Analysis of Gallica and Data BnF logs and Modelling of Behaviour Patterns

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

Bibliothèque nationale de France
Strategy and Research Delegation

Télécom ParisTech
Image, Data and Signal Department
Economic and Social Sciences Department

TeraLab
Mines-Telecom Institute, 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)

15 December 2017

Analysis of Gallica and Data BnF logs and modelling of behaviour patterns: presentation of the main results

Context and method

Gallica (http://gallica.bnf.fr) is one of the major digital libraries available for free via the Internet. It provides access to 4.3 million of documents of any type: printed documents (books, press and magazines) in image and text mode, manuscripts, sound and iconographic documents, maps and plans. Gallica receive around 1.5 million visits per month.

In the context of the Bibli-Lab research partnership between the BnF and Télécom ParisTech, and with the support of TeraLab, a new analysis of Gallica servers' connection logs was carried out, applying machine-learning methods to them[1]. The aim was not to collect information on users or their profiles but rather to use logs[2], which act as records of usage, as a basis for identifying typical clickstreams. For 15 months (April 2016-July 2017), Adrien Nouvellet, a researcher on postdoctoral contract and under the supervision of four of Télécom ParisTech's research professors[3], developed a data clusterisation algorithm enabling grouping of Gallica sessions with similarities in sequencing and duration of actions[4]. Logs analysed covered a range of durations, from a week to a month, with systematic checking of the stability of models obtained.

Such learning methods take advantage of the very factor that undermines traditional methods for gathering information on usage: the extremely high numbers of connections (45,000 visits/day on Gallica). Such numbers call the representativeness of online surveys into question – representative first and foremost of the most committed internauts, but not of all internauts.

Despite the power of the algorithms involved, machine learning also requires numerous decisions to be taken, necessitating availability of other sources of knowledge on usages and users. For this reason, the preferred methodological choice was to have statistical models dialogue with results obtained from other approaches (ethnographic observations,

[1] The whole report is avaible here: Nouvellet A., Beaudouin V., d'Alché-Buc F., Prieur C., Roueff F., *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, 2017, online : <https://hal.archives-ouvertes.fr/hal-01709264/document>

[2] Files containing all requests received by servers. Among other information required for knowledge of usage, logs contain the IP address (single connection identifier, anonymised for the purposes of this project), date and time (to the nearest second) of the request, the user's provenance (referring website), and the http request, which, in cases of requests for a Gallica document, contains its permanent identifier ARK.

[3] Florence d'Alché-Buc and François Roueff from the Image, Data and Signal (IDS) Department; Valérie Beaudouin and Christophe Prieur from the Economic and Social Sciences (SES) Department.

[4] The five "actions" identified in the logs are consultation of the homepage, consultation of collections guides (presentation of collections and blog), use of the internal search engine, consultation of a document in the Gallica interface, and downloading.

interviews, etc.[5]). Such dialogue enabled the researchers involved to: *a)* set departure parameters (definition of a session and the elementary actions composing it); *b)* check models obtained, which were highly sensitive to technical artefacts; and *c)* propose initial keys to interpretation.

The interest of the work carried out on the Gallica logs persuaded the BnF and Télécom ParisTech to add a further stage to the research, this time lasting four months (July-November 2017) and devoted to Data BnF logs as well as clickstreams between Gallica, Data BnF and BnF General Catalogue.

Part 1 / Gallica: main results

Large numbers of very short sessions among Gallica users

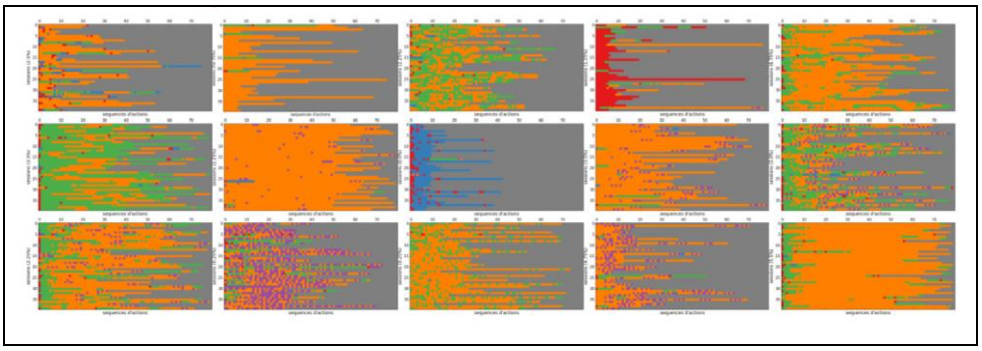
Typical of life on the web today, most Gallica users make very brief visits to the site: 50% take less than 12 seconds; 30% only make one request and only 8% of sessions consult more than four separate documents[6]. In addition, only one in every four sessions makes use of the internal search engine, a fact that has two possible explanations: *a)* a high percentage of visits are not made in order to explore collections (but rather for a single consultation of a specific document); and *b)* a number of Gallica users substitute Google for the internal search engine, not just at the beginning of but also during a session, when they want to start a fresh search.

