Gallica au crible de la science des données : l'exemple du Bibli-Lab

Philippe Chevallier (BnF)

47e congrès de l'ADBU : Les bibliothécaires, acteurs de la recherche
18 octobre 2017

2013 : création du « laboratoire d'étude des usages du patrimoine numérique des bibliothèques » (Bibli-Lab)

- Depuis 2002, à la BnF : enquêtes déclaratives auprès des usagers de **l'interface** de Gallica (entretiens et questionnaires en ligne).
- Mais de nouvelles questions, qui appelaient un renouvellement des méthodes, afin de prendre en compte **la masse des connexions** et **l'importance des usages furtifs**…

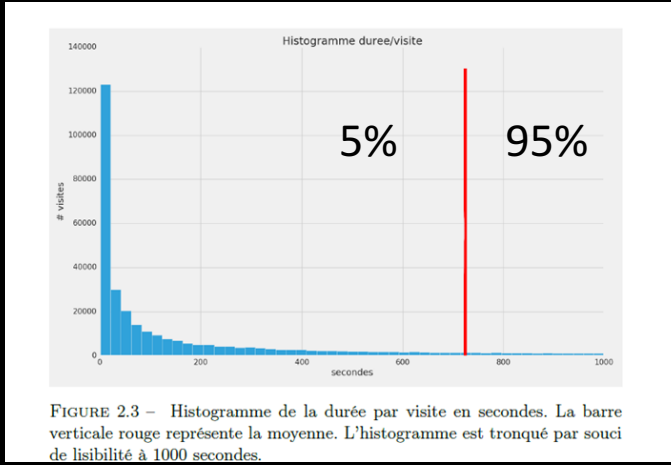
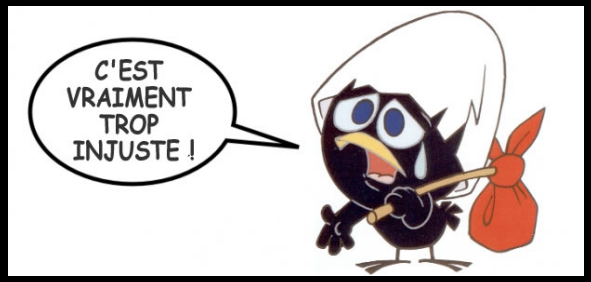


FIGURE 2.3 – Histogramme de la durée par visite en secondes. La barre verticale rouge représente la moyenne. L'histogramme est tronqué par souci de lisibilité à 1000 secondes.



Un log, c’est un événement survenu au serveur

Il contient :
- l’adresse IP de l’ordinateur qui s’est connecté
- son origine géographique
- la date et l’heure
- le site de provenance (« referer »)
- le navigateur utilisé
- la page appelée avec, dans le cas d’un document numérique de Gallica, un identifiant unique qui permet de faire le lien avec la notice bibliographique

→ Profiter de cette masse pour repérer des régularités et identifier des types de navigation grâce à un apprentissage non supervisé (« *Clustering* ») ;

→ Importance de formuler des hypothèses de travail : Que voulons-nous savoir ?

1) Gallica favorise-t-il une diversité de consultation des documents par type (imprimés, estampes, manuscrits, etc.) ?
2) Quel impact ont les actions de médiation (page Facebook, page d'accueil, etc.) sur la consultation ?

→ Mise en place d'une convention d'application du partenariat :

1) 1 Contrat de recherche postdoctoral de 15 mois (avril 2016-juillet 2017) : Adrien Nouvellet, docteur en traitement du signal
2) Encadré par deux département de Télécom ParisTech : « Image, données, signal » (F. Roueff, F. d'Alché-Buc) et « Sciences économiques et sociales » (V. Beaudouin, C. Prieur).
3) 1 comité de projet et 1 groupe de suivi mensuel impliquant la délégation à la Stratégie et la direction des services et des réseaux de la BnF

