Les optiques de la lecture distante : de l'occultation à la déconstruction de l’archive

Newseye — 15 mars 2021

Pierre-Carl Langlais
@Dorialexander
Alexander Doria (Wikipedia)

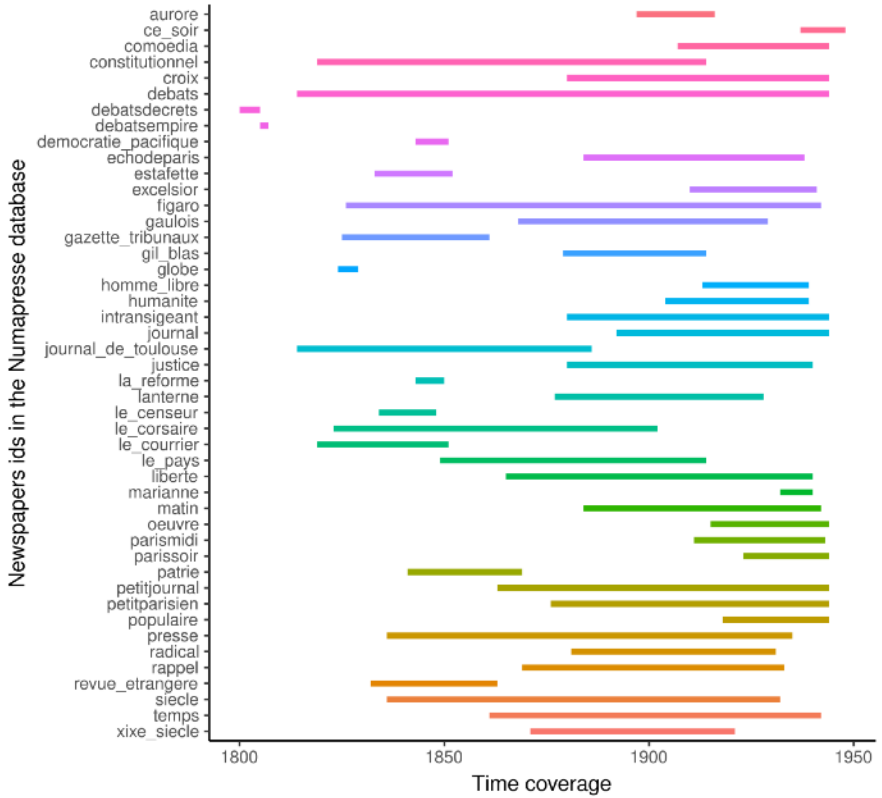
NUMA PRESSE
numapresse.org

Huma-Num
PROJET FINANCÉ PAR L'ANR
ANR
PROJECT FUNDED BY THE ANR

Les corpus de Numapresse

Depuis septembre 2019, Numapresse dispose des archives numériques détaillées de l'ensemble de la presse quotidienne nationale française disponibles sur Gallica.

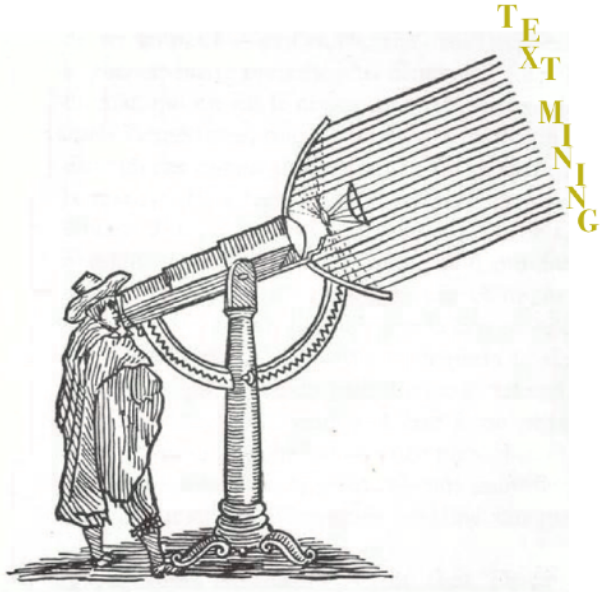
Même si les collections numérisées restent incomplètes cette intégration à grande échelle permet d'élargir le champ de nos investigation : notre objet n'est plus seulement un journal mais un écosystème médiatique



De la lecture comme un art optique

Les principaux quotidiens étudiés ont été publiés continuellement sur de très longues périodes (jusqu'à 145 ans pour le *Journal des débats*). L'évolution des formes culturelles s'intègre dans plusieurs temporalités :

- Quelques semaines pour la dissémination d'une nouvelle ou pour des collaboration de courtes durées
- Quelques mois pour les traits structurels de l'écosystème médiatique
- Quelques années pour le fonctionnement d'une rédaction.
- Quelques décennies pour l'émergence de genres journalistiques.



Pour tout prendre en compte, nous allons devoir varier nos focales…

Plan

Une présentation en trois parties :

- Lost in big data : dynamiques d’invisibilisations dans le patrimoine numérisé
- Voir l’invisible à distance : de l’archive-document à l’archive-réseau

1. Lost in big data

Dynamiques d’invisibilisations dans le patrimoine numérisé

Un patrimoine devenu “big data”…

En vingts, bibliothèques et acteurs privés ont développé de grands programmes de numérisation. Ces corpus sont de plus en plus disponibles pour des analyses quantitatives massives. La BNF a ainsi mis en ligne 1,5 to d’archives xml utilisées par le projet Europeana Newspaper ou 200 000 monographies utilisées par le projet OBVIL.

Documents de presse numérisés en mode « OCR » du projet Europeana Newspapers

Mots-clefs: Presse, gallica, Europeana Newspapers

Ce jeu de données contient les documents numériques des collections de presse traitées durant le projet européen Europea
une reconnaissance du texte (OCR, *optical chara*

Sommaire

- Contenu du jeu de données
- Contexte de production
- Formats du jeu de données
- Exemples d'utilisation
- API et jeux de données en relation

Contenu du jeu de données

Ce jeu contient la transcription réalisée par OCR d'enviro
collections de presse de Gallica traitées durant le projet E

Tous les documents numérisés des titres suivants sont pr

- Le Figaro
- L'Echo de Paris
- L'Univers
- La Presse
- L'Humanité
- Le Constitutionnel
- Le Petit Journal
- Le Siècle
- L'Action Française
- L'Intransigeant

Fiche technique

Date de création ou de mise à jour :
2015

Quantité :
1 287 500 pages, 275 000 fascicules

Langue :

TEI JSON Reconnaissance automatique des caractères (OCR) Textes

DOCUMENTS DE GALLICA PRODUITS AU FORMAT TEI PAR OBVIL

Présentation

Ce jeu de données contient le mode texte des documents de Gallica traités par l'Observatoire de la vie littéraire (Labex OBVIL). Le corpus est en français, issu majoritairement de l'édition du XIXe siècle.

TÉLÉCHARGER

Classement alphabétique (13,6 Go)

Classement Dewey (8 Go)

Classement par siècle (13.8 Go)

Métadonnées complètes (8 Mo)

FICHE TECHNIQUE

Date de mise en ligne
2018

Format
TEI JSON

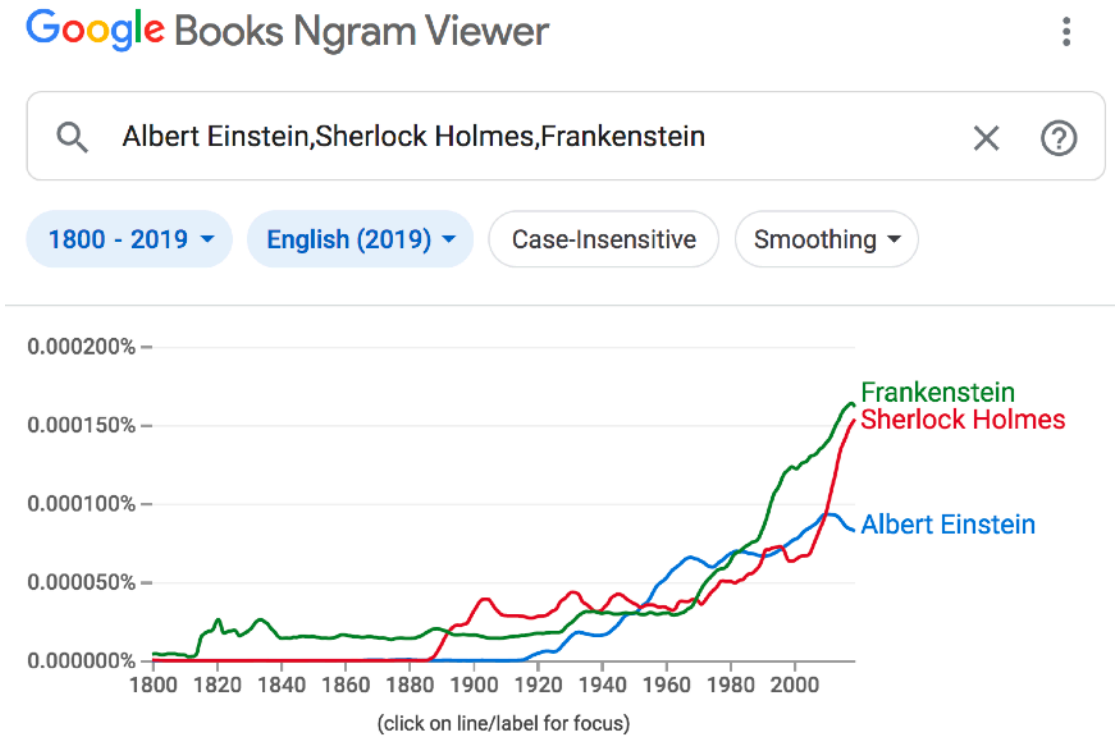
Licence
Conditions d'utilisation des contenus de Gallica

Technologies
Reconnaissance automatique des caractères (OCR)

Sujets
Textes

La fascination du *big data*…

De grandes ambitions et… beaucoup de problèmes. Les promesses inabouties de Ngram Viewer et des culturonomics.



