Bpi en juin 1995 avec 10 postes de consultation[2]. Les premières études de ce public indiquent que l'usager-type est un homme, jeune et bachelier, généralement issu de la filière scientifique, assidu de la bibliothèque et utilisateur des collections imprimées. C'est une époque où domine le versant extatique de cette nouvelle technologie ; pour reprendre l'intitulé d'un dossier du magazine *Manière de voir*, penser Internet en bibliothèque ce sera, à la fin des années 90, se situer entre l'extase et l'effroi. L'effroi, les controverses au sujet d'un internet public en bibliothèque marquent en particulier la décennie suivante pendant laquelle les bibliothèques s'interrogent sur le filtrage d'un objet qui ne lui appartient pas, dont elle n'est pas propriétaire[3]. Le périmètre des discussions est large, il touche à la fois la valeur documentaire ou pas d'internet et le respect de la liberté à l'information, particulièrement bousculée dans la période post -attentats du 11 septembre, dite du *Patriot Act* aux Etats-Unis. A la Bpi, à cette époque, l'offre consiste en 50 postes accessibles dans les espaces de la bibliothèque pour une durée limitée à 45 min (chats, messageries, son et audio sont interdits). Dans les années 2010, l'offre de la Bpi s'élargit avec l'accès wifi d'une part et une augmentation du nombre de postes en accès filaire d'autre part, toujours soumis à une limitation de durée. Le contexte de cette intensification est celui dit de la « fracture numérique », à plusieurs niveaux, celui de l'accès physique au réseau et celui de l'utilisation des ressources ; à partir de 2013 en effet, les services publics français s'engagent dans une dématérialisation massive, qui suppose des administrés autonomie et littéracie[4]. A cet égard, l'enquête Crédoc Baromètre numérique de 2016,

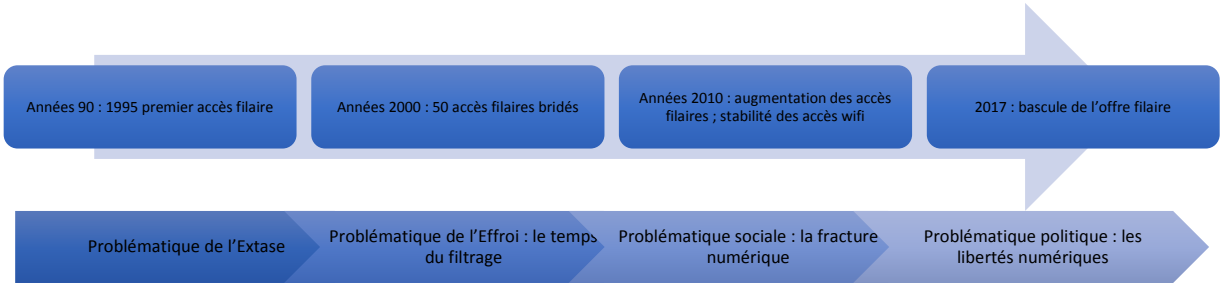
[1]. La description de l'offre d'accès Internet en médiathèque publique n'est pas toujours précise à cet égard (accès sous réserve d'inscription ou pas, pour une durée limitée ou un ensemble de services précis, etc.)

[2]. On compte à cette époque entre 200 et 300 000 internautes en France ; le premier cyber-café ouvre à Paris, quartier de la Bourse, en mai 1995, avec un tarif horaire d'accès à Internet de 60 francs. Source : Ina, Collection Journal télévisé Soir 3 (12 mai 1995).

[3]. Voir aussi sur ce point, « Babel ou le choix du caviste : la bibliothèque à l'heure du numérique », coord. Par C. Evans et F. Gaudet, in *Text-e : le texte à l'heure de l'Internet*. Paris : Bibliothèque publique d'information, 2003, coll. Etudes et recherche ; Yves Desrichard, *Cinquante ans de numérique en bibliothèque*, Paris : Editions du Cercle de la librairie, 2017, p. 53 et suiv.