The study of referring sites (address of the web page at the origin of a request) shows that Gallica users' web provenance has an influence on session "depth", measured here by the number of actions carried out on Gallica. Although, unsurprisingly enough, Google is the main referring site however many actions take place during a session, the same is not true of Facebook: the social network is best represented at the origin of sessions comprising between 2 and 4 actions (30% of sessions). In contrast to received ideas, then, sessions originating with Facebook are not necessarily simple "one-click" sessions. For sessions comprising over four actions (40% of sessions), referring sites – after Google but before Wikipedia and Facebook – include the bnf.fr domain and a first theme-based site focusing on genealogy (Geneanet). This is proof of the importance of theme-based sites (which attract a large proportion of amateur researchers[7]), at the origin of "in-depth" consultations of Gallica.

"Atypical" clickstreams

In order to have sufficiently rounded models available, a method for forming "clusters" (groups of similar sessions) was applied to sessions comprising more than five actions – only 35% of all Gallica sessions.

Example 1: 15-category classification



Although website design always supposes its "normal" use (e.g. homepage > internal search engine > consultation of document), clusters testify to the wide variety of clickstream patterns in Gallica consultations. In the first model of clusters obtained, only taking sequences of actions into account, 53% of sessions correspond to sequences of simple document consultations that did not make use of the homepage, did not download and did not use the search engine. Collection presentation pages appear in a single small (2.5%)

[5] See below for references to other studies carried out upstream of or in parallel to this research.

[6] A short session should not necessarily be seen as a failure, however: it might be a case of an internaut finding what he/she is looking for immediately by clicking on the right link (hypertext). 79% of single-task sessions on Gallica consist of a simple document request

[7] Beaudouin V., Pehlivan Z., *Cartographie de la Grande Guerre sur le Web* (Cartography of the Great War on the Web), Final Report on Phase 2 of "The Future of Online Digitized Heritage: The Example of the Great War" project, September 2016, online: <https://hal.archives-ouvertes.fr/hal-01425600/document>.

cluster, proof that such pages are not on most Gallica users' routes; their consultation constitutes a form of behaviour distinct from all the others observed.

Incorporation of the time factor provided a more accurate idea of sessions as more than simple sequences of actions: it brings together sessions that were previously assigned to different clusters but which turned out to have the same temporal "outline". As an example, consultation time for a "view" of the same document on Gallica may vary from 0 to 52 minutes, depending on conventions chosen for analysis[8]: it would seem reasonable not to assimilate such heterogeneous levels of engagement.

With the new model, the largest cluster (42%) brings together sessions taking an average of 7 minutes and alternating consultation actions (3 minutes on average) and search engine actions (2 minutes on average), with random and briefer presence of other actions – a model that is therefore closer to that of the site's designers. Other clusters also deserve attention as they enable quantification of behaviour patterns that are both simple and typical: 28% of sessions are dominated by downloading activity, often associated with consultation; 13% of sessions only comprise alternating consultation and search engine actions, with more time devoted to the former; in contrast, 3.5% comprise long search sequences (11 minutes) terminating with brief consultations. Finally, although use of the homepage during a session usually takes no longer than 30 seconds, 7% of sessions spend an average of 4 minutes on it (not necessarily continuously) within a varied sequence of actions that even occasionally includes consultation of collection presentation pages. This final result shows that, even though it may scarcely be apparent, there are people who consult the homepage and know how to make good use of it for careful and varied consultations. The largest clusters should not be allowed to conceal forms of behaviour that are non-negligible at quantitative level and of interest as regards development of audiences and uses.

If we now return to all sessions, however many actions they include, it can be seen that 4% of them never get past Gallica's homepage (13% of single-action sessions) and do not actually enter the site itself.

Little diversity in types of documents consulted during a session

In order to measure the diversity of documents consulted, logs were provided with the metadata of documents contained in the Open Archives Initiative (OAI) repository. Correcting the assertion made in the 2016 online questionnaire on this point[9], it is clear that daily-press publications are the type of document most often consulted, ahead of monographs and images, even though, compared to the number of press titles available in Gallica, the consultation ratio is the lowest.

