Le potentiel de recherche inédit des corpus géants

Perspectives sur la publication et les revues savantes

Pierre-Carl Langlais
Paris-Sorbonne

Présentation

1. **Text mining & grands corpus de recherche**. Une histoire déjà longue…

2. **De la bibliométrie au études scientifiques quantitatives**. L'émergence d'un nouvel écosystème de recherche

3. **Les collections sont-elles des données comme les autres ?** La construction d'un nouveau service à la recherche

4. **Des grands corpus au grands modèles de langues**. Une ressource stratégique

1. Text mining & grands corpus de recherche

Une histoire déjà longue

Les grands corpus de recherche : une obsession ancienne

L'indexation des archives périodiques n'a jamais été satisfaisante... jusqu'à aujourd'hui. Déjà en 1828, l'un des premier historiens de la presse français, Deschiens réalisait avec horreur que le catalogage des articles journalistiques et savants sur la révolution française lui prendrait plusieurs vies.


COLLECTION

DE

MATÉRIAUX

POUR

L'HISTOIRE

DE LA RÉVOLUTION DE FRANCE,

DEPUIS 1787 JUSQU'A CE JOUR.

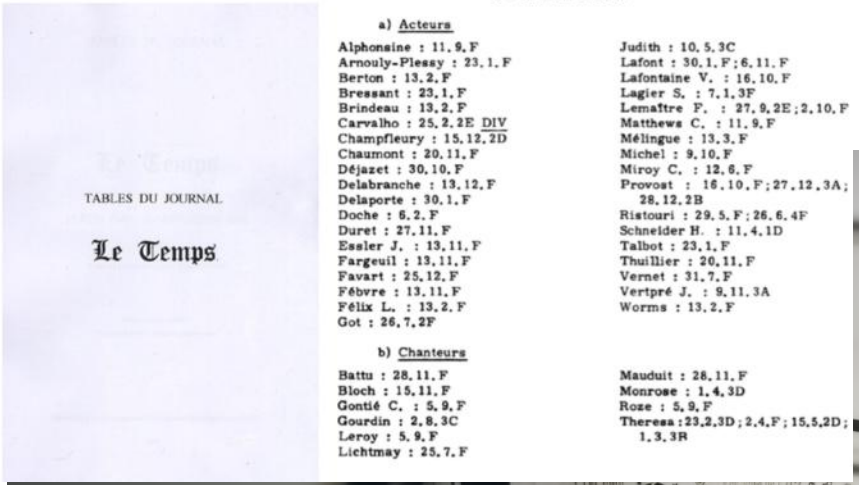
BIBLIOGRAPHIE

DES JOURNAUX.

« J'ai eu le projet de donner une table générale de ma collection ; mais à peine ce travail était-il commencé que j'ai reconnu qu'il ne fallait pas moins de 25 à 30 volumes, pour les tables des deux premières divisions. J'ai bien vite renoncé à pareille entreprise » (p. XII-XIII)

Les grands corpus de recherche : une obsession ancienne

Au cours du 20e siècle, les méga-corpus sont principalement appréhendés via les microfilms et les index. Et ce complexe technique affecte toujours les corpus numériques aujourd'hui : pour des raisons de coûts, les numérisations sont souvent d'abord réalisées à partir des microfilms.



TABLES DU JOURNAL

Le Temps

e - INTERPRETES

a) Acteurs

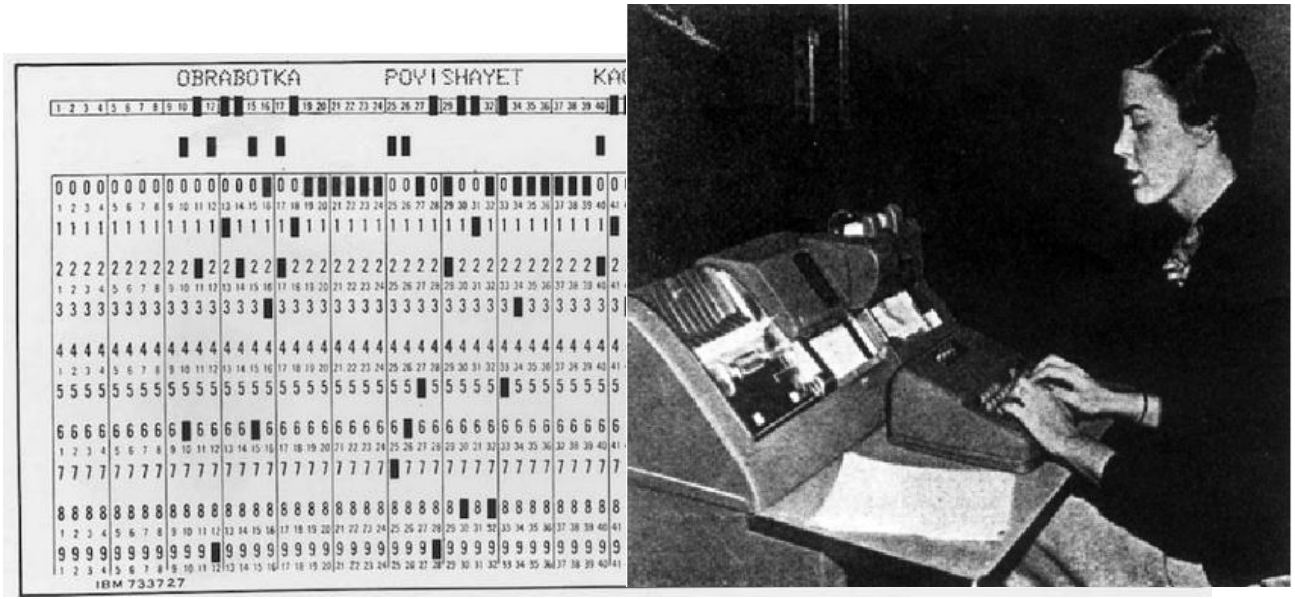
Alphonsine : 11, 9, F
Arnouly-Plessy : 23, 1, F
Berton : 13, 2, F
Bressant : 23, 1, F
Brindeau : 13, 2, F
Carvalho : 25, 2, 2E DIV
Champfleury : 15, 12, 2D
Chaumont : 20, 11, F
Déjazet : 30, 10, F
Delabranche : 13, 12, F
Delaporte : 30, 1, F
Doche : 6, 2, F
Duret : 27, 11, F
Essler J. : 13, 11, F
Fargeuil : 13, 11, F
Favart : 25, 12, F
Février : 13, 11, F
Félix L. : 13, 2, F
Got : 26, 7, 2F
Judith : 10, 5, 3C
Lafont : 30, 1, F ; 6, 11, F
Lafontaine V. : 16, 10, F
Lagier S. : 7, 1, 3F
Lemaître F. : 27, 9, 2E ; 2, 10, F
Matthews C. : 11, 9, F
Mélingue : 13, 3, F
Michel : 9, 10, F
Miroy C. : 12, 6, F
Provost : 16, 10, F ; 27, 12, 3A ; 28, 12, 2B
Ristouri : 29, 5, F ; 26, 6, 4F
Schneider H. : 11, 4, 1D
Talbot : 23, 1, F
Thuillier : 20, 11, F
Vernet : 31, 7, F
Vertpré J. : 9, 11, 3A
Worms : 13, 2, F

b) Chanteurs

Battu : 28, 11, F
Bloch : 15, 11, F
Gontié C. : 5, 9, F
Gourdin : 2, 8, 3C
Leroy : 5, 9, F
Lichtmay : 25, 7, F
Mauduit : 28, 11, F
Monrose : 1, 4, 3D
Roze : 5, 9, F
Theresa : 23, 2, 3D ; 2, 4, F ; 15, 5, 2D ; 1, 3, 3R