[4]. Définie ainsi par l'Ocde : " Aptitude à comprendre et à utiliser l'information écrite dans la vie courante, à la maison, au travail et dans la collectivité en vue d'atteindre des buts personnels et d'étendre ses connaissances et ses capacités", in *La*

rapporte que « *environ 15% des adultes se sentent incapables d'entreprendre des démarches administratives en ligne* ». On peut signaler sur ce point que certaines bibliothèques d'Amérique du Nord vont plus loin en matière de lutte en faveur de la réduction de la fracture numérique en proposant le prêt de spots wifi, comme c'est le cas à Toronto ou à Chicago. Évoquons encore le retour récent de la problématique de la protection de la vie privée des usagers de l'Internet public en bibliothèque dans un contexte politique tenté par une politique accrue de surveillance : le Library Freedom Project préconise en effet de recourir à un relai TOR afin d'anonymiser et de protéger la vie personnelle des usagers qui utilisent la connexion Internet de l'établissement (et l'idée commence visiblement à faire son chemin en France), comme le signalait dans un récent billet Calimaq sur son blog SiLex.



Graphique 1 : Evolution de l'offre d'accès Internet à la Bpi mise en contexte

La nouvelle offre de la Bpi depuis le 7 avril 2017 est d'une autre nature et se caractérise par trois traits : d'une part, une augmentation conséquente du nombre de postes – multiplié par trois, d'autre part, la suppression d'une durée de connexion limitée et du système de réservation des postes qui accompagnait ce partage du temps de connexion ; et enfin, une offre élargie sur tous les postes, à la fois vers l'Internet public et vers ce que nous appelons l'Autre Internet, qui fédère les accès aux ressources pour lesquelles la bibliothèque souscrit des abonnements payants. C'est dans le cadre de cette bascule de l'offre – privilégiant l'internet public – que le service Etudes et recherche a enquêté sur la base de plusieurs dispositifs d'analyse[5] – des plus éprouvés comme l'enquête par observation et par entretiens – aux plus expérimentaux – comme le *Data mining*[6].

Log de navigation et/ou trace de consultation : le pari du *Data Mining*

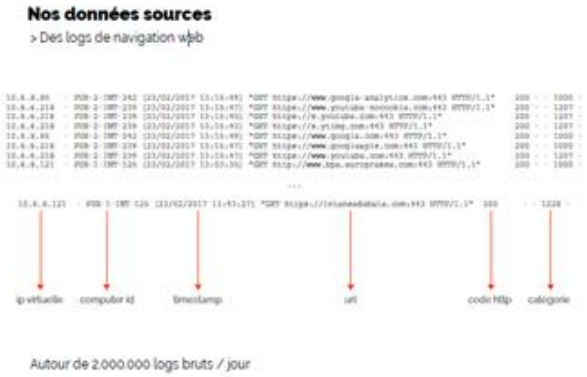
L'exploitation de données enregistrées quotidiennement par le serveur d'accès à Internet de la Bpi (journal de logs du proxy) peut-elle constituer une source d'information pour la connaissance de l'usage d'internet dans l'espace public de la bibliothèque ? Autrement dit, nous nous sommes demandés si des données techniques et fonctionnelles, nécessaires à l'affichage des pages web appelées par les Bpi-nautes, pouvaient fonctionner comme des traces de consultation, voire comme des indices de projet d'usage de l'Internet à la Bpi (voir graphique 2). Cette tentative, ou cette hypothèse, de transformation de données nécessaires à la réalisation d'une tâche en des connaissances sur cette tâche elle-même, constitue le coeur des travaux relevant de la discipline du

littératie à l'ère de l'information, 2000. Pour une approche critique, le numéro "New Literacy Studies", *Langage & Société*, n°133, septembre 2010, numéro dirigé par Béatrice Fraenkel et Aïssatou Mbodj.

