Compter les mots pour remonter le temps : Gallicagram et Gallicagrapher, deux outils d'exploration des archives numérisées de la BnF

Les archives numérisées de la Bibliothèque nationale de France, Gallica, constituent un trésor de données ouvertes. Les auteurs ont développé deux logiciels de lexicométrie, qui mesurent et permettent de visualiser l'évolution de l'usage des mots au cours du temps, et facilitent l'accès au contexte des occurrences.

BENJAMIN AZOULAY
Administrateur-élève des affaires maritimes, ministère de la Mer et École normale supérieure Paris-Saclay

BENOÎT DE COURSON
Doctorant au Max Planck Institute de Freiburg (Allemagne)

WILL GLEASON
Développeur indépendant

La Bibliothèque nationale de France (BnF) dispose d'un des plus riches fonds imprimés au monde. Elle est en effet dépositaire du dépôt légal : chaque livre ou numéro de presse publié depuis 1537 doit y être déposé pour archivage et désormais pour consultation par les chercheurs.

Depuis 1997, la BnF entreprend une numérisation massive de ses fonds et les verse en libre accès sur la plateforme Gallica. Cette « bibliothèque virtuelle de l'honnête homme[1] » est aussitôt devenue un outil de travail incontournable pour les chercheurs en humanités.

De par son volume (400 000 livres et 4,5 millions de numéros de presse océrisés[2] en français), Gallica se prête à merveille à la lexicométrie, c'est-à-dire au traitement quantitatif des textes. L'accès transparent aux données est ici un atout précieux. Développés à cette fin, les sites Gallicagram et Gallicagrapher visent à offrir aux chercheurs en humanités une interface Web pour exploiter et visualiser ces données.

Les outils lexicométriques existants n'appliquent guère les principes de la « science ouverte ». L'outil de référence en la matière, Google Books Ngram Viewer[3], est peu utilisé par les chercheurs, embarrassés par la constitution opaque de son corpus et déçus de ne pouvoir accéder au contexte des occurrences[4]. Plus souple et transparent, Frantext[5] permet des relevés syntaxiques dans près de 6 000 œuvres (soit 266 millions de mots), mais son volume est insuffisant pour des analyses quantitatives diachroniques[6], et Frantext (par ailleurs payant) a récemment cessé d'assortir les données récoltées de graphiques.

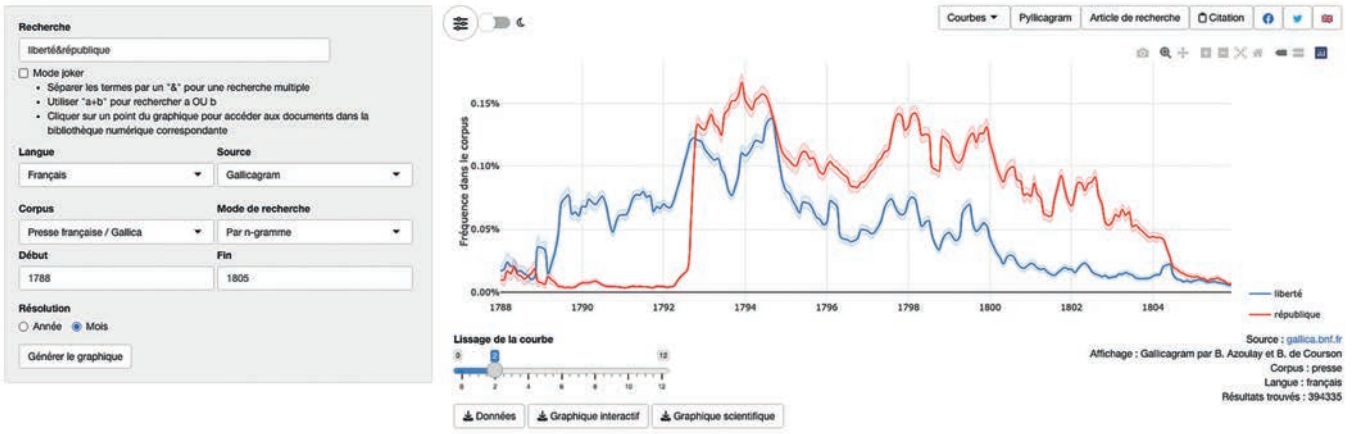
Gallicagram et Gallicagrapher : deux logiciels pour démocratiser le traitement automatique de Gallica