La préhistoire de la numérisation : des essais précoces et… décevants

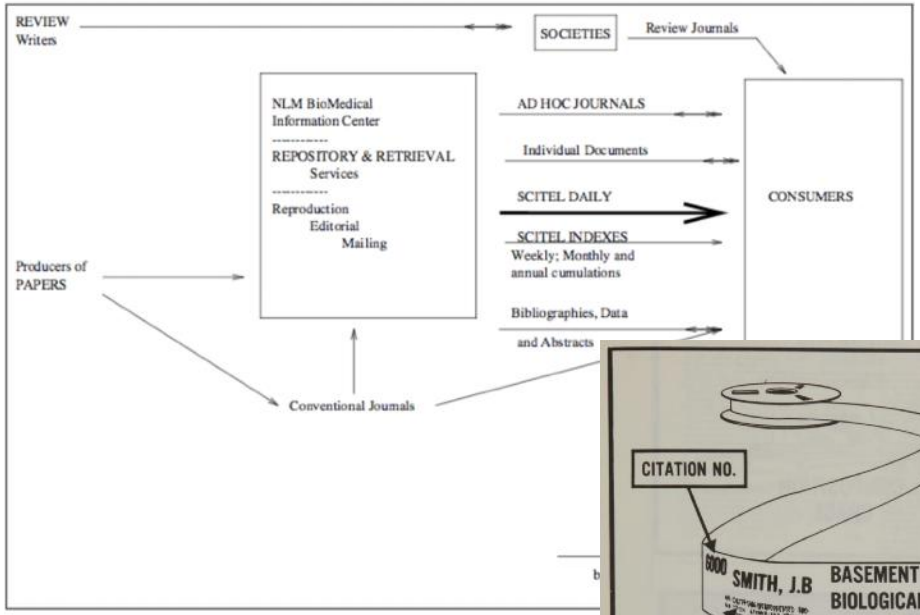
Dès ses débuts, l’ordinateur est pensé comme une machine à lire des textes — en partie sous l’effet de la résolution du code Enigma sur la structuration du champ informatique. Et si au fond, les langues, les formes d’expressions et les représentations culturelles étaient un code comme un autre.

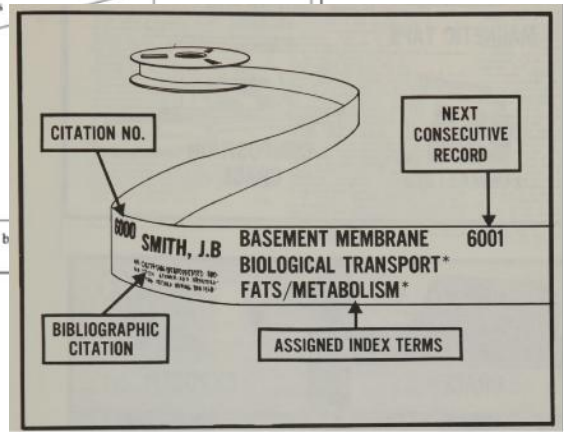


Les débuts du traitement du langage naturel du premier été de l’IA : l’expérience de Georgetown

La préhistoire de la numérisation : des essais précoces et… décevants

Dès le début des années 1950, l'administration fédérale américaine envisage de créer un grand « système d'information centralisé » publiant jusqu'à 1 millions d'articles scientifiques par an. Le *Citation index* de Garfield n'était qu'une pièce de ce vaste projet qui ne verra finalement pas le jour.





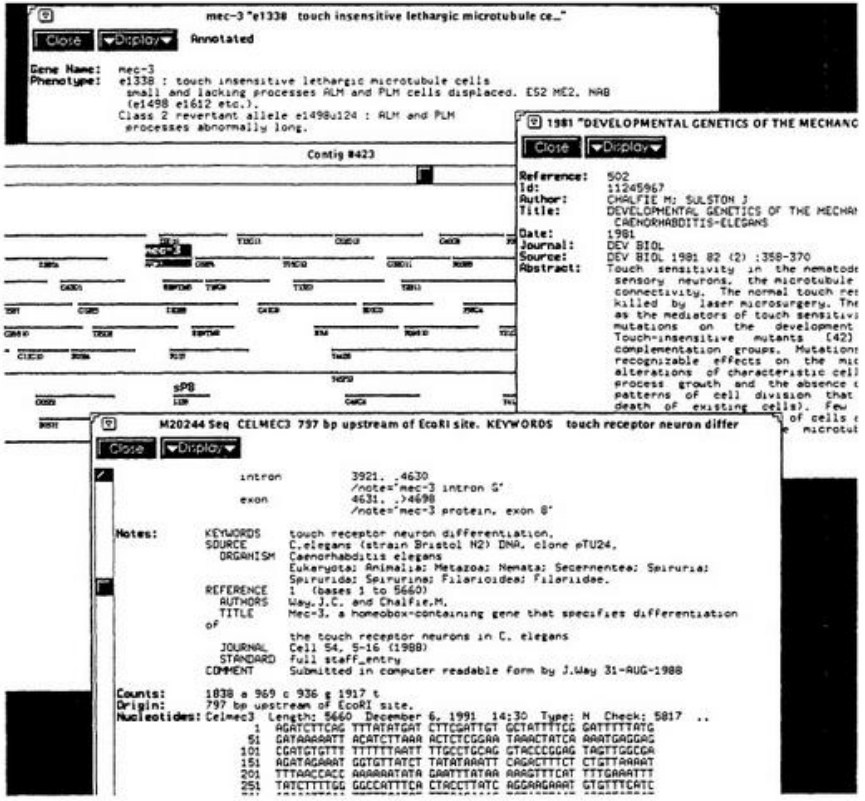
Des essais précoces et… décevants

Les infrastructures scientifiques ne communiquent pas entre elle. L'échange des données est techniquement possible mais long et compliqué. À défaut il faut souvent faire le déplacement pour consulter un corpus : la pesanteur des corpus numériques contraste alors avec la fluidité du papier. Des projets très ambitieux comme la New York Times Information Bank restent à usage internet et on une portée très limitée.

Un terminal de la New York Times Information Bank en 1975 © NY Times

Une si lente révolution numérique

Internet marque une résolution radicale au pris d'une grande simplification : un protocole de communication unique pour toutes les machines. En quelques années à peine, toutes les infrastructures scientifiques sont dépassées et préparent leurs migrations.