[5]. Cette diversité de méthodes était en outre rendue nécessaire par les difficultés rencontrées à la Bpi de longue date pour enquêter auprès des usagers de l'offre Internet, voir sur ce point l'intervention de Christophe Evans aux Rencontres numériques des 27 et 28 mars 2017.

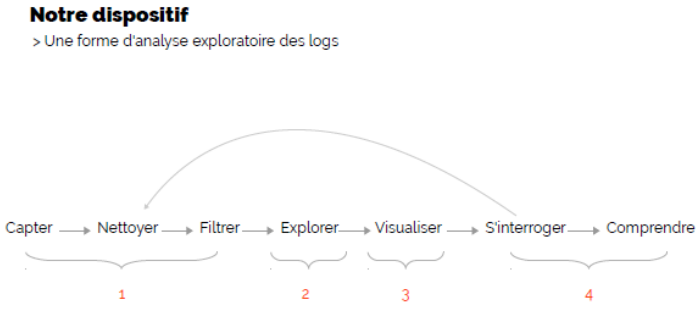
[6]. Une traduction pourrait être *fouille de données* ; quand les données sont issues du web, on parle plutôt de *Web Mining*. Les études menées depuis 20 ans à partir de l'exploitation des journaux de logs relèvent de la sous-branche du *Web Usage Mining*, Francony 2016. Nos collègues de la BnF ont également adapté la méthode de fouille de données à leur besoin de connaissance des pratiques en ligne développées à partir de la bibliothèque numérique Gallica, voir sur ce point les travaux de BibliLab.

Data mining. Nous nous sommes rapprochés de spécialistes de ce domaine travaillant à Paris Tech - Institut des sciences et des techniques de Paris - nous avons en particulier établi une convention de recherche avec une équipe mixte de sociologue et d'ingénieur Dana Diminescu et Quentin Lobbé pour explorer cette piste de travail.



Graphique 2 : extrait du journal de logs (source : Quentin Lobbé et Dana Diminescu)

Chaque jour, le serveur d'accès Internet de la Bpi enregistre deux millions environ de ligne de logs, ces deux millions constituent-elles autant de consultations de sites web ? Pas vraiment : pour une page qui s'affiche sur l'écran du Bpi-naute, ce sont jusqu'à cinq lignes de logs différentes qui sont écrites dans le journal du serveur : autrement dit, le journal de logs est très bruité du point de vue de l'analyse des usages par des lignes non directement visualisées par l'usager mais servant à afficher une page ou encore à la compter. Il faut donc filtrer les lignes du journal en isolant celles correspondant aux URL visualisées par le Bpi-naute ; il faut également transformer une ligne de logs en champs structurés d'informations interprétables : l'URL du domaine consulté bien sûr mais aussi l'étage du poste de consultation par exemple ou encore la date et l'heure de consultation : c'est l'aspect *mining* du *data mining*, rendant l'interprétation possible. Le dispositif d'analyse exploratoire des logs qui a été conçu par nos collègues de ParisTech comporte quatre phases dont seule la première est automatisable : elle permet de passer de 2 millions quotidiens de lignes de logs à 250 000 lignes de logs analysables par jour.



Graphique 3 : dispositif d'analyse des logs (source : Quentin Lobbé et Dana Diminescu)