Gallicagram[7] est un logiciel permettant de visualiser l'évolution de l'usage des mots au cours du temps, en fouillant, parmi d'autres corpus, la presse et les livres numérisés de Gallica. La croissance des données interrogées à partir de la Révolution rend le corpus de presse (certainement le plus intéressant pour l'historien) particulièrement fiable entre 1789 et 1950.

Disposant d'un corpus ouvert, l'historien est à même de savoir dans quoi il cherche, et d'accéder au contexte des occurrences. Pour ce faire, Gallicagrapher[8] exploite les API[9] de Gallica et présente le contexte immédiat de chaque occurrence, directement dans le logiciel, sur le modèle de Frantext. Cela facilite l'analyse des résultats correspondant aux courbes affichées, ce qui permet, par exemple, de lever les ambiguïtés sur les homonymes et les erreurs de reconnaissance optique des caractères.

Les deux logiciels cherchent à appliquer les principes de la science ouverte et collaborative. Ils sont

1. **https://gallica.bnf.fr/edit/und/a-propos**
2. Le terme « océrisation » dérive de l'abréviation OCR : *Optical Character Recognition*, c'est-à-dire en français : « Reconnaissance optique des caractères » (ROC, peu utilisé). Techniquement, il s'agit du traitement d'une image (le texte est scanné, comme par une photocopieuse) sur laquelle on fait intervenir un logiciel de reconnaissance de caractères : le logiciel déchiffre les formes et les traduit en lettres.
3. **https://books.google.com/ngrams**
4. François Héran, « Les mots de la démographie des origines à nos jours : une exploration numérique », *Population*, vol. 70, 2015, p. 525-566.
5. **https://www.frantext.fr**
6. Nous développons ce point dans l'article de Benoît de Courson, Benjamin Azoulay, Clara de Courson, Laurent Vanni et Étienne Brunet, « Gallicagram : les archives de presse sous les rotatives de la statistique textuelle », *Corpus*, nº 24, 2023. **https://doi.org/10.4000/corpus.7944**
7. **https://shiny.ens-paris-saclay.fr/app/gallicagram**
8. **https://www.gallicagrapher.com/**
9. Une API (*Application Programming Interface* ou « interface de programmation d'application ») est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.