La fascination du *big data*…

« Les *culturonomics* désignent l'application de la collecte et de l'analyse de données massive à l'étude de la culture humaine. Les livres sont un début, mais nous devons également intégrer les journaux, les manuscrits, les cartes, les œuvres d'art et une myriade d'autres créations humaines »

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17]*†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between

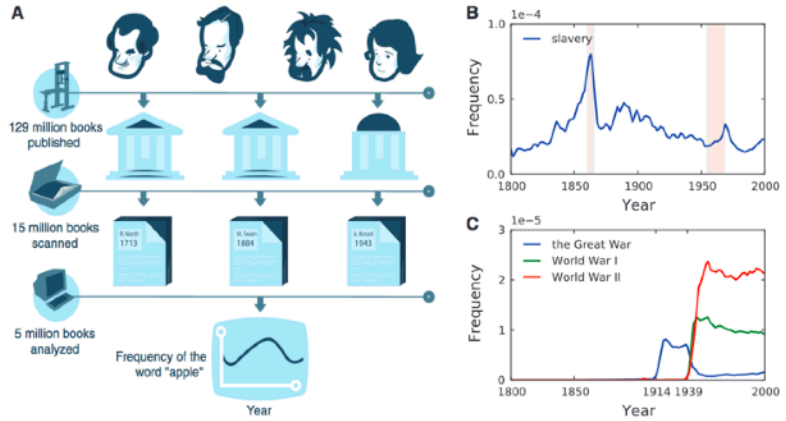
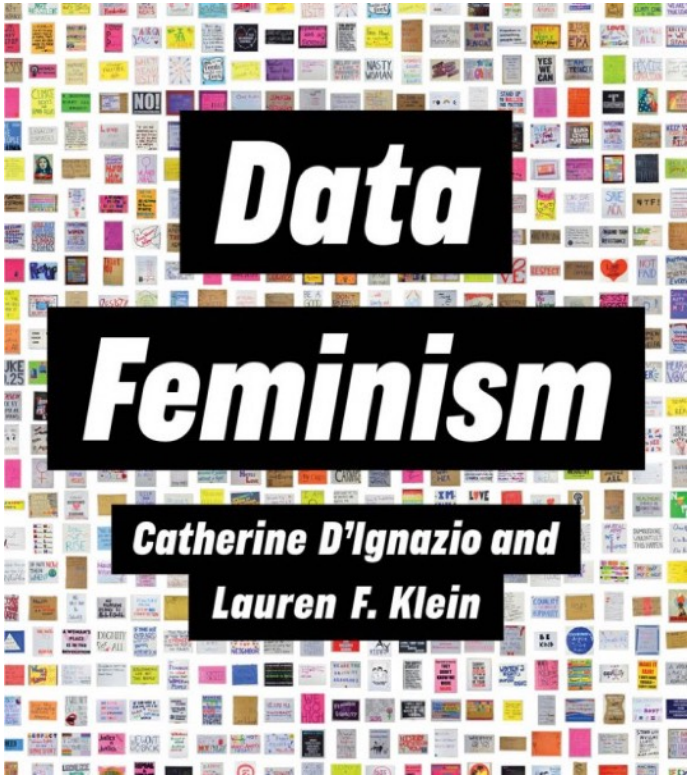


Fig. 1. Culturomic analyses study millions of books at once. **(A)** Top row: Authors have been writing for millennia; ~129 million book editions have been published since the advent of the printing press (upper left). Second row: Libraries and publishing houses provide books to Google for scanning (middle left). Over 15 million books have been digitized. Third row: Each book is associated with metadata. Five million books are chosen for computational analysis (bottom left). Bottom row: A culturomic time line shows the frequency of "apple" in English books over time (1800–2000). **(B)** Usage frequency of "slavery". The Civil War (1861–1865) and the civil rights movement (1955–1968) are highlighted in red. The number in the upper left (1e-4 = 10^{-4}) is the unit of frequency. **(C)** Usage frequency over time for "the Great War" (blue), "World War I" (green), and "World War II" (red).

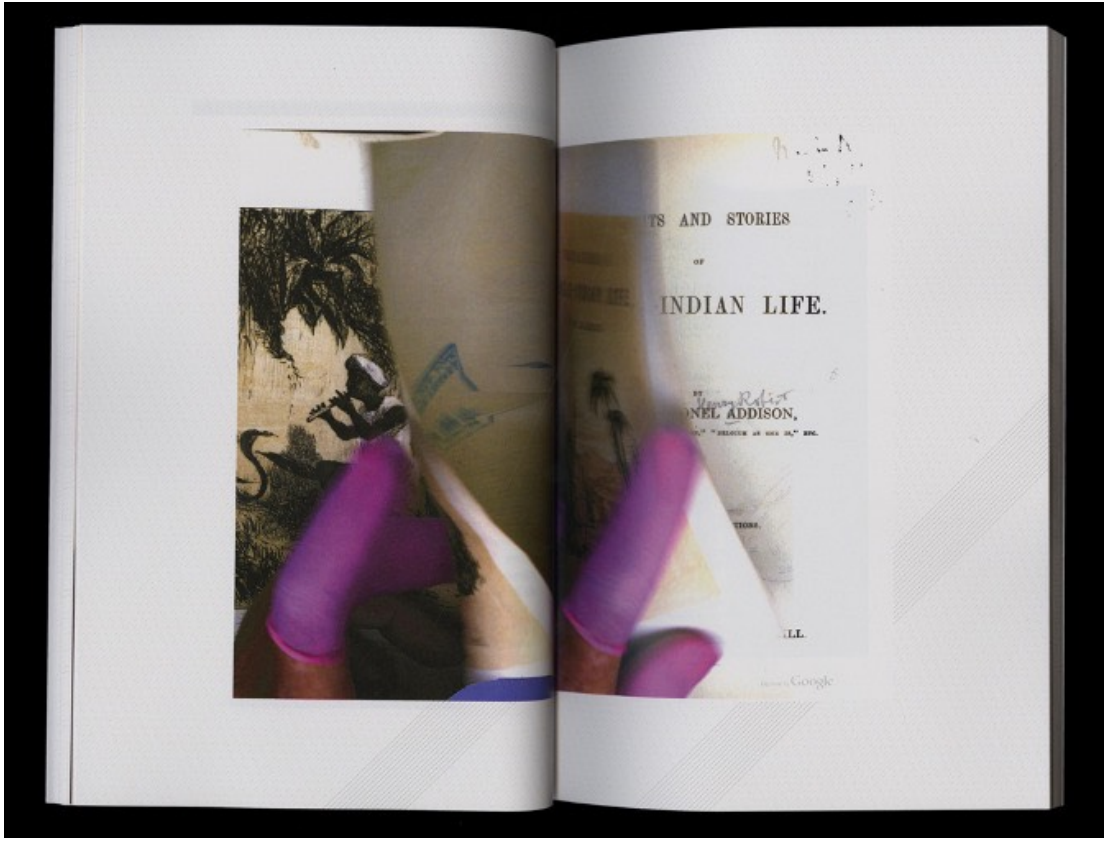
La fascination du *big data…*

« Big Dick Data projects fetishise large size and prioritise it, along with speed, over quality, ignore context and inflate their technical capabilities. They also tend to have little consideration for inequalities or inclusion in the process. Mark Zuckerberg aiming to supersede human senses with AI might be considered one such project, along with software company Palantir's claims about massive-scale datasets. Big Dick Data projects aren't necessarily wholly invalid, but they suck up resources that could be given smaller, more inclusive projects. »



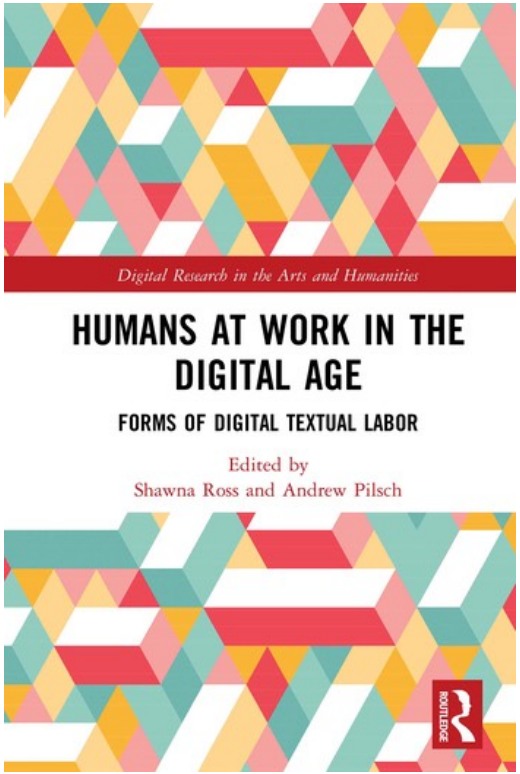
…face à l’écueil de la matérialité

La présence occasionnelle de « mains » dans Google Books souligne que la numérisation n’est pas un bouton magique mais un processus complexe trop souvent dissimulé (avec un recours fréquent au « digital labor »)