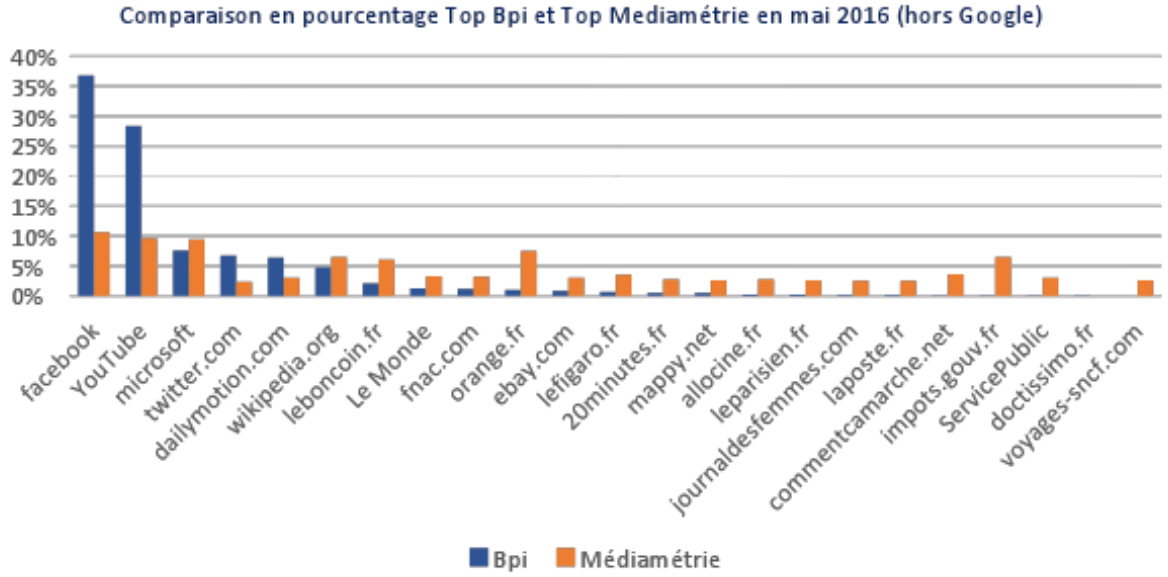
Un ensemble de trois briques logicielles libres[7] permet de nettoyer les données, de les filtrer, de les agréger par nom de domaine pour tenter d'isoler ce que peut être une trace de consultation menée

[7]. Logstash pour filtrer les logs (suppression des images, stats / css, des publicités grâce à la liste adblocker augmentée) et agréger les logs de connexion / Elasticsearch : moteur de recherche open source, structuration des logs par champs: | date | url | nom de domaine | extension (.com .fr .ru …) | catégorie du site (olfeo) | catégorie du site (bpi) | session id | étage | secteur | poste / Kibana : interface de visualisation à la volée.

sur un poste public de la Bpi. Evidemment, ces briques logicielles reposent sur des heuristiques qui ne sont pas aujourd'hui complètement stabilisées[8] : une première version du dispositif a été livrée en septembre 2016 qui nous a permis de traiter trois mois de logs de mars à mai 2016, dont l'analyse est présentée dans ce document. Dans cette première phase, nous avons été confrontés à plusieurs types de questions et difficultés. S'il paraît à première vue assez facile d'éliminer du journal de logs des lignes qui relèvent de traces nécessaires à la machine seulement, la frontière entre traces d'usage volontaires et traces d'usage involontaires devient vite très poreuse. C'est pourquoi une boucle rétroactive est indiquée dans ce schéma entre les phases d'interrogation et les heuristiques programmables de nettoyage ou de filtrage (voir graphique 3).

L' « Effet bibliothèque » : ni tout à fait le même, ni tout à fait un autre (Internet)

La première question que nous avons adressée aux données a permis de comparer les sites web consultés par les Bpi-nautes avec les sites web consultés par les internautes français en général. La question sous-jacente à cette comparaison consiste à savoir si un « effet bibliothèque » joue sur les modalités de consultations d'internet. Nous avons donc comparé la répartition des consultations à la Bpi sur les 50 premiers sites les plus consultés selon l'analyse menée par Médiamétrie dans son compte rendu des audiences d'Internet mensuel.