Interface de Gallicagram : **https://shiny.ens-paris-saclay.fr/app/gallicagram**

Page 1 de 645 > >>

Document	Gallica	Contexte gauche	Pivot	Contexte droit
(5) *Le Mémorial des Pyrénées : politi* 1888-08-07	Image	INFORMATIONS UES	GRÈVES	A PARIS Les terrassiers La réunion quotidienne des ouvriers grévistes terrassier
(7) *Le Petit républicain : journal quoti* 1888-08-03	Image	donner leur adhésion à la	grève	des terrassiers. les cochers de fiacre Après deux discours des citoyens Winant e
(4) *Les Chantiers de l'Exposition univ* 1888-08-01	Image	La	Grève	et l'Exposition de 1889. —Les Chantiers de Pari». — Les Travaux de Paris. — Une
(12) *Le Cri du peuple : journal politiqu* 1888-08-01	Image	LA	GRÈVE	DES MINEURS (De notre correspondant) Saint-Etienne, 30
(1) *La Lanterne de Boquillon / par A. I* 1888-08-19	Image	mirent en	grève	et les garçons de café «c extras », battus par Lozé II à Aboukir, se jurèrent de
(7) *Le Clairon du Lot : journal monarc* 1888-08-07	Image	que les progrès faits par la	grève	dans d'autres corporations ne pouvaient qu'encourager les terrassiers à persévér
(8) *Le Cri du peuple : journal politiqu* 1888-08-21	Image	La	grève	doit se transformer, comme s'est transformée l'industrie et la guerre elle-même
(9) *Le Cri du peuple : journal politiqu* 1888-08-02	Image		GRÈVE	ET SYNDICAT Huit mille ouvriers se sont mis en
(10) *Le Radical algérien : paraît tous l* 1888-08-05	Image	E» c est comme ce a partout 1 V La	grève	des mineurs. — On écrit do Saint Etienne
(7) *Le Cri du peuple : journal politiqu* 1888-08-03	Image	UNE NOUVELLE	GRÈVE	Les verriers à vitres de Saint-Etienne Le verre à vitre. — L'usine Velin. — Un r
(8) *La Presse* 1888-08-02	Image	Une	grève	se résume toujours pour les travailleurs par ces trois phases: Plus de salaires,
(9) *L'Intransigeant* 1888-08-06	Image	Le jour où toutes les corporations du bâtiment se mettraient simultanément ^n	grève	, on verrait ce que pèse le capital devant le travail, et il faudrait bien que l
(8) *Le Radical* 1888-08-09	Image	LE DEVOIR DES PATRONS La	grève	des terrassiers semble entrer enfin dans la voie de l'arrangement
(10) *Le Cri du peuple : journal politiqu* 1888-08-11	Image	La	grève	des ouvriers verriers. — Im- portante réunion ouvrière. — Pas de désordres
(7) *Le Cri du peuple : journal politiqu* 1888-08-14	Image	Les terrassiers en	grève	ne font que réclamer do leurs exploiteurs les salaires et les conditions que le

Fonction « contexte » de Gallicagrapher : **https://www.gallicagrapher.com/** Gallicagram utilise l'API de Gallicagrapher pour proposer une présentation analogue sous ses graphiques.

Open Source[10], ce qui permet le réemploi du code dans le cadre de projets tiers. Gallicagram est aussi *Open Data* : la base de données constituée par le décompte des milliards de mots des corpus numérisés de Gallica (presse et livres français) est accessible par API[11]. Les deux logiciels collaborent puissamment à travers leurs API respectives : Gallicagram fournit ses données au graphique affiché par Gallicagrapher, qui lui renvoie le contexte des occurrences (figure ci-dessus).

Ceux-ci étant destinés à une population de chercheurs inégalement à l'aise avec l'informatique, l'ergonomie est un enjeu central. Gallicagrapher ne fait au fond qu'enrober les API de Gallica pour présenter de façon plus intuitive et plus maniable des données que le chercheur aurait pu trouver manuellement.

Mais il y a fort à parier que sans ce tour de force ergonomique, il n'en aurait pas eu le courage. Notons aussi que le développement de telles applications Web interactives a été facilité par les avancées récentes des langages de programmation (respectivement Shiny et React pour les deux logiciels).

Des obstacles persistants à l'ouverture des données

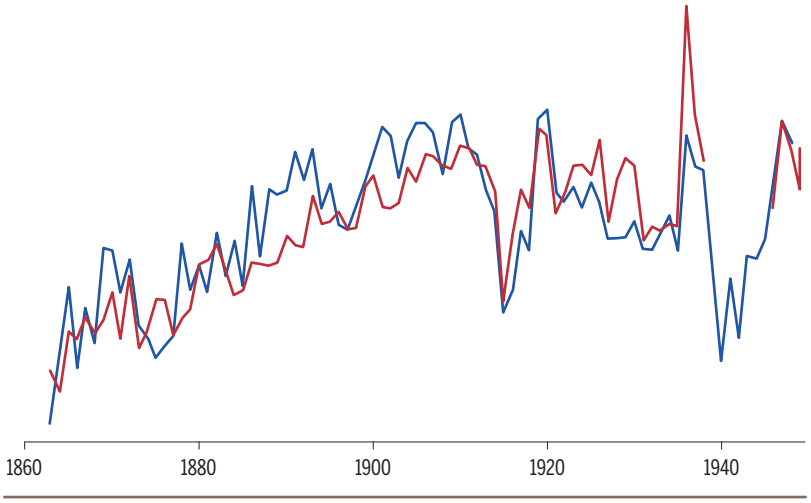
Ces logiciels se sont développés en tirant profit des données ouvertes de Gallica – mais aussi, avouons-le, en contournant les barrières à l'entrée de sources qui le sont moins.