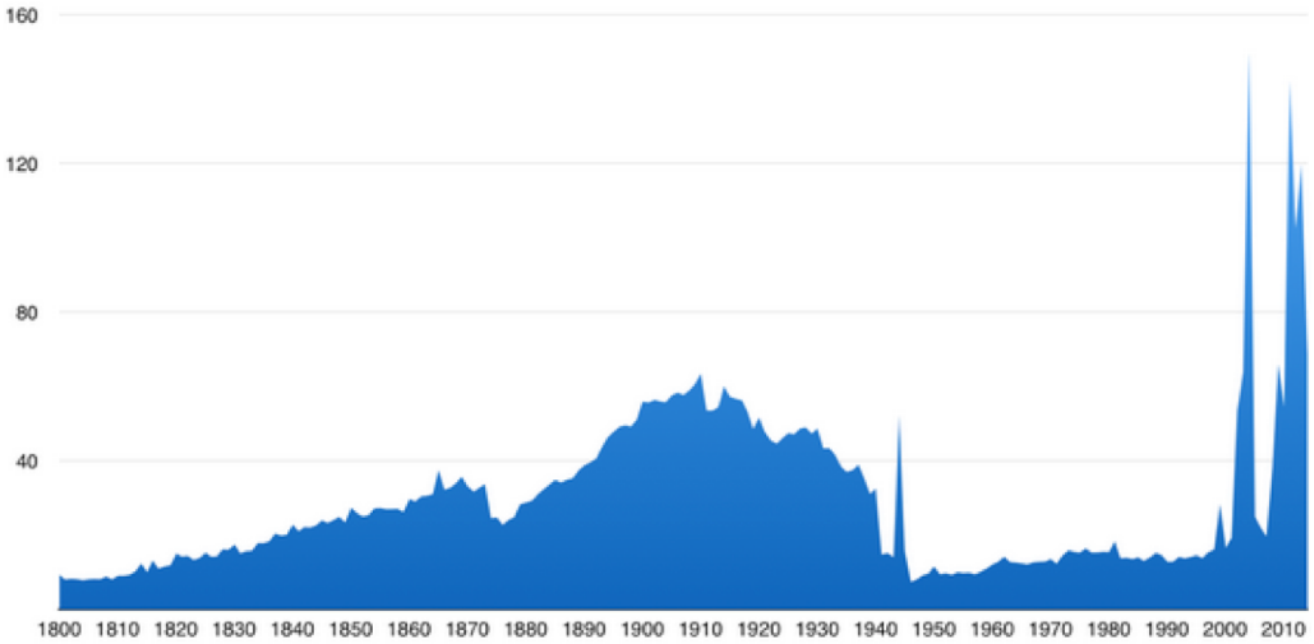


…face à l'écueil de la matérialité

Un récent tweet viral d'Internet Archive a levé un coin de voile sur les conditions de production de l'archive numérisée et du *digital textual labor* avec des divisions du travail sociales, genrées et ethniques.

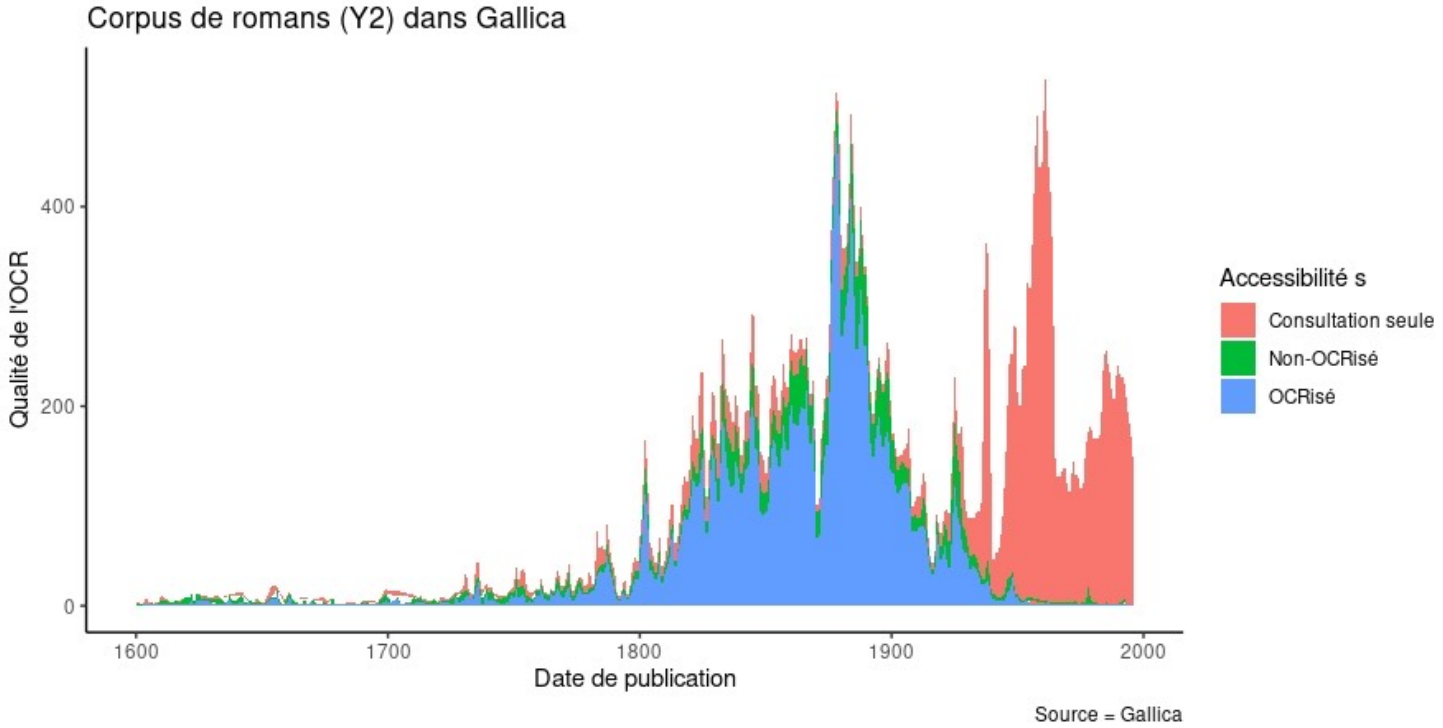


Les biais de collections



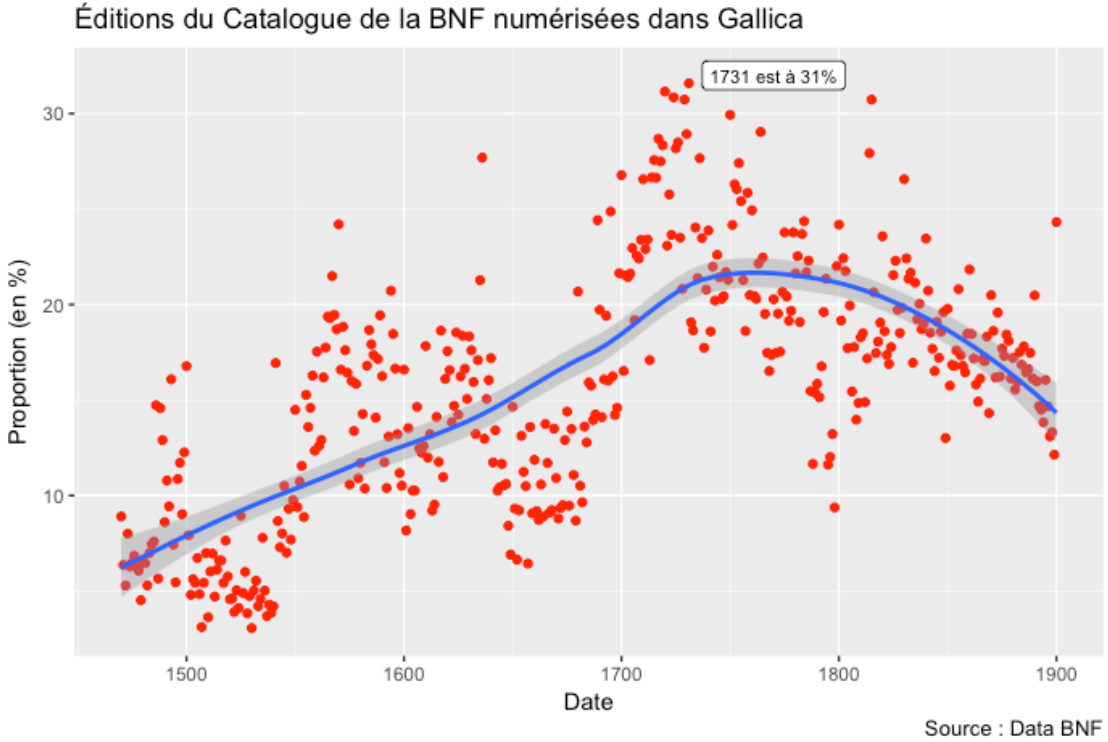
Le trou noir du web entre 1945 et 1990 dans Europeana

Les biais de collections



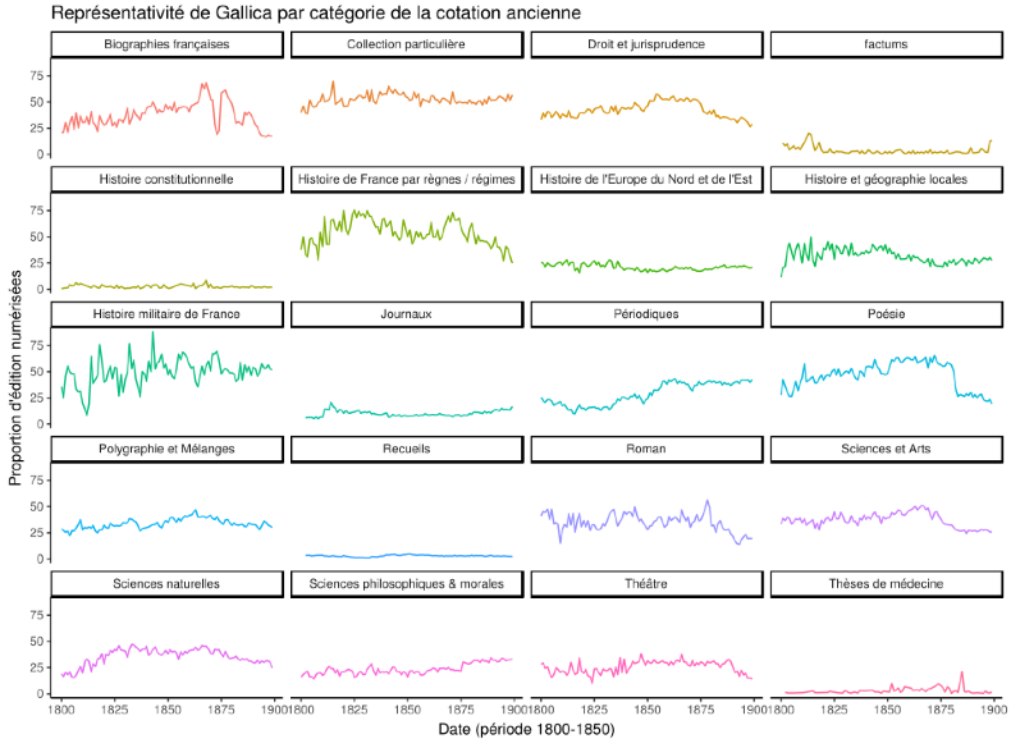
Le mille-feuille des conditions d’accès aux documents sur Gallica