This is surprising: despite the exploration facilities provided by the web's interfaces and the much-vaunted role of serendipity, Gallica consultations are still largely single-type. This is the case with 45% of sessions in which more than 5 documents are consulted, with predominance, as might be expected, of sessions focusing on daily-press publications and monographs alone. Sessions analysed, although longer than average, bear witness to users consulting the site in "silo" fashion (as if organisation of collections and research practices were still compartmentalised), as was shown by the 2012 study of document requests in the BnF's "*Rez-de-jardin*" reading rooms[10]. One challenge for the Gallica interface will be to encourage "bouncing" from one type to another (from a manuscript by Apollinaire to a recording of his voice, for example). Only 3% of sessions covering more than 5 documents explore almost all document types.

The most frequent actions carried out during a session vary greatly depending on documents requested: during sessions exclusively devoted to press publications, Gallica users spend most of their time consulting (i.e. making requests containing an ARK) and much less of it downloading, unlike sessions devoted to manuscripts, where downloading takes up most of users' time[11]. This type of analysis can help prioritise developments of

[8] The video ethnography carried out in parallel to this analysis evidenced the existence of very lengthy consultations of single views (see Rollet N., Beaudouin V., Garron I., *Vidéo-ethnographie des usages de Gallica* (Video ethnography of uses of Gallica), Final Report on Phase 2, April 2017, online: <http://www.bnf.fr/documents/mettre_en_ligne_patrimoine_ethnographie.pdf>, which led to revision of the definition of a session: a session on Gallica ends when the time between two requests exceeds 60 minutes (where, for other web services, received practice advises 10 minutes). We should bear in mind, however, that logs tell us nothing about user activity outside Gallica.

[9] TMO Régions, *Enquête auprès des usagers de la bibliothèque numérique Gallica* (Survey of Gallica Digital Library Users), survey report, 10 April 2017, online: <http://www.bnf.fr/documents/mettre_en_ligne_patrimoine_enquete.pdf>.

[10] Pardé Th., "Les usages documentaires dans une bibliothèque de Recherche" (Documentary usages in a Research Library), *Bulletin des Bibliothèques de France* (BBF), 2015, no.5, pp.112-119, online: <http://bbf.enssib.fr/consulter/bbf-2015-05-0112-002>.

[11] This result is clarified by qualitative analysis: interviewees stated that they consulted the press in Gallica as its reading tools were well suited to the purpose (zoom functions in particular), whereas they downloaded monographs, which could be read just as

functionalities by type of document. Tools for exploring press collections and designed to facilitate refinement (e.g. search by named entities) would therefore appear necessary.

Social networks: the impact that methods of publication on Facebook has on direct users

Communication on social networks has a proven impact on Gallica users, and, as we have seen, Facebook is well represented among referring sites. Analyses confirmed this with regard to a specific event: the peak period for consulting Voltaire's works on the anniversary of his death in 2016 was compared with the success of a publication on Gallica's Facebook page the same day (1,055 likes and 568 shares). Furthermore, although numbers of publications per week on the page have remained relatively stable since its creation, numbers of reactions per publication have increased significantly.

A study on types of links to Gallica included in publications has shown that they affect the number of "clicks": an active link in an image generates 25 times as many visits to Gallica than an active link in the text (with indication of the URL). This result led the team responsible for the Facebook page to modify its publication methods.

Part 2 / Data BnF and the general catalogue

The aim of the Data BnF project (<http://data.bnf.fr>) is to make the BnF's data more useful on the Web. It involves transforming existing data, enriching and interlinking the dataset with internal and external resources, and publishing HTML pages for browsing by users and search engines. The raw data is also available in RDF following the principles of linked open data architecture

Data BnF consultations follow a particular pattern as far as times of day are concerned: peak times for consulting the interface occur during the day (at 11 a.m. and 3-4 p.m.), whereas the peak period for Gallica is during the evening (between 8 and 10 p.m.). In-depth visits to Data BnF may also be quantified: 800 sessions a day include over 4 actions (5% of the total) and 250 sessions (1.5%) consult more than 4 entity pages. "*Auteur*" is by far the most consulted entity, although "*Thèmes*", "*Spectacles*" and above all "*Œuvres*" (the second most consulted) are consulted proportionally more often than their presence in Data BnF would suppose.

Over the course of a month, 12% of two million authors were consulted – a considerable figure: the function providing a breakdown of authors consulted at least once in a month enables formulation of the hypothesis (which must nonetheless be tested out over a longer time period) that most authors will have been consulted after enough time has passed and that none of them really leads the others. There would therefore be no "dark zone" in Data BnF, whose listing of all webpages seems to function as it should.