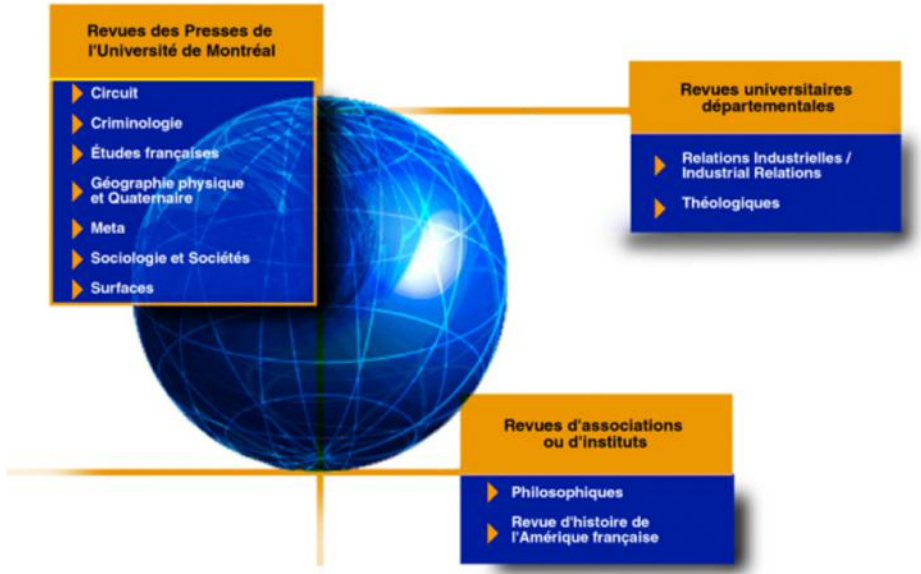
Un cas d'école : le Worm Community System bascule sur Internet après 1994 et inspire une définition de référence des infrastructures scientifiques (Star & Ruhleder, 1996)

Une si lente révolution numérique

Les corpus scientifiques et patrimoniaux deviennent universellement accessibles à partir du début des années 2000 et changent immédiatement complètement les conditions de recherche.

Pour vous procurer et installer les modules d'extension (*plug-ins*) nécessaires à la visualisation des textes en format SGML et PDF (respectivement *Panorama Viewer*, de Interleaf, et *Acrobat Reader*, de Adobe), consultez la rubrique Logiciels à télécharger.

Ce site est optimisé pour les navigateurs Internet Explorer 4 (ou supérieur) et Netscape 4 (ou supérieur). Résolution minimale recommandée : 800 x 600 en 65535 couleurs (16 bits).



Le portrail érudit en 2000 avec un double-format de publication qui se maintient jusqu'à aujourd'hui : langage à balises (SGML puis HTML/XML) et PDF

Une si lente révolution numérique

Les corpus scientifiques et patrimoniaux deviennent universellement accessibles à partir du début des années 2000 et changent immédiatement complètement les conditions de recherche.

Early CANADIANA online

A project to provide enhanced access to Canada's published heritage

ENGLISH

Notre MÉMOIRE en ligne

Un projet visant à favoriser l'accès au patrimoine imprimé du Canada

FRANÇAIS

En 2003, un chercheur identifie une occurrence beaucoup plus précoce qu'attendu du hockey sur glace (dès 1825)

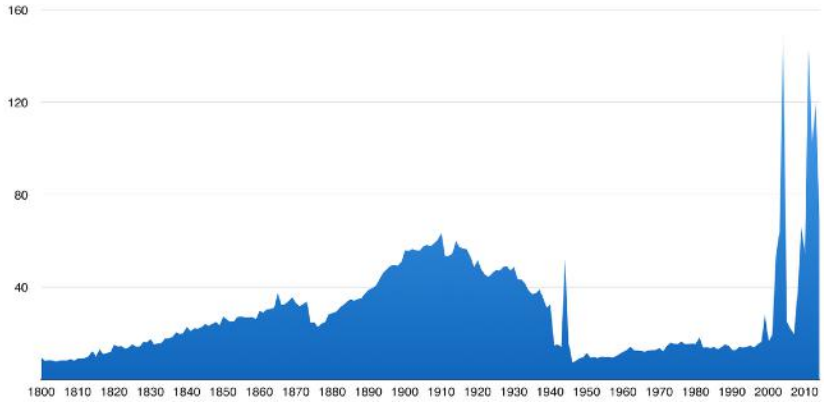
Une si lente révolution numérique

Les premières bibliothèques numériques de presse reconduisent le dispositif du NYT Information Banks en transcrivant à la main une sélection d'article — ce qui apparaît rapidement comme une tâche impossible.

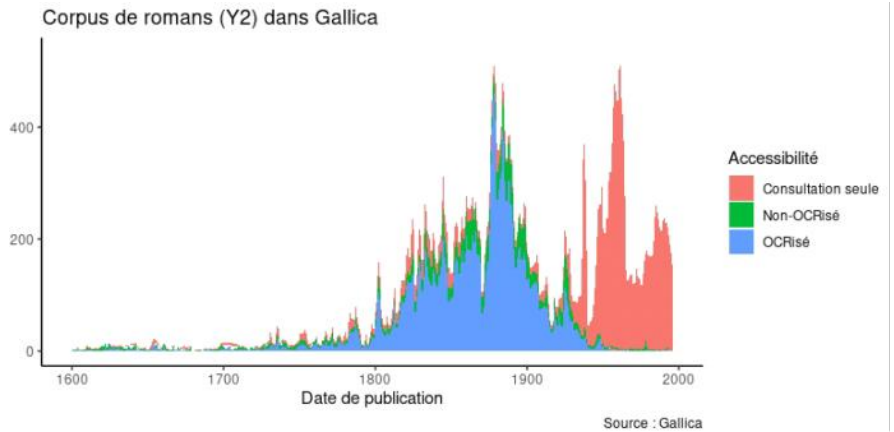
Les outils de reconnaissance optique des caractère permettent de reconnaître le texte de tout le journal. Malgré les erreurs fréquentes, c'est un changement profond de conception du texte journalistique : il n'y a pas de sélection du texte en amont et l'article redevient ancré dans sa page.

Vidéo de présentation du projet European Newspaper en 2015 : le texte sort littéralement du journal mais n'en est jamais totalement coupé.

Une si lente révolution numérique



Le « trou noir du domaine public » chez Europeana



Les romans publiés après 1930 sont majoritairement disponibles en consultation seule à la BNF.

En dehors du Canada, les lois sur la protection du droit d'auteur sont assez contraignantes et il existe peu d'exceptions pour les usages de recherche. Concrètement, les bases patrimoniales sont paradoxalement mieux renseignées sur la seconde moitié du 19e siècle que sur la seconde moitié du 20e siècle.