Les biais de collections



Même avant 1945 la représentativité de Gallica évolue…

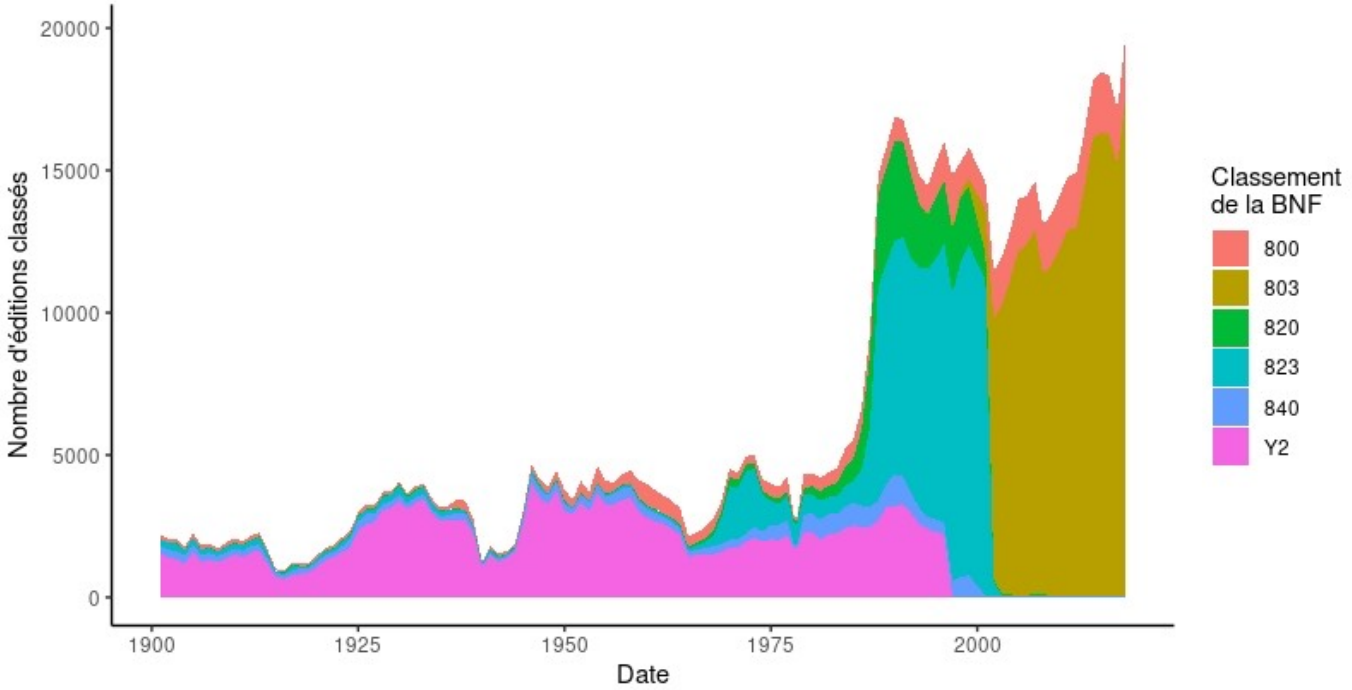
Les biais de collections



…avec d'importantes variabilités selon les thématiques.

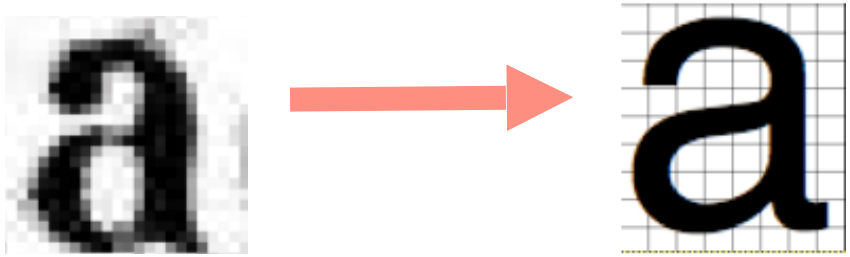
Les biais de collections

La documentation des archives patrimoniales n'est jamais simple : elle hérite d'une longue histoire et des classements successifs adoptés par les bibliothèques



Classifications des romans dans Gallica : de Y2 (jusqu'en 1996) jusqu'au Cadre de classement de la BNF)

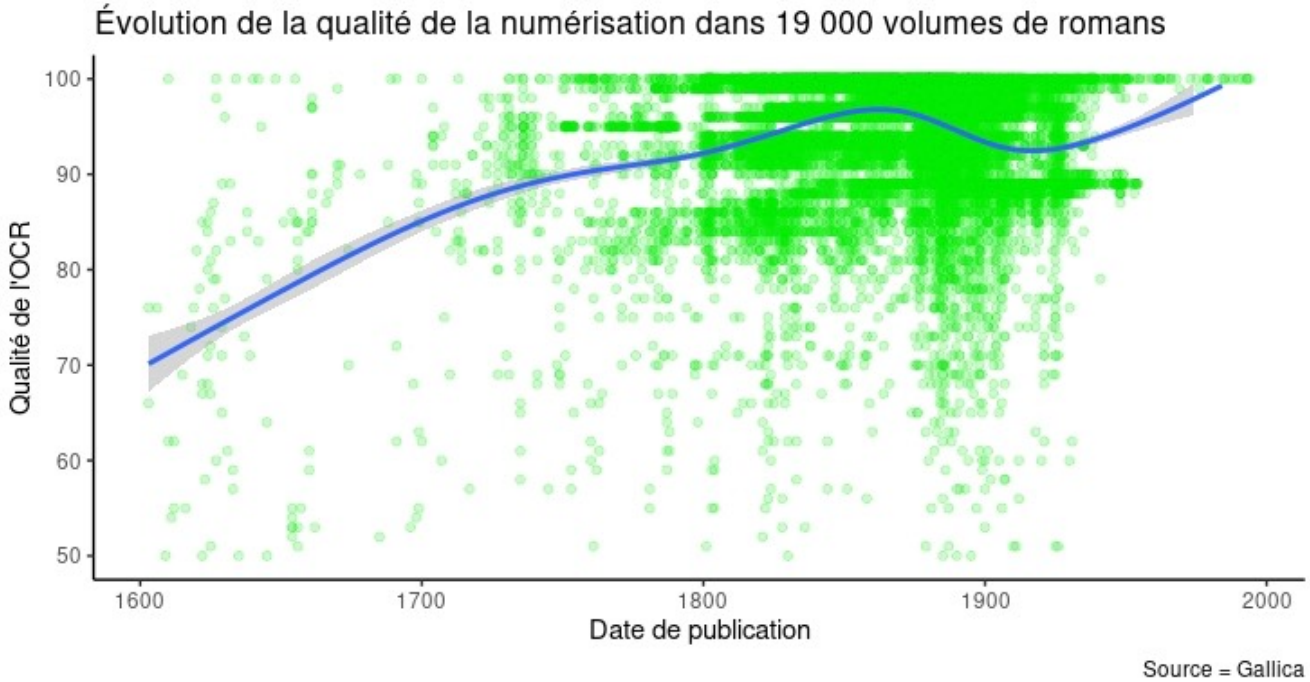
Tous les textes… mais pas toutes les lettres.



Les milliards de pages numérisés ne peuvent pas l'être à la main : on utilise des outils automatisés les OCR avec leurs propres biais (ici avec le S long de l'ancien français).

Tous les textes… mais pas toutes les lettres.

La qualité de l'océrisation est inégalement répartie selon les corpus : les logiciels ont avant tout été conçus pour des imprimés contemporains et fonctionnent de plus en plus difficilement pour les périodes anciennes.