Graphique 4 : Les 20 sites les plus consultés selon Médiamétrie et selon le dispositif Bpi

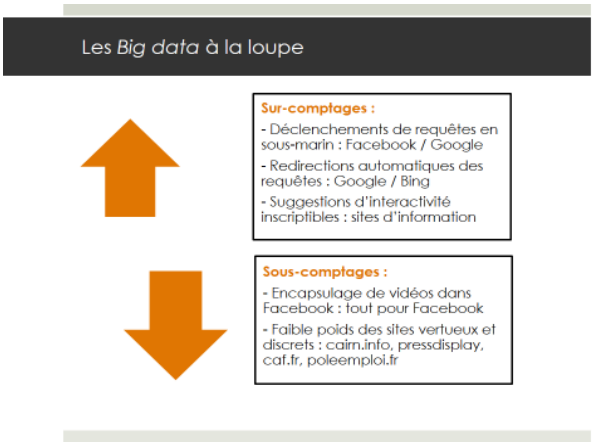
Ce graphique permet de visualiser la concentration très forte des consultations Bpi sur les deux plateformes Facebook et Youtube (3 fois plus consultées à la bibliothèque qu'à domicile)[9] ; de façon corollaire, est tout aussi notable la quasi-disparition, en bibliothèque, de la longue traîne en tant que phénomène de consultation faible : tous les autres sites font en effet l'objet de scores de consultations extrêmement faibles, sans doute sporadiques à l'échelle du mois. On sait bien sûr que les deux plateformes YouTube et Facebook constituent en elles-mêmes des espaces du web aux contenus et aux usages extrêmement hétérogènes à l'intérieur desquels à la fois des phénomènes de longue traîne

[8]. En informatique, on entend par heuristique « *un raisonnement formalisé de résolution de problème (représentable par une computation connue) que l'on tient pour plausible, mais non pour certain, et qui conduira à la détermination d'une solution satisfaisante du problème* », Jean-Louis Le Moigne, *La modélisation des systèmes complexes* (Dunot, 1991). Ces heuristiques dépendent également de la configuration de l'offre d'accès à Internet qui elle-même est encore en cours d'expérimentation à la Bpi (trois modalités d'offre ont été testées entre avril 2016 et octobre 2017).

[9]. Le graphique 4 ne rend pas compte des scores de Google, en tête bien sûr de tous les palmarès : ils auraient « écrasé » toutes les autres données, rendant l'analyse encore plus difficile

et d'ultra-consultations sont à l'œuvre. Le problème est que l'exploration plus fine des consultations à l'intérieur de ces plateformes n'est pas possible en raison du protocole sécurisé *https* utilisé par ces deux plateformes. Il ne serait d'ailleurs pas permis sans autorisation de la Cnil. Aujourd'hui, l'unité de consultation sur laquelle nous travaillons n'est pas l'URL consultée mais le nom de domaine concaténé et agrégé[10].

Reste qu'intrigués par la sur-fréquentation de ces deux plateformes à la Bpi, nous avons tenté d'observer ces *big data* à la loupe... en fait à la trace, devrait-on dire. Nous avons nous-même réalisé des parcours de consultation sur les postes internet de la Bpi et annoté très scrupuleusement tout ce que nous faisions : nous avons ensuite confronté notre carnet de bord avec les traces que nous retrouvions à travers notre application, et là, sans véritablement découvrir de scoop, nous avons tout de même mieux compris à quel point les *Big data*, la production massive des données de connexion, profitaient essentiellement aux *Big* producteurs de données. Des phénomènes de sur-comptage ont pu être observés : par exemple, des lignes de connexion à Facebook étaient enregistrées dans une session alors que pendant le parcours annoté ce site n'avait pas été appelé[11]. Plus les sites proposent une information élaborée de façon gratuite, comme les sites d'information, plus les données de connexion sont abondantes et indiquent que des requêtes se déclenchent en sous-marin, c'est-à-dire sans être effectivement formulées de façon directe et volontaire par l'usager. Inversement, des phénomènes de sous-comptages ont pu être observés : il y a un rapport de 1 à 40 entre la consultation du boncoin.fr et celle de la caf.fr, c'est-à-dire que 2 minutes de consultation du site leboncoin.fr génèrent 40 fois plus d'occurrences du site LeBonCoin que 2 minutes sur la caf.fr, polemploi, cairn ou encore pressdisplay.