Nous nous sommes surtout heurtés à Retronews, un service développé par BnF-Partenariats pour

10. **https://github.com/regicid/docker_gallicagram**, **https://github.com/gleasonw/gallica-grapher**

11. **https://github.com/regicid/pyllicagram**

Superposition de la fréquence occurrentielle de « grève(s) » dans la presse de Gallica avec le nombre de grèves effectivement recensées



monétiser certaines archives de presse. Outre l'aspect étonnant du rétablissement des droits d'auteur s'agissant de documents tombés dans le domaine public, la valeur ajoutée du site est faible (« ré-reconnaissance » optique des caractères des fichiers numérisés et interface graphique plus moderne). Si Retronews n'avait été qu'un doublon de Gallica, le mal serait bénin ; mais la BnF a engagé un authentique versement de ses archives numériques vers l'opérateur, et ce pour une bonne moitié des journaux de la collection. L'exploitation d'une option cachée du site Retronews nous a permis d'extraire massivement le texte de ces journaux, sans lesquels notre base serait largement incomplète. Lorsque l'on effectue une recherche avancée sur Gallica, le contexte des occurrences issues de documents de Retronews est masqué[12]. Pourtant, les API de Gallica pouvaient y accéder – ce blocage semble donc délibéré. La fonction « contexte » de Gallicagrapher permet d'en rétablir l'accès en exploitant cette bizarrerie.

Par ailleurs, l'exploitation statistique de corpus soumis aux droits d'auteur se situe dans une zone grise juridique. Gallicagram ne diffuse que des nombres (les fréquences d'occurrence des mots au cours du temps), et non les textes eux-mêmes. Les ayants droit ne sont donc pas lésés, et l'on peut penser que notre usage tomberait dans l'exception au droit d'auteur « fouille de textes »[13], récemment introduite en droit français[14]. Mais l'incertitude demeure, et ce flou pousse aujourd'hui de nombreux chercheurs à l'autocensure. Il est heureux que Retronews, contacté après l'extraction des données (*scraping*)[15], ait décidé de ne pas chercher noise à deux jeunes développeurs bénévoles. Mais il serait préférable qu'une doctrine plus claire apaise la conscience et les nuits des chercheurs. ■

L'exploitation statistique de corpus soumis aux droits d'auteur se situe dans une zone grise juridique. Gallicagram ne diffuse que des nombres (les fréquences d'occurrence des mots au cours du temps), et non les textes eux-mêmes. Les ayants droit ne sont donc pas lésés.

12. Par exemple : **https://t.ly/ghX1**

13. Voir dans ce numéro l'article de Didier Thebault, p. 85.

14. Articles L. 122-5-3 et R. 122-23 du Code de la propriété intellectuelle. Décret n° 2022-928 du 23 juin 2022 portant modification du Code de la propriété intellectuelle. Ordonnance n° 2021-1518 du 24 novembre 2021.

15. Le *Web scraping* est une technique permettant l'extraction des données d'un site *via* un programme, un logiciel automatique ou un autre site. L'objectif est donc d'extraire le contenu d'une page d'un site de façon structurée. Le *scraping* permet ainsi de pouvoir réutiliser ces données.