Travailler collectivement sur les données

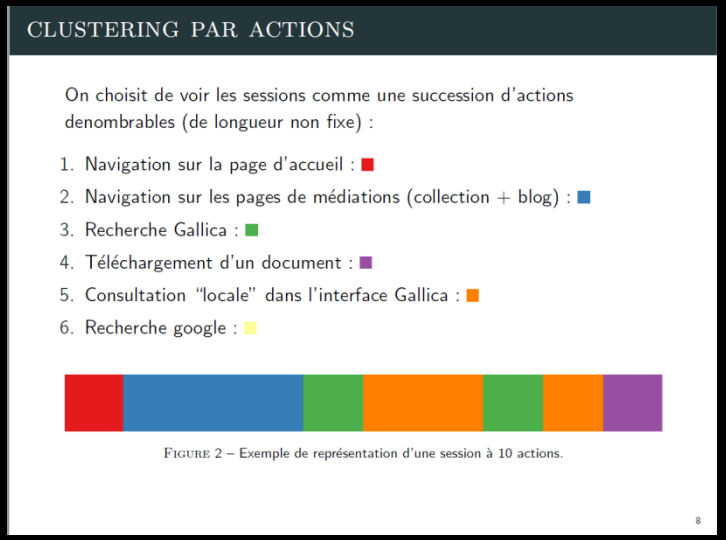
→ Importance d'un dialogue Institution / Chercheurs, mais aussi Science des données / Sciences sociales :

1) Documenter les données : comment ont-elles été produites ?

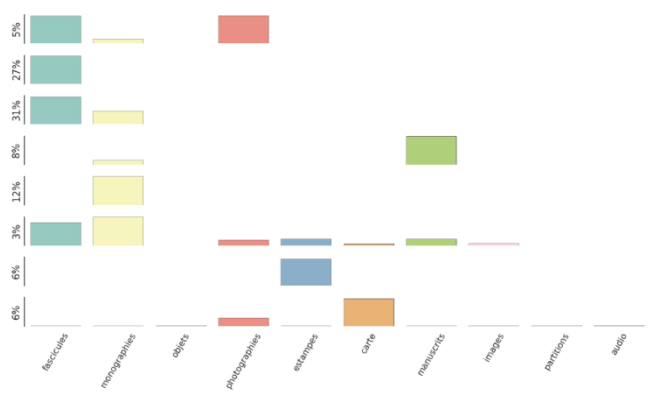
2) Nettoyer les données : bots, artefacts, tests, etc.

3) Organiser les données : définir une durée de session, définir les actions-types de Gallica, etc.

→ *Toujours confronter les données avec d'autre sources d'informations*



CENTROIDS DES CLUSTERS PAR TYPE



TEMPS MEDIAN POURCENTAGE

clusters	homepage	médiation	recherche	téléchargement	consultation
2 : monographies	14	21	28	17	21
3 : fascicules	1	3	16	22	58
4 : partitions	3	7	21	30	39
5 : tout	2	7	24	20	46
10 : monographies bis	2	3	11	29	56
11 : fascicules/mono	2	3	27	19	49
12 : manuscrits	4	8	12	65	11
13 : manuscrits/mono	3	5	17	32	43
14 : photo/cartes	3	9	20	25	43
ALL	2	6	19	27	45

TABLE 5 – Pourcentage du temps median de chaque action pour chaque cluster de type