Graphique 5 : synthèse des observations sur les parcours annotés révélant sur-comptages et sous-comptages

Ces observations ne bouleversent pas les grands équilibres et les abyssales différences de consultations mais elles permettent de comprendre que la modestie du volume des consultations enregistré par les sites publics tient pour une part à des modes de développement des bases de données et des interfaces faiblement sensibles voire insensibles à l'enjeu du marquage de leur présence sur le web. On comprend aussi que les *big data* ne sont pas forcément pertinentes pour analyser l'usage de tous les types de sites, de tous les types de parcours web pourrait-on dire.

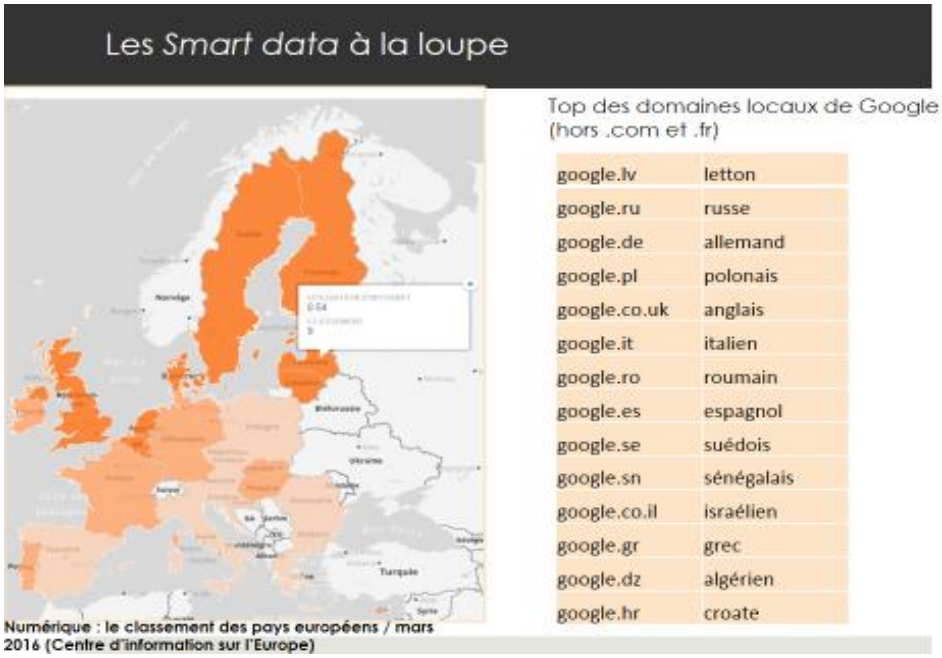
[10]. Ce type de traitement avait déjà été testé en mis en oeuvre en 2004 par Matthieu Renault sur les logs Bpi, mis à jour en 2017 par Chaïma Berrachedi, sous la direction de Pierre Senellart et Quentin Lobbé.
[11]. Inscription relevant de techniques de fabrication d'audience artificielle, de type *likejacking* (littéralement : détournement de « J'aime »).

C'est pourquoi nous avons exploré un autre angle d'analyse sans nous laisser éblouir par l'ultra-consultation de Facebook et YouTube à la Bpi[12] ; nous avons cherché à saisir l'éventuelle diversité qualitative des consultations, en dehors des seuls critères quantitatifs. Pour cela, un examen par niches a été conduit, en particulier sur deux segments : celui des sites de rencontres et celui des sites d'information non-francophones.