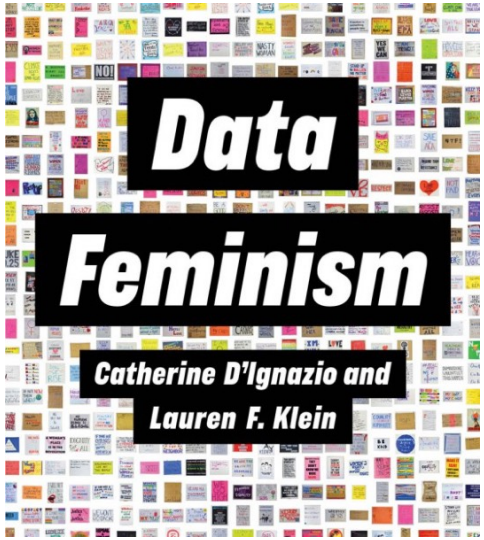
2. Voir l’invisible à distance

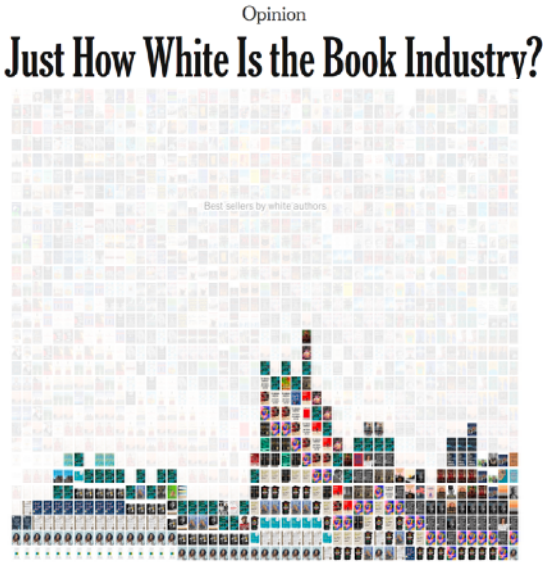
De l’archive-document à l’archive-réseau

Les archives invisibles

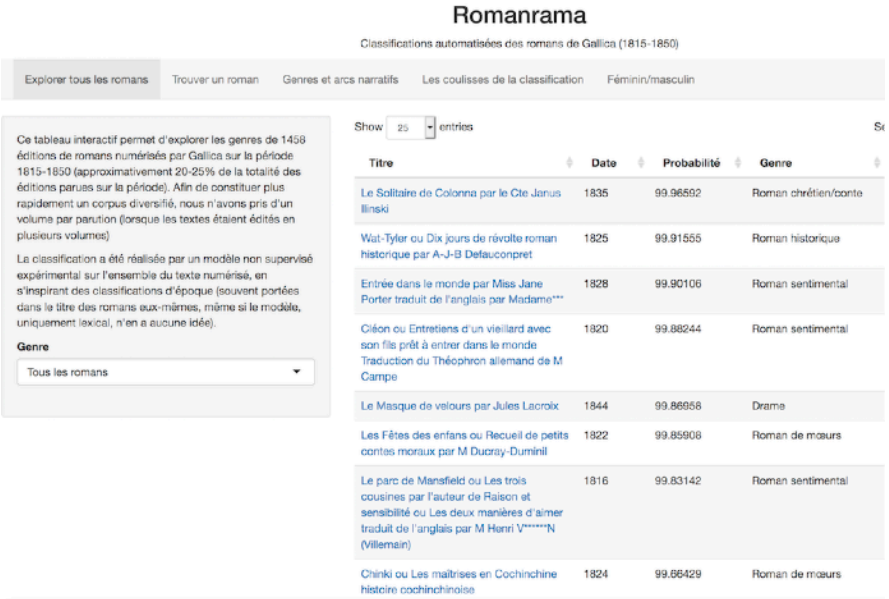
L’enjeu des humanités computationnelles aujourd’hui n’est plus seulement de ne pas déformer les archives existantes mais de questionner les conditions de production des archives et de l’archivage et de se donner les moyens de « produire » des archives occultées

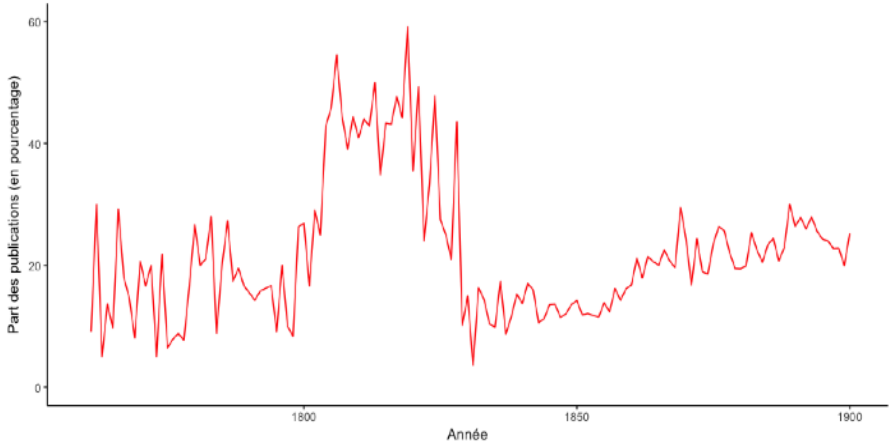
« In calling for a conceptual reorientation from the axis of distant and close to a space defined by multiple dimensions of scale, I seek to make the case that quantitative methods can be used to probe the research questions about gender, race, and their intersection with labor that have thus far proved di cult (although certainly not impossible) to explore. » — Lauren F. Klein (« Dimensions of Scale »)





Surmonter les filtres de lectures

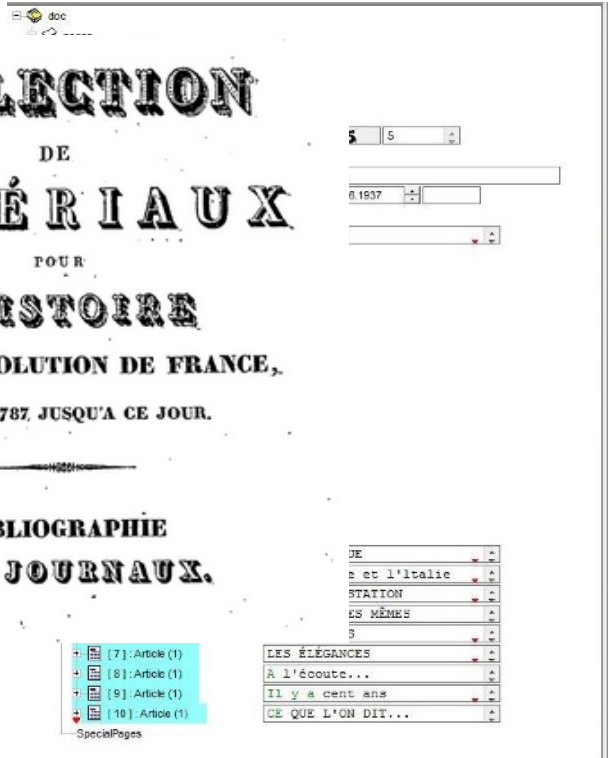


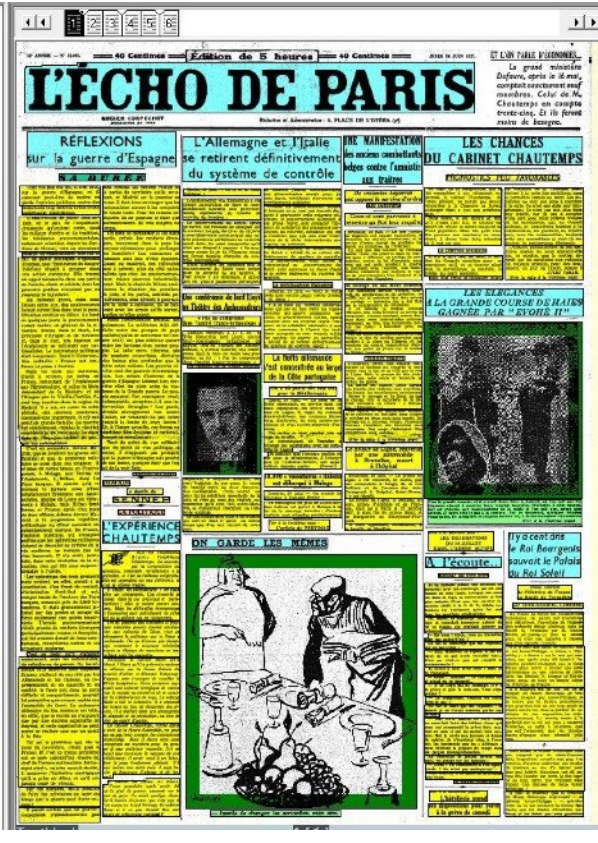


En donnant accès un grand nombre de publications oubliées la numérisation s'accorde avec une vision « décentralisée » de l'histoire culturelle, potentiellement en rupture avec les représentations canoniques.

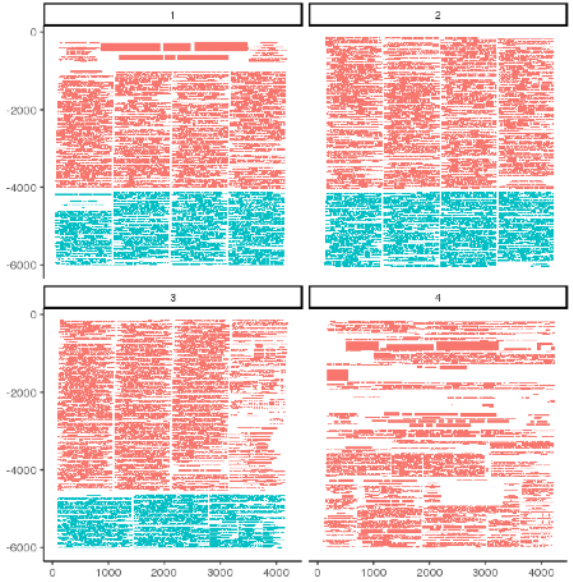
Inventer des archives : le cas de la presse

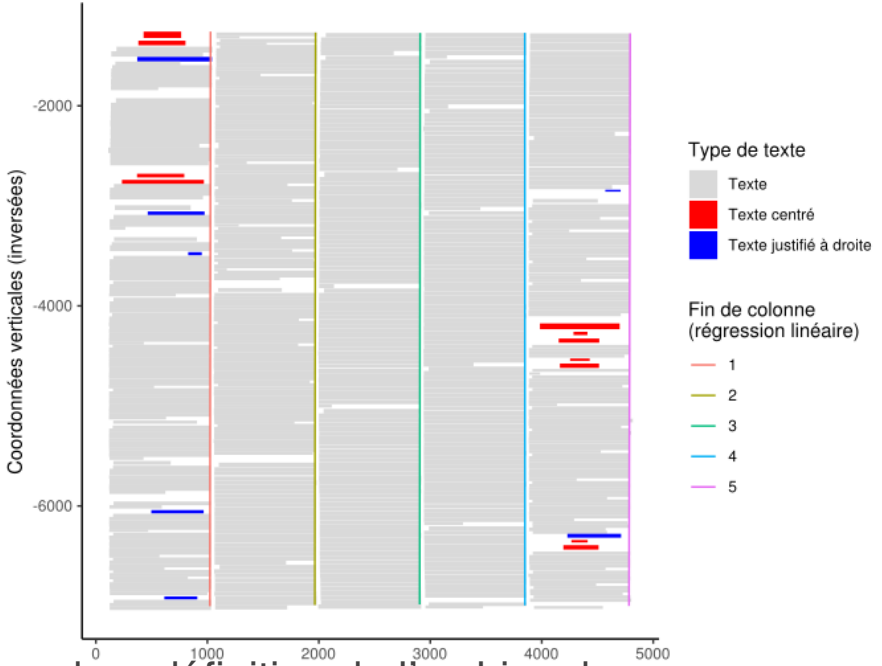
Dans le cas de la presse, l'indexation des archives n'a jamais été satisfaisante… jusqu'à aujourd'hui. Les essais n'ont pas manqué de Deschiens (1828) aux Tables du temps, avec des résultats au mieux limités.