2. De la bibliométrie au études scientifiques quantitatives

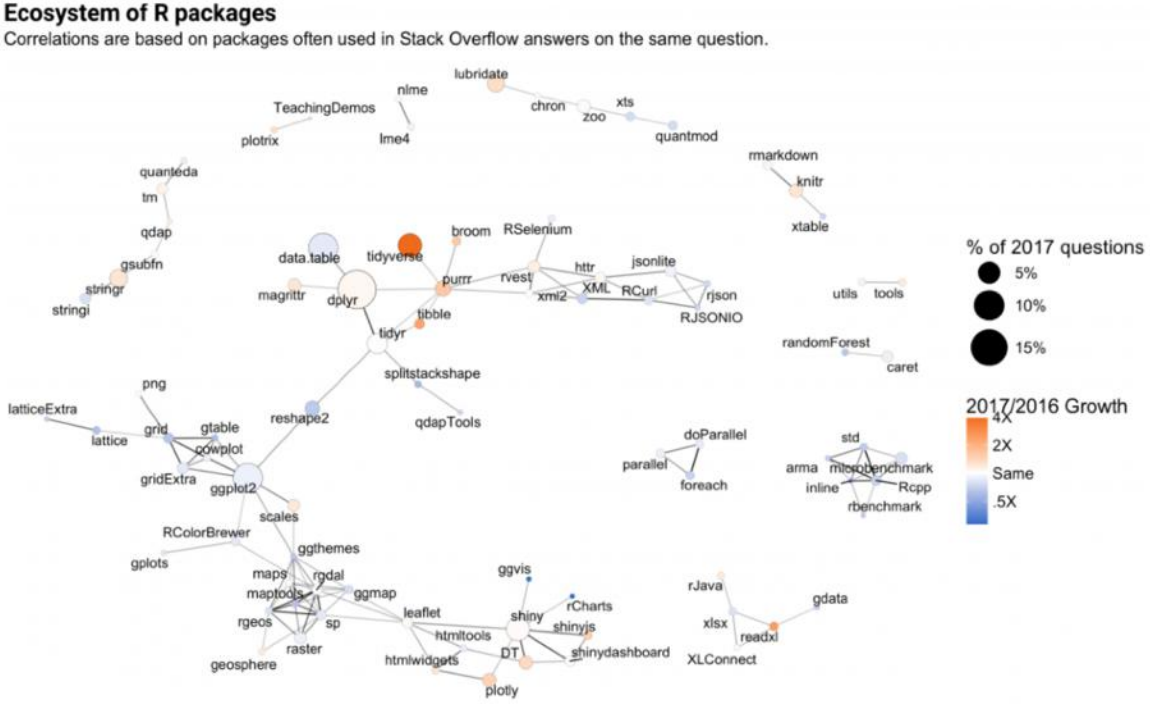
L’émergence d’un nouvel écosystème de recherche

L’émergence d’un nouvel écosystème de recherche

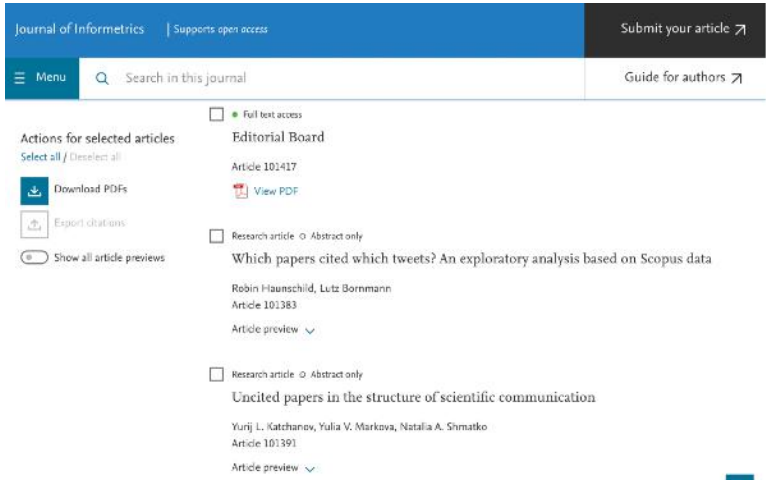
Le champ de la bibliométrie s’est historiquement structuré autour du Science Index de Garfield devenu Web of Science. Depuis une dizaine d’années, ces grandes bases commerciales fermées sont concurrencées par de nouveaux acteurs commerciaux et non-commerciaux, mais aussi par l’émergence de grandes plateformes régionales

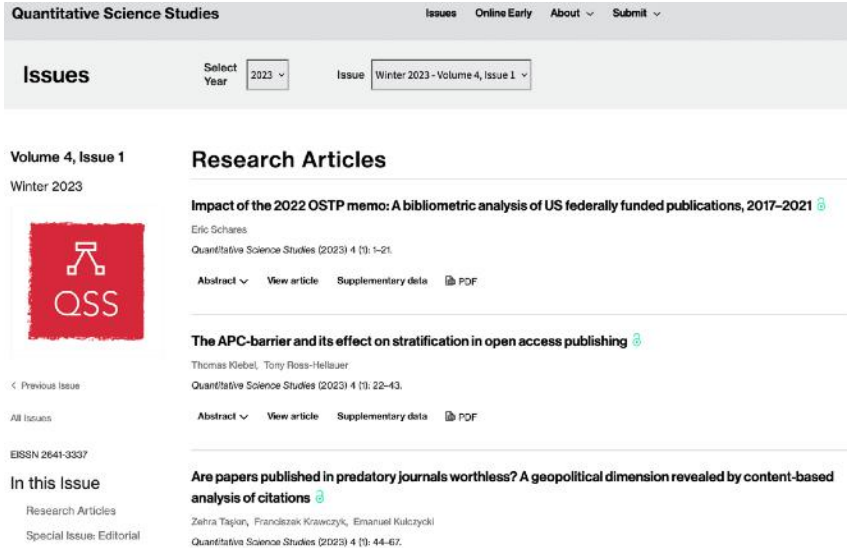
L’émergence d’un nouvel écosystème de recherche

Internet n’est pas seulement le lieu d’échange du texte mais aussi du code. À partir du début des années 2010, on assiste à une véritable vague de programmes et d’extensions spécialisés, souvent construits par leurs chercheurs eux-mêmes ou de proches collaborateurs techniques, en fonction des besoins spécifiques de leur champ de recherche.



L’émergence d’un nouvel écosystème de recherche

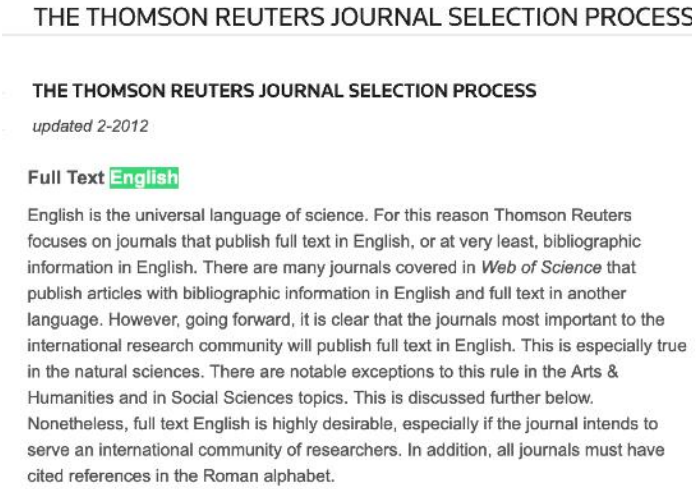




Depuis quelques années, la disponibilité de sources alternatives en libre accès a provoqué l’émergence d’une discipline plus large : les Quantitative Science Studies.