Graphique 6 : synthèse des observations sur deux requêtes - sites de rencontres et sites d'information

Notre première surprise a été de voir se hisser dans le top des sites de rencontres consultés à la Bpi non pas les sites de rencontres *mainstream* comme *meetic* ou *adopteunmec.com* mais des sites de rencontres plus marqués par la recherche d'endogamie come *inshalla.com* ou *afrointroductions*. La même sur-représentation dans les consultations Bpi de l'intérêt porté à l'actualité non-francophone est également observable dans le segment constitué par les sites d'informations hors .fr et .com. On observe une intéressante diversité linguistique dont le tableau ci-dessus donne un aperçu avec des communautés à faible nombre de locuteurs comme pour le letton (graphique 6). Ces signaux faibles nous ont engagé vers l'exploration de *smarts data*, appelées ainsi parce qu'elles sont construites sur la base d'une hypothèse de recherche qui se donne pour objectif de considérer les données dans leur contextes de production, susceptibles de devenir ainsi intelligibles et interprétables. Nous avons donc interrogé la base de données de nos logs de consultation en nous focalisant sur un segment de sites à la fois *mainstream*, courant, et exprimant une diversité linguistique et avons retenu le corpus des domaines locaux de Google en excluant les domaines .com et .fr, là nous avons vu se hisser en première place des domaines les plus consultés le Google en letton à la Bpi, venant confirmer la place des consultations en cette langue dans les emprises de la bibliothèque.

[12] . Résultat confirmé par l'analyse systématique menée par Marie Pierru, élève ingénieur à ParisTech (Rapport d'étude sur les logs Bpi, méthode des K-Means, 2017, *np*).



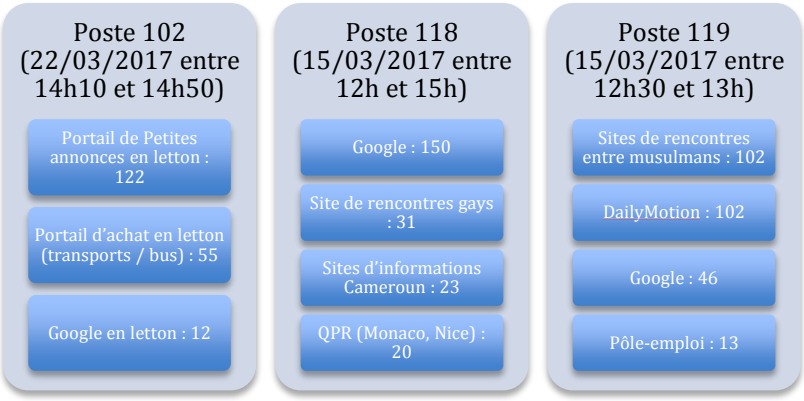
Graphique 7 : requête sur les domaines locaux de Google et mise en évidence d'une communauté d'usages en langue lettone

Pour comprendre cette présence inattendue, nous avons examiné l'indice de pénétration du numérique au sein des pays européens mis en place depuis deux ans, l'indicateur DESI (Digital Economy and Society Index) : nous avons relevé que la Lettonie se place, en mars 2016, au 9e rang des pays européens en matière d'utilisation d'internet par les citoyens, bien devant la France, située au 17e rang pour cet indicateur. L'analyse de l'usage d'internet public à la Bpi permet de révéler des pratiques d'Internet beaucoup plus diversifiées que ce que laissaient supposer les premières approches essentiellement quantitatives.

Ces *smarts data* sont cependant très dépendantes des catégorisations de sites que l'on utilise et des critères de catégorisations que l'on retient. La limite, on le voit, de la démarche que nous avons menée actuellement, est de considérer l'analyse du seul point de vue des sites consultés eux-mêmes et non du point de vue des Bpi-nautes : c'est à la recherche de traces des parcours sur Internet que nous consacrons désormais nos efforts dans la phase 2 de notre recherche qui a commencé fin février 2017. Par ces parcours, peut-être observerons-nous des articulations entre une sur-consultation des GAFA et une exploration plus singulière d'un espace moins investi du web[13]. Ainsi l'examen exploratoire[14] de quelques parcours montre-t-il que la présence quasi systématique dans tous les parcours des sites phares Facebook, Youtube ou Google ne doit pas pour autant laisser croire à une uniformité de tous les projets d'usage.