Inventer des archives : le cas de la presse





Le projet Numapresse est investi dans un processus de redéfinition de l'archive de presse avec l'introduction de nouveaux découpages (par article, par paragraphes…) et la production de nouveaux corpus qui n'ont jamais été édités isolément (comme les romans-feuilletons)

Vers une « bibliothèque » de modèles

Un exemple de fiche : le modèle des genres journalistiques de 1840-1860 http://www.numapresse.org/generotheque/items/show/4

Genres journalistiques "1840-1860"

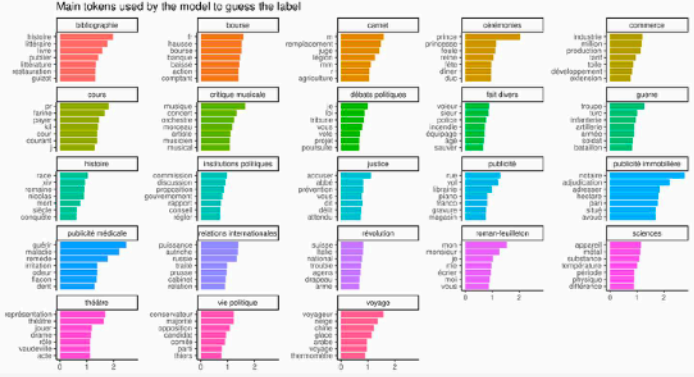
Corpus d'entraînement

1 444 blocs de textes de plus de 100 mots collectés à partir d'une sélection aléatoire de 500 exemplaires de presse numérisés par Gallica entre 1840 et 1860. Le corpus comprend quatre titres actifs sur l'ensemble de la période (le *Journal des débats*, la *Presse*, le *Siècle* et le *Constitutionnel*), deux titres partiellement publiés (le *Pays*, créé en 1849 et le *Courrier français*, disparu en 1851), ainsi que titres avec des numérisations parcellaires sur cette période (le *Figaro*, la *Démocratie pacifique*, le *Corsaire* et l'*Estafette*)

Du fait de la sélection aléatoire les titres les plus tardifs ou les titres avec un volume de publication plus faibles sont moins représentés (par exemple 5 textes pour l'Estafette avec des archives présentes pour la seule année 1852).

Le corpus a été annoté manuellement par blocs de 250 blocs de textes dans un tableur. En raison des imperfections du processus de numérisation, certains blocs de textes ont été subdivisés (s'ils comportaient plusieurs nouvelles différentes). Les textes difficiles à classifier ont été écartés.

Catégorisation



Description

Le modèle 1840-1860 appartient à la série des modèles "générationnels" de Numapresse couvrant les genres journalistiques de la presse quotidienne nationale française du début du 19e siècle à la Seconde Guerre Mondiale.

Modèle

Télécharger le modèle au format R

Corpus

Télécharger le corpus d'entraînement

Format original

Modèle SVM enregistré avec R et Tidysupervise (format .rda)

Auteur

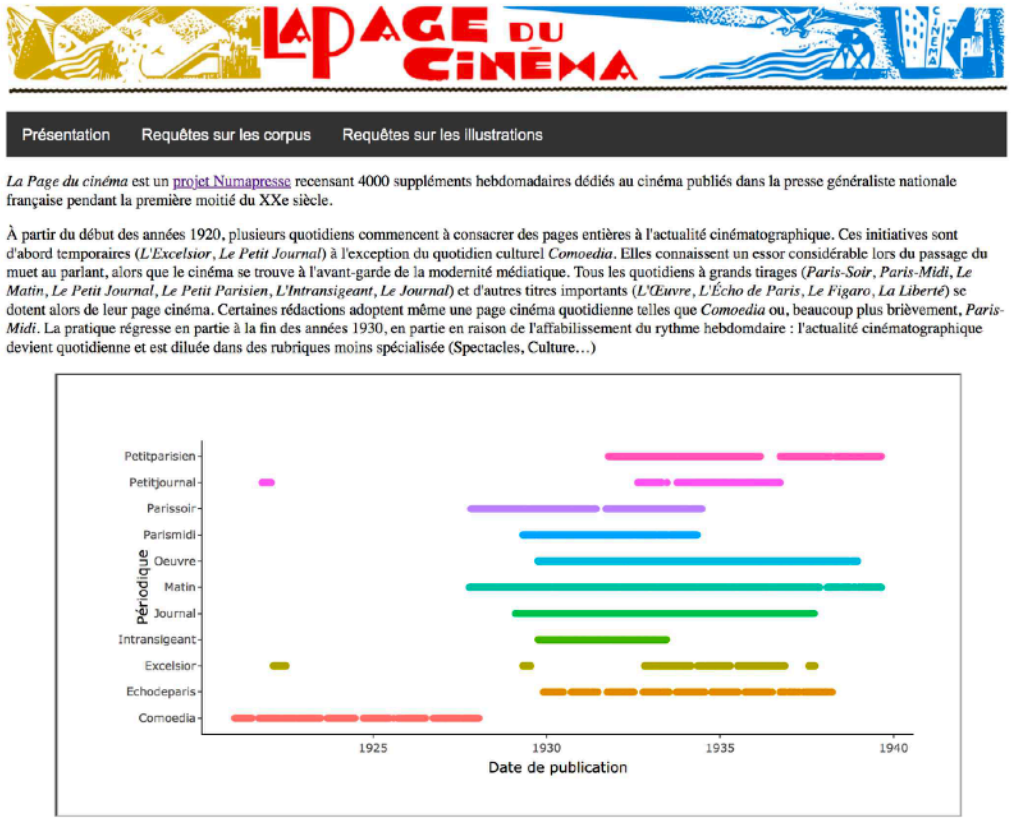
Pierre-Carl Langlais


LA PAGE DU CINÉMA
LA PROPAGANDE A REBOURS
L'activité en France
critiques
L'activité à l'Étranger
ECHOS
SALLES
BABEL
DICTION
RIALTO-CINÉMA
LA VALSE DE L'ADIEU
PARAMOUNT
THÉATRES RÉUNIS TRIVOLI
A MARIVAUX
LA VIE PRIVÉE D'HÉLÈNE DE TROIE
CINÉMA-ABEILLE
BEN-HUR
GAUMONT-PALACE
HATEZ-VOUS LE ROI DES ROIS A L'OPÉRA
AU CAMEO LA GRANDE ÉPREUVE
LE GORILLE
L'ÉTUDIANT PAUVRE AU CINÉ MAX LINDER


théâtre : 47.68%
cinéma : 26.14%
critique littéraire et artistique : 15.4%
cinéma
critique littéraire et artistique
économie
justice
liste des spectacles
négociations internationales
none
programme radio
publicité_affiche
publicité_annonce
reportage
sport
théâtre
vie politique