Un cas emblématique : le multilinguisme

Jusqu'à la fin des années 2010 il n'existait pas d'études fiables sur le multilinguisme en recherche. Le Web of Science encourageait explicitement les revues à adopter la langue anglaise et ignorait massivement les productions en d'autres langues.



THE THOMSON REUTERS JOURNAL SELECTION PROCESS

THE THOMSON REUTERS JOURNAL SELECTION PROCESS

updated 2-2012

Full Text English

English is the universal language of science. For this reason Thomson Reuters focuses on journals that publish full text in English, or at very least, bibliographic information in English. There are many journals covered in *Web of Science* that publish articles with bibliographic information in English and full text in another language. However, going forward, it is clear that the journals most important to the international research community will publish full text in English. This is especially true in the natural sciences. There are notable exceptions to this rule in the Arts & Humanities and in Social Sciences topics. This is discussed further below. Nonetheless, full text English is highly desirable, especially if the journal intends to serve an international community of researchers. In addition, all journals must have cited references in the Roman alphabet.

95,86%

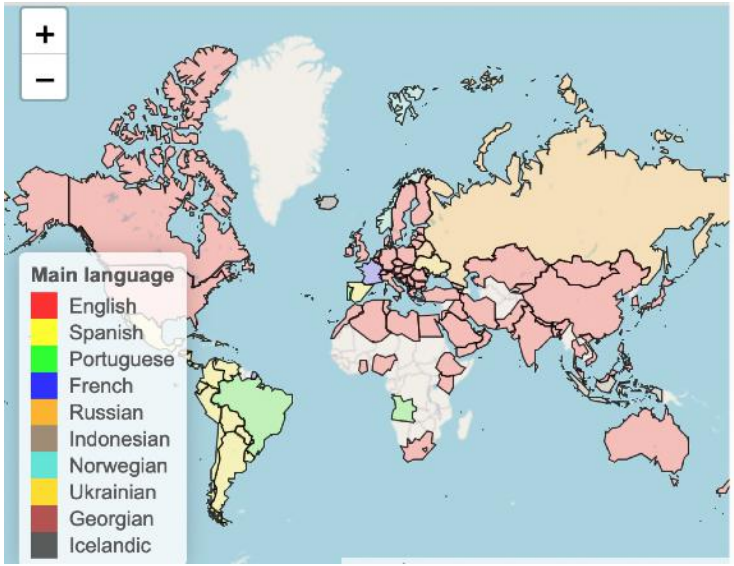
(Web of Science)

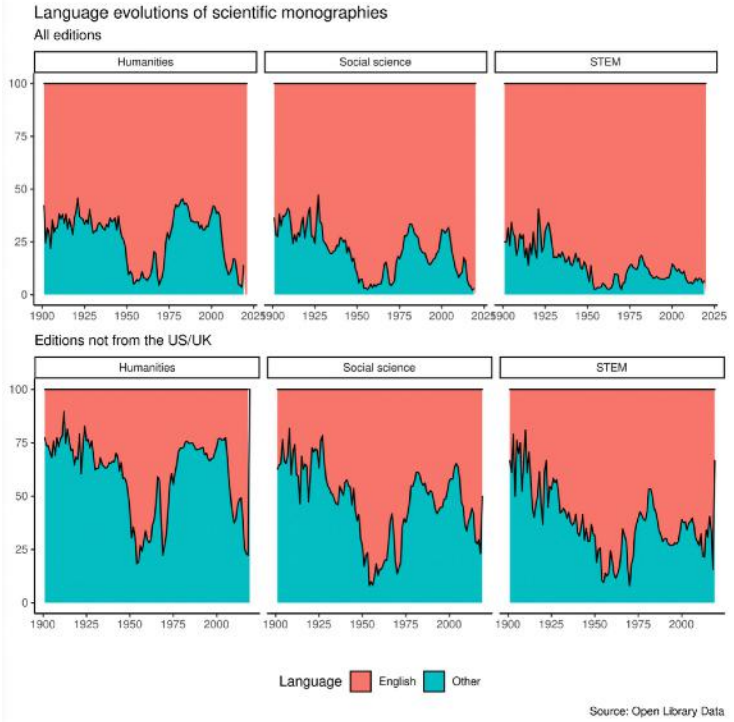
84,35%

(Scopus)

Un cas emblématique : le multilinguisme

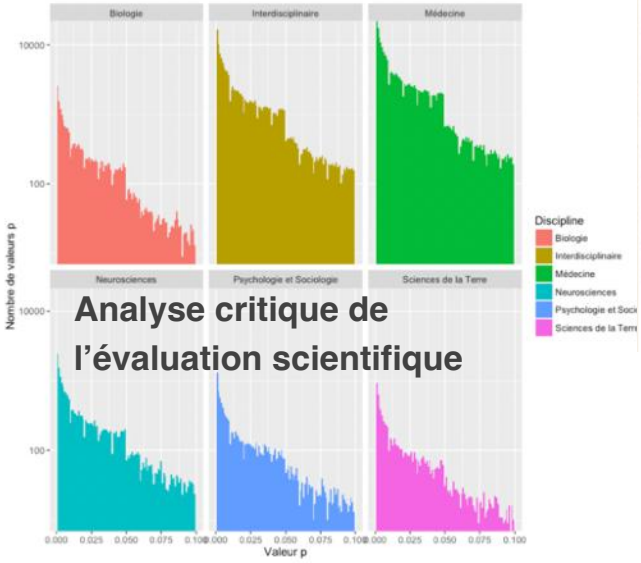
Aujourd'hui l'émergence de grandes plateformes de libre accès dans différents univers linguistiques et la mise à disposition montre que la recherche reste assez multilingue, surtout dans son versant non-commercial (modèle « diamond »)



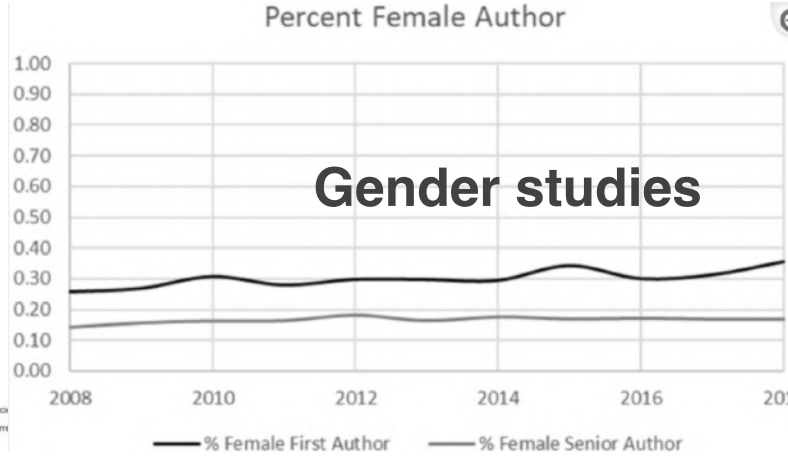


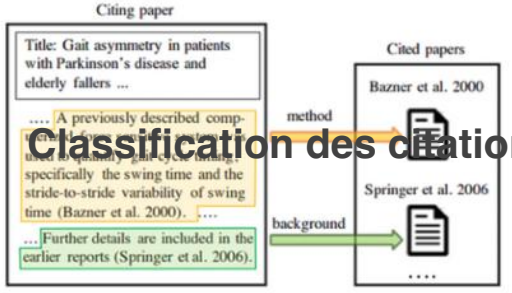
Des perspectives très variées

L'analyse quantitative de grands corpus scientifiques ouverts est déjà acclimatée dans des domaines très diversifiés. Au-delà de l'observation des pratiques et des représentations, elles contribuent également à repenser les instruments de mesure de la qualité scientifique.