SESSIONS 1ER REFERERS

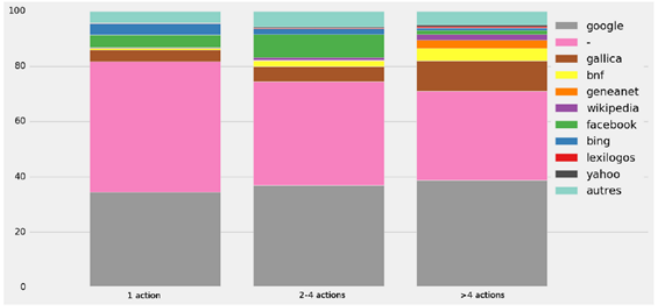
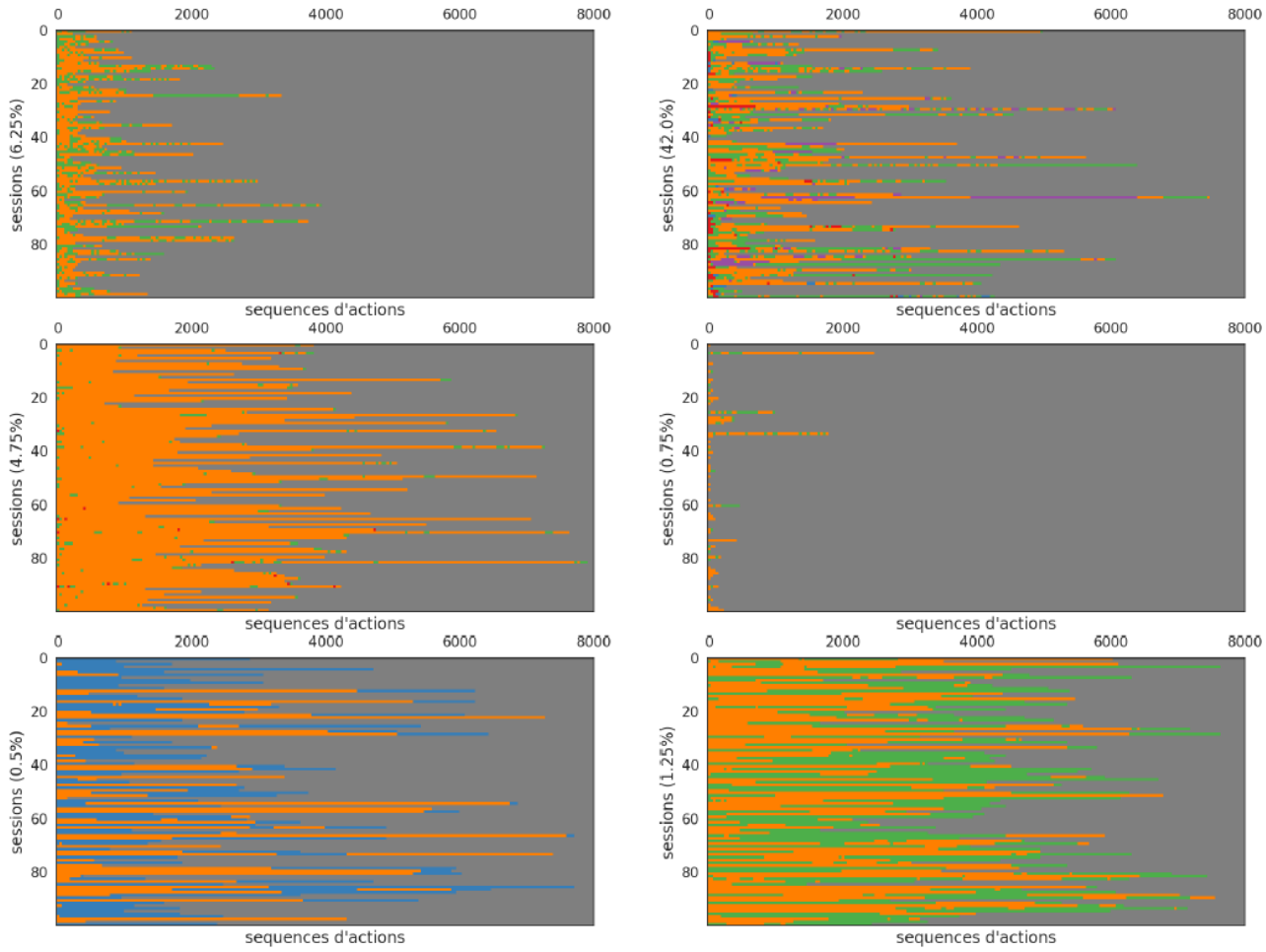


FIGURE 1 – Proportion des 1er referers pour chaque catégorie de sessions.

Tous les tableaux et graphiques ont été réalisés par Adrien Nouvellet (chercheur post-doctoral à Télécom ParisTech), sous la supervision de Florence d'Alché-Buc, Valérie Beaudouin, Christophe Prieur et François Roueff.


0
2000
4000
6000
8000
0
20
40
60
80
sessions (6.25%)
sessions (42.0%)
sessions (4.75%)
sessions (0.75%)
sessions (0.5%)
sessions (1.25%)
sequences d'actions

Ce que n’est pas le Big Data…

And the winner is (not)…

Conclusion : approcher les usages et les usagers de Gallica en croisant les méthodes

Méthode	Échantillon	Atout	Limite
Fouille des données de connexion (logs)	x 100,000,000	« Tout » prendre et tirer partie de la masse	Documenter, nettoyer, modéliser, interpréter
Questionnaire	x 1,000	Quantifier des propositions	Les questions sont des présomptions d'usages
Entretiens	x 10	Comprendre les « raisons de faire »	Difficulté de raconter des pratiques de plus en plus complexes
Ethnographie	x 10	« Voir » la pratique	Singularité du cas concret