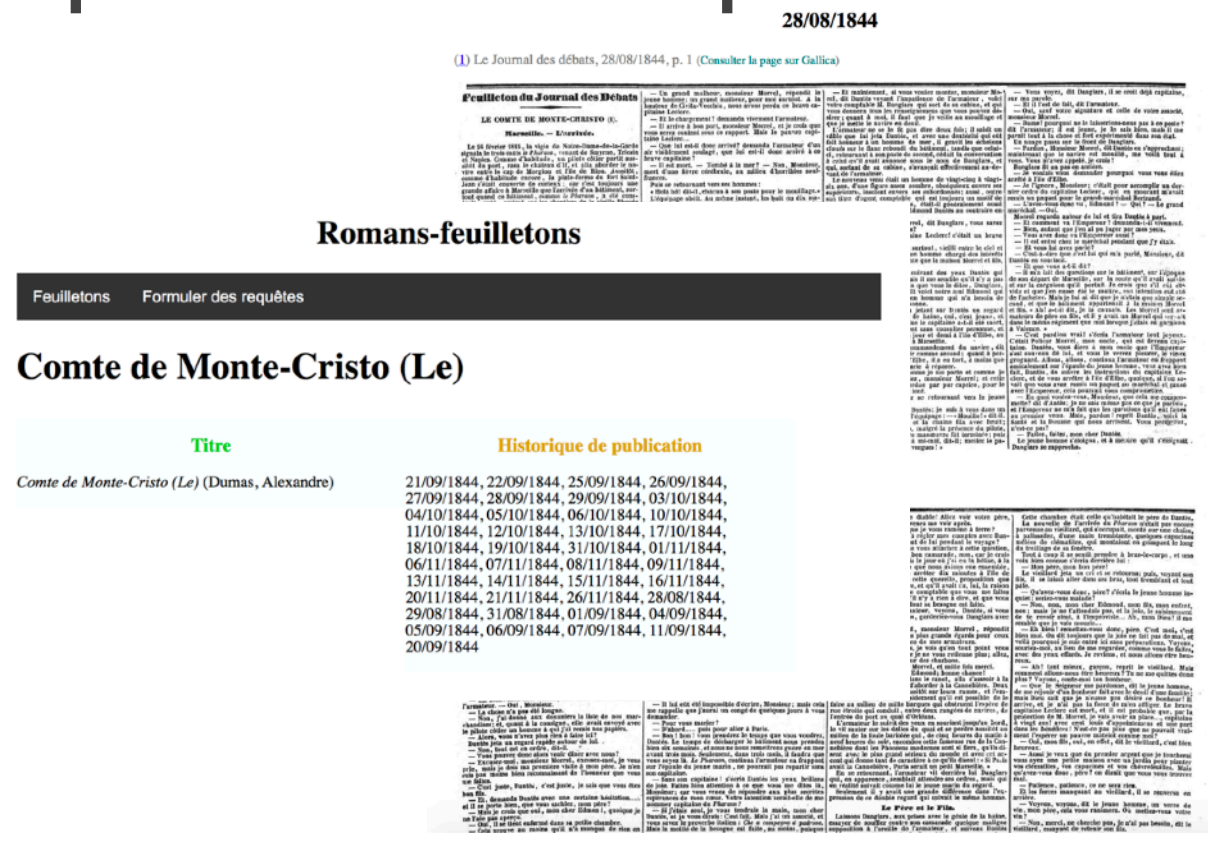
Créer des corpus contextualisé par IA

La classification de masse rend également possible l'exploration de corpus quasi-exhaustifs à l'image du projet « La Page de cinéma » qui contient désormais 4000 suppléments cinéma hebdomadaires publiés dans 10 quotidiens différents : http://www.numapresse.org/exploration/cinema_pages/presentation.php



Créer des corpus contextualisé par IA

Nos premiers essais de « redécoupage » de l'archive de presse : une application de lecture des romans-feuilletons qui isole les feuilletons du reste du journal et les présente dans la continuité : http://www.numapresse.org/exploration/feuilleton/feuilleton_main/comte_de_montecristo_le.html

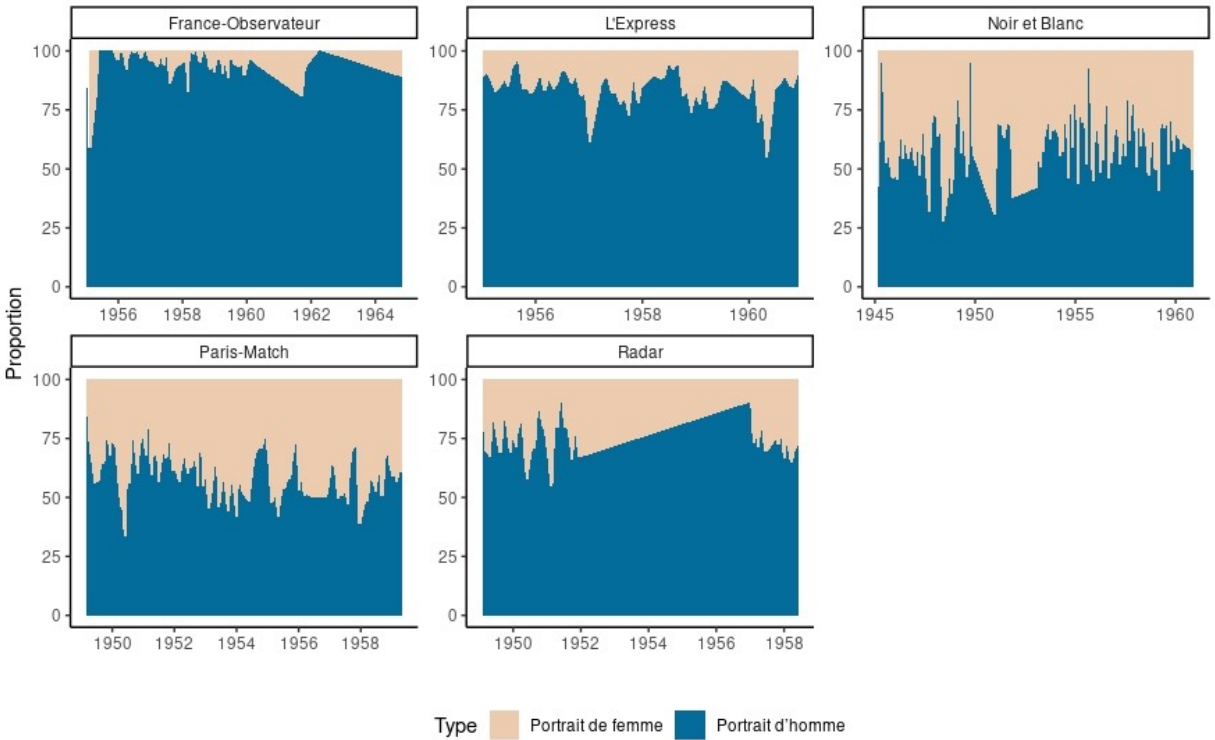


Créer des corpus contextualisé par IA

Pour les illustration, l'impact de la classification est encore plus profond : ces documents ne sont tout simplement pas indexés par défaut http://www.numapresse.org/exploration/cinema_pages/query_illustration.php

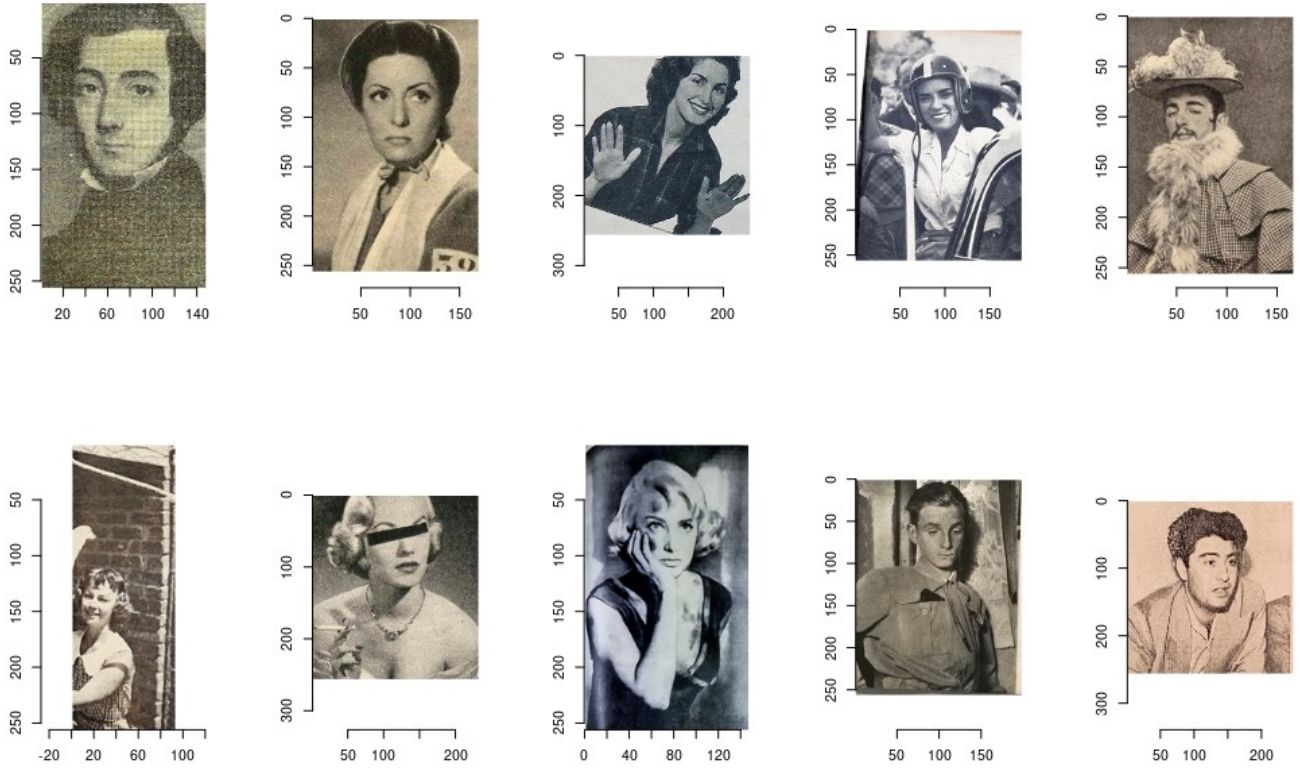
Trouble dans les catégories

Les biais des modèles peuvent être détournés pour analyser des biais sociaux. Ces graphes montre les taux de « féminisation » dans cinq hebdomadaires des années 1950. varient très fortement selon les corpus ce qui reflète assez nettement le rapport entre les politiques éditoriales et les normes de genre