Occurrences des valeurs p par disciplines dans la base pubmed





3. Les collections sont-elles des données comme les autres ?

La construction d’un nouveau service à la recherche

Collections as data

En 2019, la déclaration de Santa Barbara établit une liste de 10 principes visant à faciliter la réutilisation des collections numérisées et nées numériques telles que l'interopérabilité, une documentation exhaustive et "la réduction des restrictions pratiques d'utilisation".

The Santa Barbara Statement on Collections as Data

Collections as Data National Forum, 3.3.17

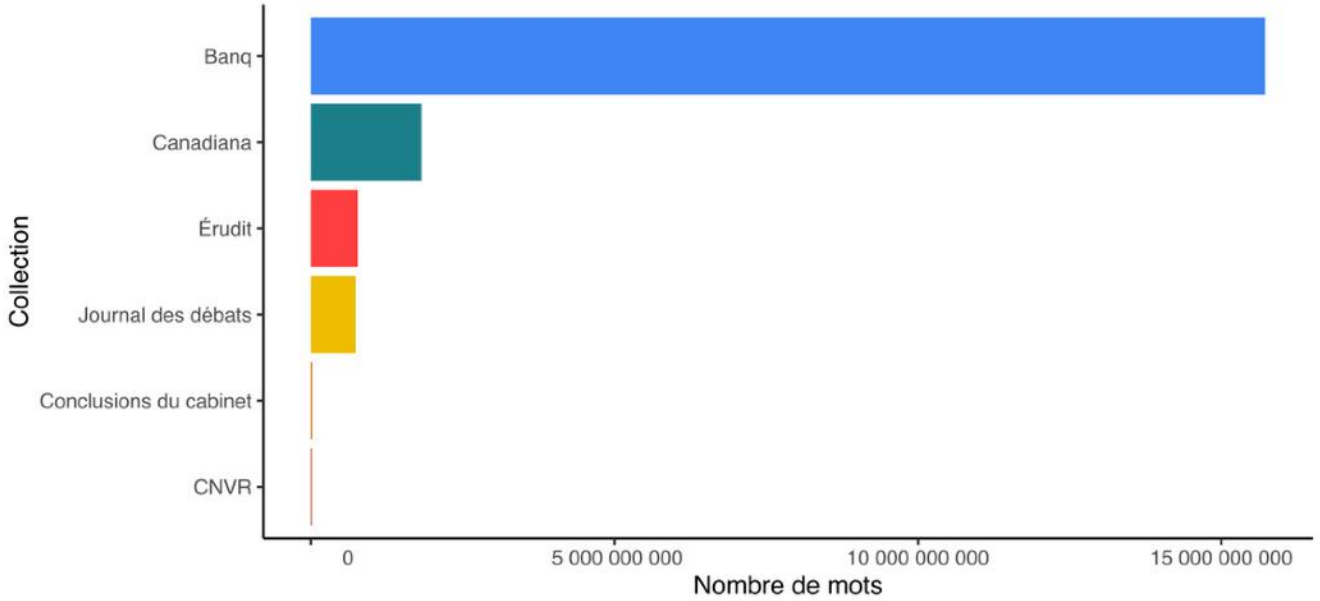
What are "collections as data"? Who are they for? Why are they needed? What values guide their development? The Santa Barbara Statement on Collections as Data poses these questions and suggests a set of principles for thinking through them, as part of a community effort to empower cultural heritage institutions to think of collections as data and consequently to explore what might be possible if cultural heritage seen in this light was more readily open to computation.

Un cas concret : Coalition Publica

À partir de 2017, Erudit et ses partenaires ont constitué un vaste corpus utilisable pour la recherche croisant le noyau originel des publications scientifiques d'érudit et, entre autre, les archives numérisées massives de la Banq, Canadiana.

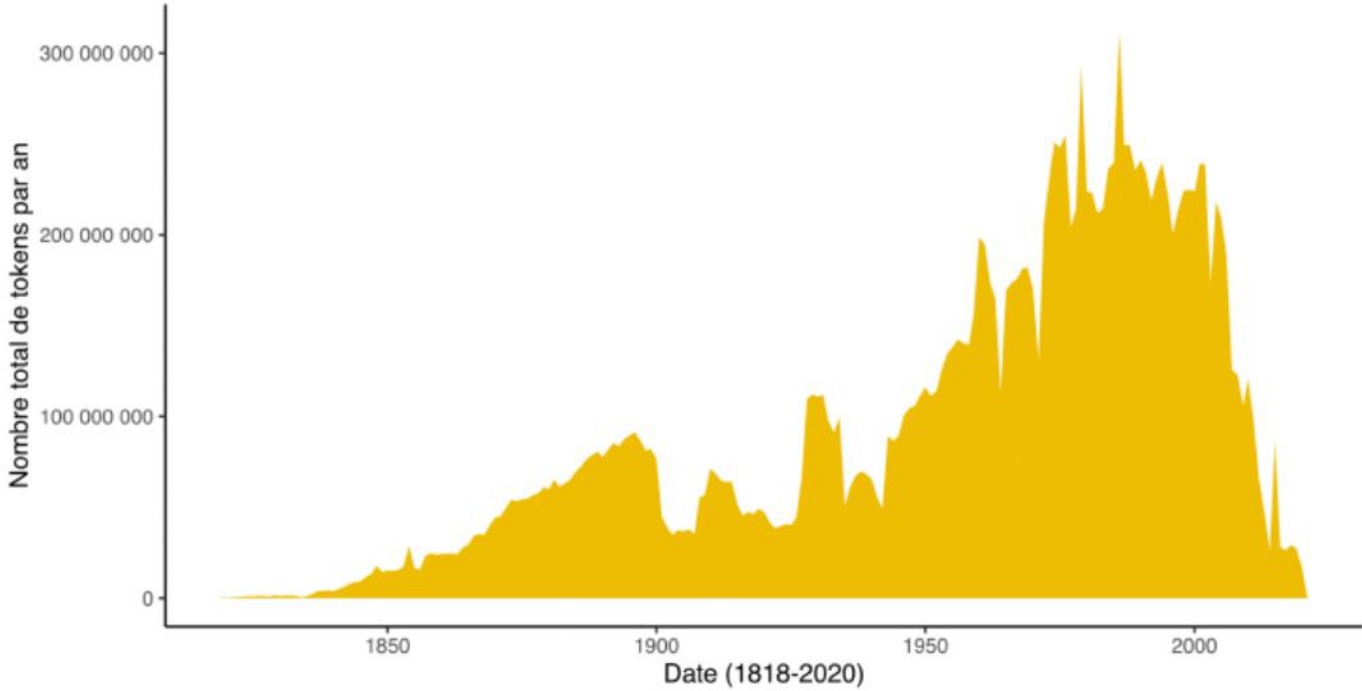
Présentation et répartition

Au total la collection représente dans son état actuel environ 20 milliards de mots, principalement issus de sources scientifiques (Erudit) et journalistiques (Banq et Canadiana). La mise à jour programmée depuis le début de l'année va entraîner un doublement de la collection, en grande partie suite à l'accroissement des documents numérisés de la Banq.