As initially desired by the Metadata Department, clickstreams between Gallica, Data BnF and BnF General Catalogue were then subjected to analysis with a view to understanding how today's internauts access BnF documents and metadata on the web: how they enter applications and travel – or otherwise – from one to another. The large number of consultations of Gallica *alone* results in only 4% of all sessions combined on the three sites being multisite. Among such multisite sessions, Data BnF is by far the most frequently in evidence (87%), ahead of Gallica (69%): clear proof that it is successful in playing its role as a "pivot" between the BnF's main documentary and bibliographic services.

Data BnF acts as a gateway to Gallica documents before acting as one to data contained in BnF General Catalogue: half of all bounces are between Data BnF and Gallica, with only one out of every three bounces between Data BnF and the catalogue. Direct access to documents is therefore the prime motive among users who, after visiting Data BnF, continue their search in the bnf.fr domain. Such bounces sometimes occur unexpectedly from Gallica to Data BnF or, more rarely, from BnF General Catalogue – probably due to the browser's navigation tabs, as there are no direct links from Gallica documents or BnF General Catalogue to Data BnF.

Grouping multisite sessions into "clusters" confirms that:

easily offline (see Rollet N., Beaudouin V., Garron I., *Je pars d'un sujet, je rebondis sur un autre: pratiques et usages des publics de Gallica* (I start off on one subject and bounce to another: Gallica users' practices and usages), exploratory qualitative study, Final Report on Phase 1, September 2016, online: <http://www.bnf.fr/documents/mettre_en_ligne_patrimoine_entretien.pdf>. The 2016 survey by TMO Regions also drew attention to the increase in careful consultation in Gallica users' stated practices (TMO Régions, *op. cit.*).

- Data BnF is not simply a gateway to other services: in over a quarter of multisite sessions, lengthy sessions on Data BnF precede resulting consultations of Gallica (13%) or BnF General Catalogue (5%);
- 10% of multisite sessions take full advantage of all three services and may, by general agreement, be referred to as "expert";
- The most "expected" sorts of sessions – brief consultations of Data BnF that switch to long consultations of BnF General Catalogue (initial idea of a "documentary pivot") – actually account for only 3.5% of multisite sessions.

Finally, as with Gallica, most Data BnF sessions (54%) only consult a single entity type, most often "*Auteurs*". Bi-type sessions are usually sessions in which one type is consulted for a longish time before a brief switch to the second, but with no real alternation of different types during a session ("*Auteurs*" / "*Œuvres*" alternations only occur in 1% of sessions).

Avenues for work

Measuring evolutions in usage as the interface evolves

For Gallica, all processing was carried out on logs prior to 15 June 2016, i.e. before redesign of the homepage. Renewing such analysis would enable very direct measurement of the impact that evolution of the interface has on behaviour patterns.

Exploring segments of the public important to development of the interface

Analysis of logs enables identification and quantification of usages denoting high levels of engagement (time spent, number of actions carried out, diversity of types of documents consulted, etc.). Even though they are in a minority as far as overall numbers are concerned, they are far from being negligible in terms of absolute value, reflecting richly varied usages of the interface that it would be useful to understand better. For example, although most Gallica sessions are single-type, 3% of sessions covering more than 5 documents go from one type to another: how are such clickstreams to be better understood? What actions and referring sites characterise them? What currently fosters such exploration of diversity?

Having a more refined semantic classification of documents available

For Gallica: apart from document types and other information contained in the OAI, and so far used separately (author, theme, etc., which are not always relevant or provided for all documents), it would be useful to be able to analyse research topics and their evolution over the course of clickstreams by having a record of the content of documents consulted available (e.g. via a number vector). An initial attempt at classification was made based on words in OAI notes; another method was also suggested ("word embedding", learned of from the corpus of articles in Wikipedia) but was not implemented due to lack of time.

For Data BnF: in order to gain a more detailed understanding of the makeup of Data BnF's audience and confirm or otherwise the "long tail" effect of consulting it, the fame of authors and/or works might be assessed in both conventional and automated fashion through use of information internal (number of documents "about this author") or external (Wikipedia) to the site.

Incorporating the notion of user

Beyond the notion of "session", it would be useful to analyse user behaviour patterns with the help of a domain cookie with a short lifespan (1 month, for example). The same user might well make varied use of our interfaces, alternating short sessions and long sessions, etc. This possibility will be investigated in 2018.