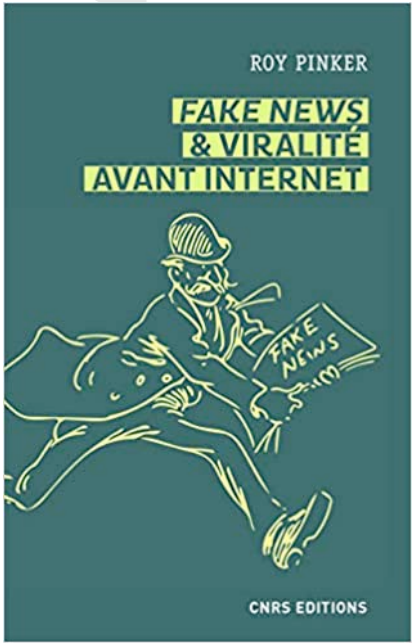


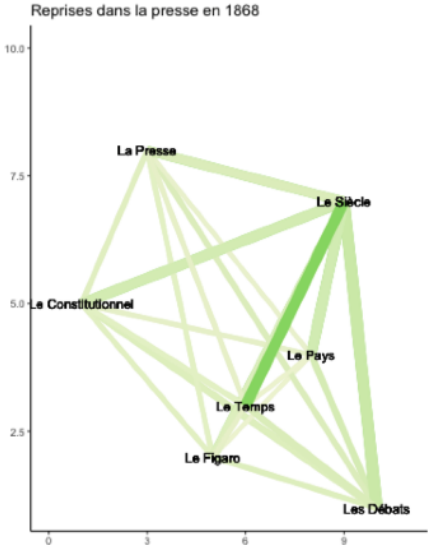
Trouble dans les catégories

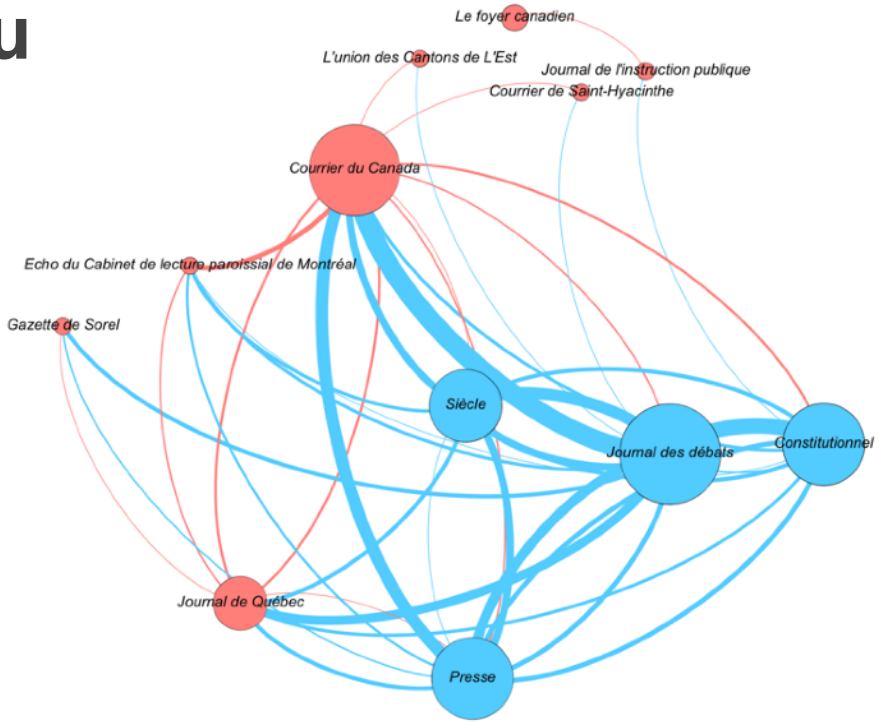
En cherchant des images associant des traits féminins et masculins (> 30% dans les deux cas) nous parvenons à isoler des cas limites : représentations historicisées, ambiguës voire *queer*



Vers l'archive-réseau

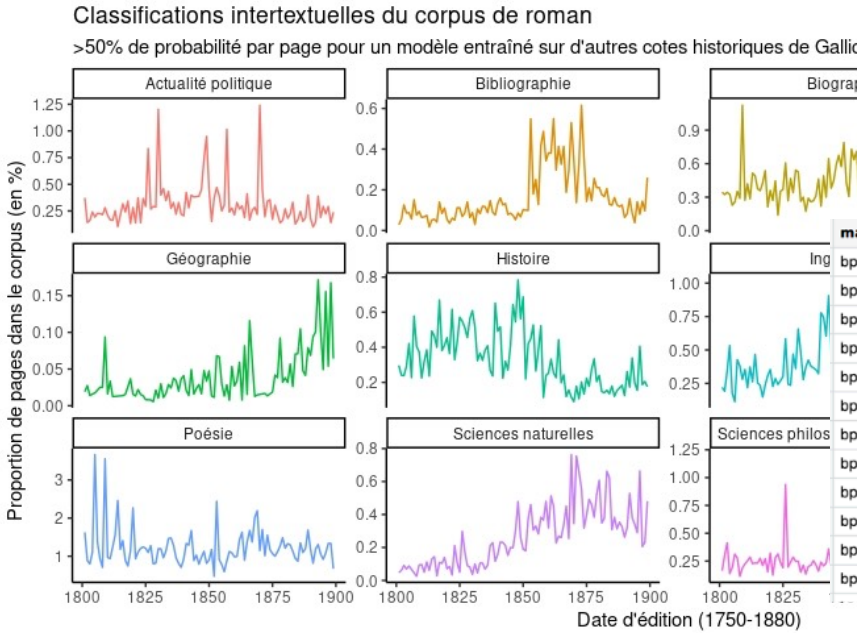






L'éclatement de l'unité documentaire du journal permet d'opérer d'autres rapprochements. D'autres projets de Numapresse portent ainsi sur l'identification automatique des nouvelles virales, ce qui permet de reconstituer des chaînes de transmission successive.

Vers l’archive-réseau

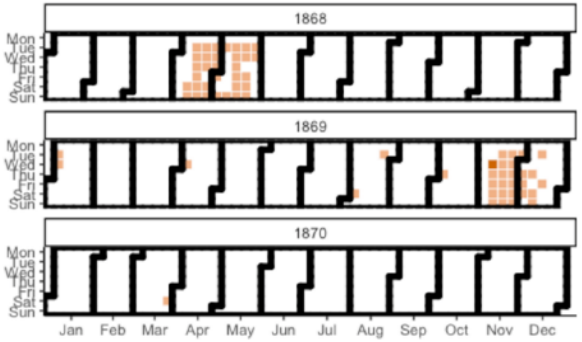


main_id	titre	nom_auteur	prenom_auteur	page	texte
bpt6k6138352h_p1	Ascanio	Dumas	Alexandre	1	NA
bpt6k6138352h_p2	Ascanio	Dumas	Alexandre	2	NA
bpt6k6138352h_p3	Ascanio	Dumas	Alexandre	3	NA
bpt6k6138352h_p4	Ascanio	Dumas	Alexandre	4	_p_ ASCANIO.
bpt6k6138352h_p5	Ascanio	Dumas	Alexandre	5	_p_ PARIS. IMPRIMÉ PAR BÉTHUNE ET PLON, _p_ EL
bpt6k6138352h_p6	Ascanio	Dumas	Alexandre	6	_p_ ASCANIO, _p_ PAR _p_ ALEXANDRE DU MAIS. _
bpt6k6138352h_p7	Ascanio	Dumas	Alexandre	7	NA
bpt6k6138352h_p8	Ascanio	Dumas	Alexandre	8	_p_ ASCANIO. _p_ CHAPITRE PREMIER. _p_ CE QU'C
bpt6k6138352h_p9	Ascanio	Dumas	Alexandre	9	_p_ 2 ASCANIO. _p_ Aussi, comme on s'en doute bi
bpt6k6138352h_p10	Ascanio	Dumas	Alexandre	10	_p_ ASCANIO. S _p_ Les habitants du Petit et du Gr
bpt6k6138352h_p11	Ascanio	Dumas	Alexandre	11	_p_ .4 ASCANIO. _p_ son corps, malgré les saints re
bpt6k6138352h_p12	Ascanio	Dumas	Alexandre	12	_p_ ASCANIO. 4 _p_ La fenêtre n'était donc défend

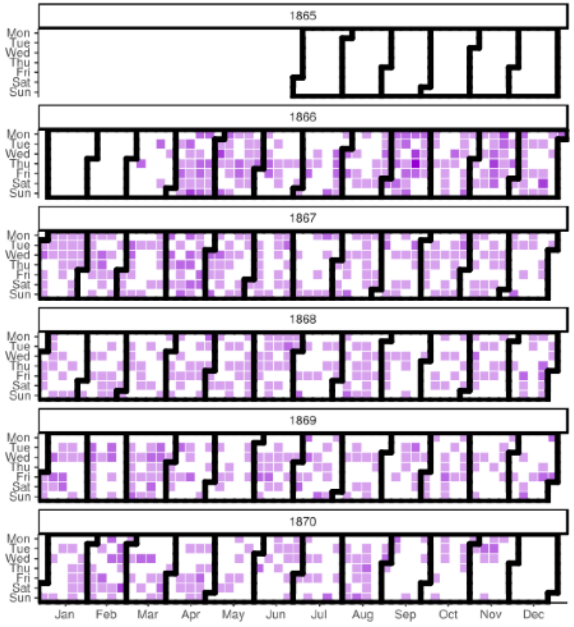
L’éclatement des cadres usuels de l’archivage affecte aussi les monographies. Les projets de cultural analytics comme le notre privilégient l’indexation au niveau de la page voire de sous-éléments internes à la page (comme les illustrations).

Vers l’archive-réseau

L’extraction des signatures rend possible l’identification de profils sociologiques à des périodes où la documentation externe fait défaut (avant 1880). Dans la *Liberté* de 1865 à 1870, les femmes occupent des positions beaucoup plus précaires que les hommes.



Judith Gauthier



Wilfried de Fonvielle

Conclusion

Du journalisme à la littérature et au-delà

Ce que l’archive change à la classification

Un possible effet retour des études en histoire culturelle avec l’intelligence artificielle : mieux informer les biais sous-jacents aux outils contemporains.

Au-delà des images réelles : vers l’exploration des images possibles.

De nouveaux projets d’intelligence artificielle sont capables de créer des images qui n’ont jamais existé mais qui font sens dans l’univers culturel d’où elles sont extraites. Ici, il y a une représentation très rétrofuturiste des Iphone par décennie.