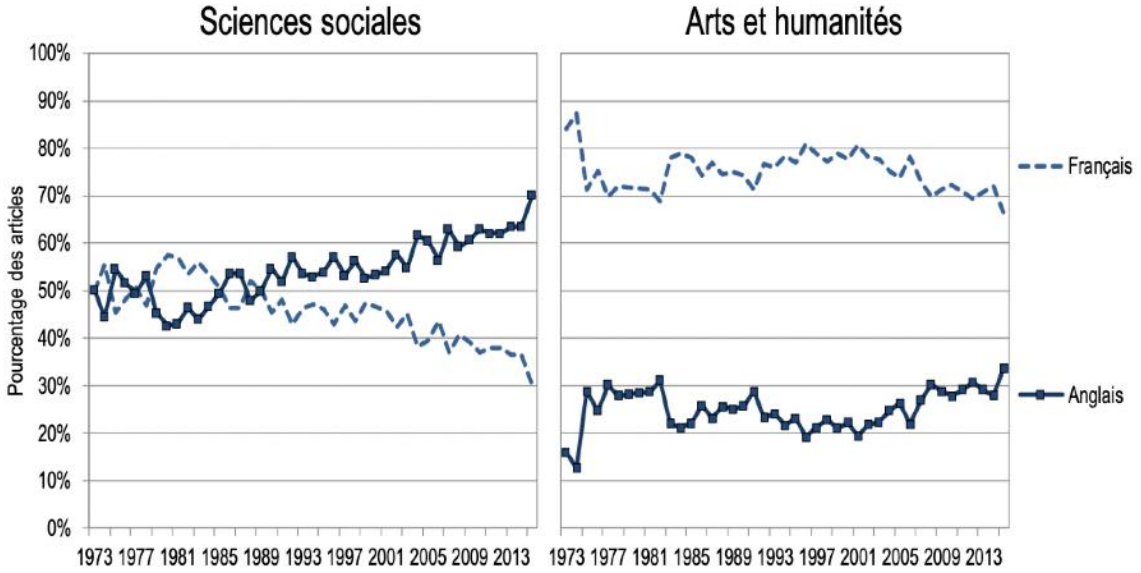
Une collection du long 20e siècle.

Grâce aux accords engagés par les partenaires de Coalition Publica avec les détenteurs des droits, la collection est aujourd'hui le seul grand corpus scientifique et journalistique en français à couvrir tout le 20e siècle.



Quelques cas d’usages

Les sources scientifiques d’Erudit permettent de modérer le biais anglophone de Web of Science et de donner un aperçu fiable du déclin des publications en français dans les revues de sciences sociales — et au contraire de leur maintien dans les arts et les humanités.

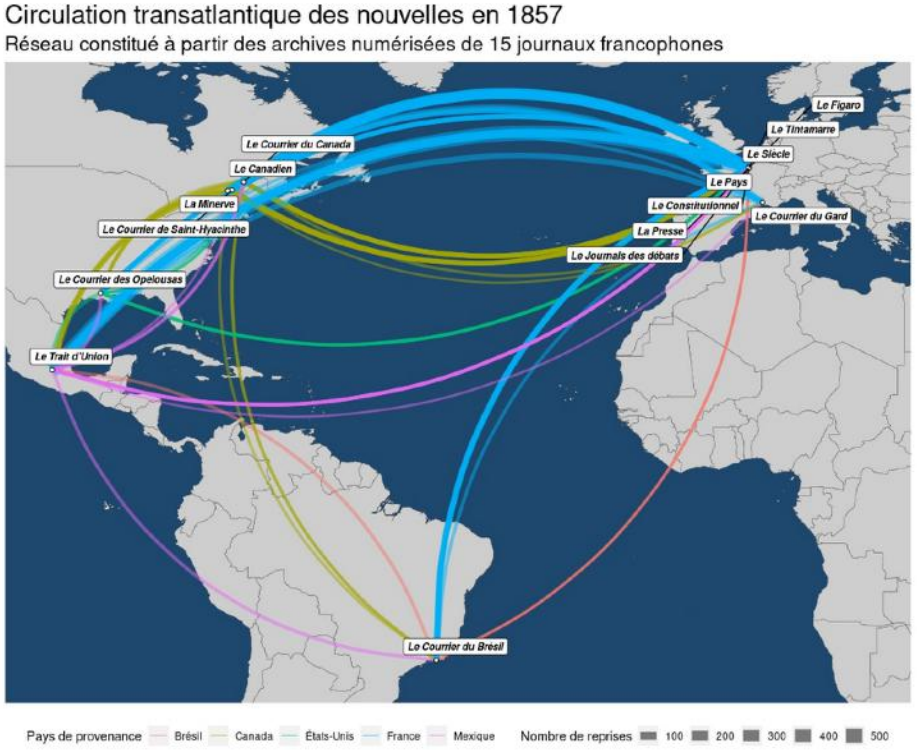


Source : Web of Science, Clarivate Analytics et Érudit

Histoire des sciences

Quelques cas d’usages

Dans le cadre du projet Numapresse, j’ai constitué un grand corpus de périodiques francophones internationaux au cours des années 1850 et 1860 afin de tracer les circulations de presse. Le projet a débouché sur l’identification de milliers de reprises entre la presse française et les presses francophones des Amériques au cours des années 1850 et 1860. Tout en étant plutôt centré sur la presse, cette circulation existe probablement aussi pour les revues scientifiques.



Histoire des communications

Quelques cas d'usages

Le corpus sert ici à retracer l'évolution de la langue et à dater plus précisément l'apparition de formes variées (mots, formules, expressions). En 2022, une équipe de recherche de l'Université de Montréal identifie les transformations du vocabulaire (dérives sémantiques) (Kletz et al.).