[13]. Ainsi par exemple ceux qui recourent massivement à Meetic n'ont pas forcément besoin de le faire dans un lieu public et ceux qui passent par la Bpi pour faire des rencontres en ligne développent très vraisemblablement des parcours sur le web bien spécifiques. Autrement dit, pour le champ de nos interrogations, la catégorie « sites de rencontres » n'est pas finalement si pertinente, si ce n'est pour révéler cette hétérogénéité.

[14]. Réalisé sur un corpus différent de celui présenté précédemment : corpus de mars 2017, avec traçage possible des sessions.



Graphique 8 : le site Google est présent dans toutes les sessions mais dans des proportions et au sein de logiques hétérogènes (les données chiffrées renvoient à un nombre de lignes de logs filtrées et concaténées).

Conclusions sur l'Internet "immobile" et en public

L'internet proposé sur les postes de la Bpi n'est pas complètement équivalent à celui auquel donnent accès des abonnements contractés depuis le domicile, par exemple auprès de fournisseurs privés, parce que précisément il ne s'agit pas d'un accès privé à domicile mais d'un accès public et en public. En effet, les postes Internet Bpi sont disposés dans les espaces publics de la bibliothèque au vu et au su de tous et ce contexte d'usage pèse sur ce qu'il est possible de consulter individuellement en public ; à cet égard, la Bpi, comme bien d'autres établissements, filtre l'accès à Internet dans ses emprises sur la base d'une charte soumise à l'approbation de ses usagers. Deuxième caractéristique de cet Internet Bpi public et en public, il s'agit d'un Internet partagé, le nombre de postes mis à disposition ne répondant pas à tous les besoins[15]. A cet égard, si elle est encore exploratoire, cette analyse des logs de connexion à la Bpi peut peut-être compléter la connaissance en matière d'audience et d'usages de l'internet en France enregistrées classiquement par les méthodes centrées sur l'internet à domicile (approches dites *user-centric*) ou réalisées à partir des sites producteurs (approches dites *site-centric*). Notre approche que l'on pourrait qualifier de *biblio-centric* a peut-être pour intérêt de donner à voir des usages moins souvent répertoriés puisque développés dans des contextes d'absence d'équipement personnel, quelles qu'en soient les raisons[16]. Cependant, pour peu que nous parvenions à restituer des parcours ou des communautés d'intérêt[17], les analyses qui pourront être menées à partir des *Big data* aujourd'hui disponibles à la Bpi serviront vraisemblablement plus la connaissance des usages du web, voire du web lui-même, que celle des usagers et des usages de la bibliothèque... [18]

[15] . Des besoins qu'il semble bien difficile de circonscrire ; si quantitativement, on peut estimer qu'il concerne environ 13% des usagers (Enquête TMO 2016 et enquête Bpi 2015), la variété des profils et des projets d'usage est réelle, voir Rapport d'enquête qualitative (entretiens et observations) menée par Anaïs Crinière, Agnès Camus-Vigué et Christophe Evans, *à paraître*.

[16]. On notera, à cet égard, que dans son *Baromètre du Numérique*, le Crédoc ne produit plus de données sur les connexions à Internet réalisées depuis les cybercafés ou les bibliothèques (la dernière donnée produite date de 2011 et indique que 15% des personnes se connectant à Internet le font à partir de l'un ou l'autre de ces lieux).

[17]. Travail d'exploration des logs Bpi en cours mené par Dana Diminescu et Quentin Lobbé à partir du corpus des sites de traduction en ligne.

[18]. Pour une argumentation complète, voir par exemple MENGER, Pierre-Michel (dir.) ; PAYE, Simon (dir.). *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*. Paris : Collège de France, 2017. Disponible sur Internet : http://books.openedition.org/cdf/4987.