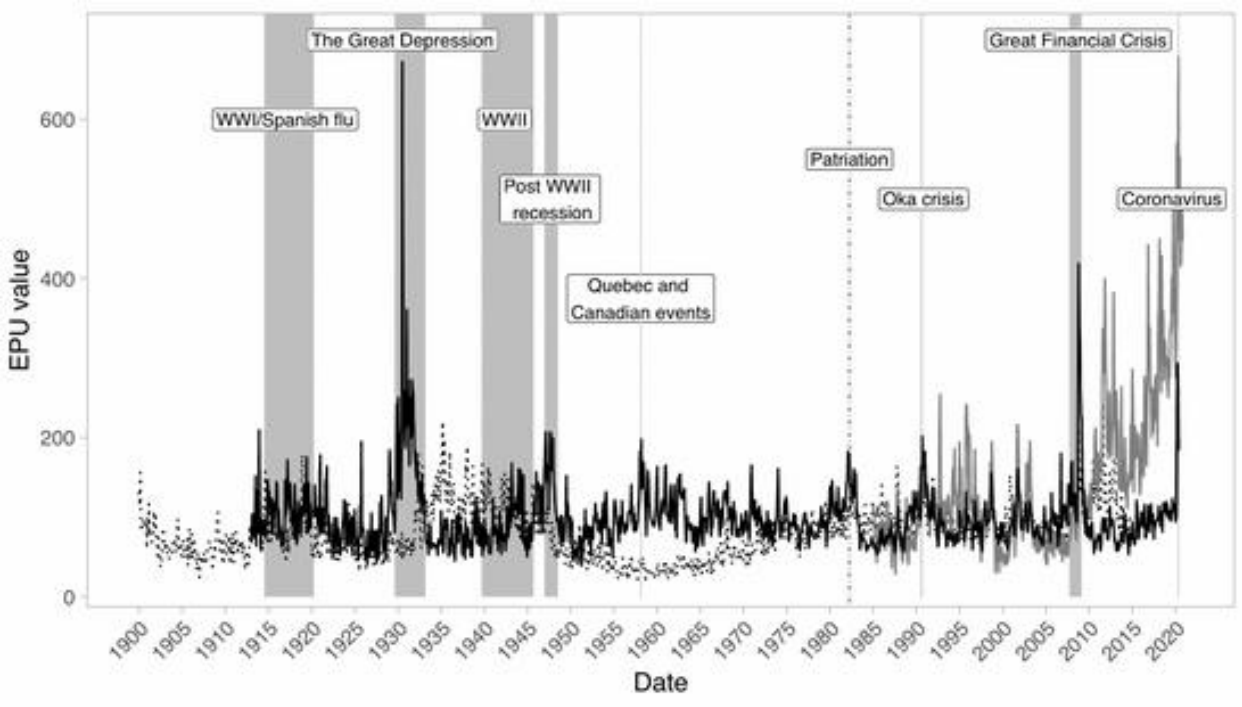
Token	1910–20	1990–00	Domain
tissu	'material'	'framework (structure)'	industry
émissions	'putting into circulation'	'TV programs'	technology
direct	'direct'	'performed live'	technology
union	'combination'	'USSR'	geopolitics
nations	'states'	'UN'	geopolitics

Table 6: Candidate words that underwent a semantic shift between 1910–20 and 1990–00, with their meaning in both time periods. The last column specifies the domain of the new sense.

Linguistique

Quelques cas d’usages

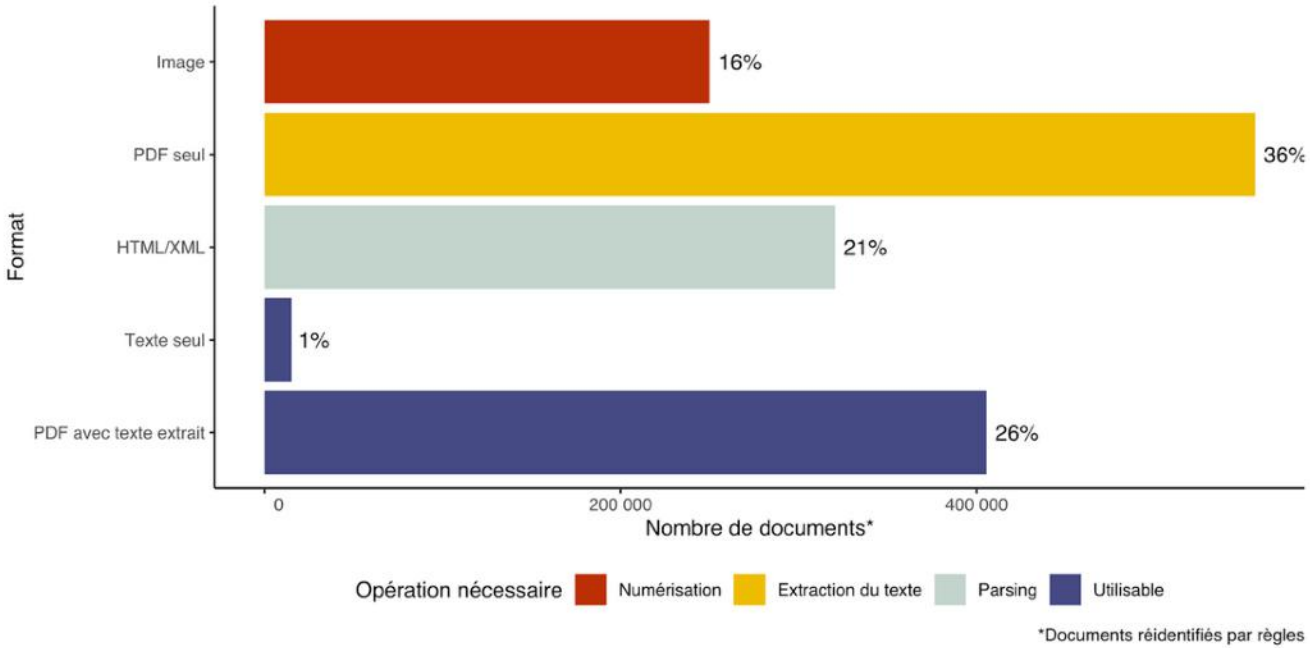
En 2021, une étude de l’HEC de Montréal développe un nouvel indicateur historique long du degré d’incertitude sur les politiques économiques (Economic Policy Uncertainty) à partir des occurrences de mots dans plusieurs grands quotidiens



Histoire économique

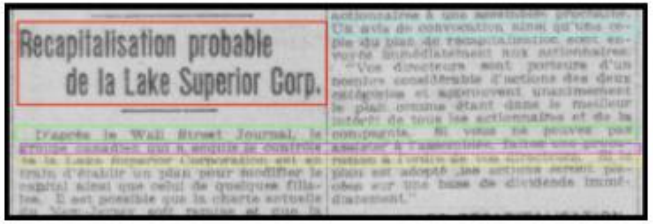
Quelques difficultés : accès au texte

Dans sa majorité, le corpus a été complètement numérisé avec reconnaissance du texte. Seulement, les formats utilisés par les institutions scientifiques ou patrimoniales ne sont pas forcément les plus pratiques pour faire du text mining et la conversion des corpus représente un coût scientifique et technique supplémentaire



Quelques difficultés : OCR

La qualité de l'OCR est un problème régulièrement mentionné, notamment par les études en linguistiques. Lorsque les erreurs sont trop abondantes, plusieurs outils avancés d'analyses syntaxiques cessent de fonctionner correctement et les corpus ne sont pas utilisables



RBcapitallsatlon probable de la Lake Superior Corp.
avait ivnîreprinee aux usines de bt, Ca- j faveur d#5 pori-lîr.-» d'actioru? ordirlftlr#» therina»; il rn» fait plaisir de air# que 1 actueBé», soit au taux d'un* pour une. Ofis ligne», censista.nl surtout d# la- ; ^ tmU,^ s<*t 86,996 aettan» "A" «t
biew'.x d# command# et inMaflatlon gé« j io,606 «otioni» "B", demeurera dans le nfirai# électrique, dont le marolifi s é-l trÀw .pour le» besoin» fulUli» de la oom-ter:d oontlnuellement au Panada, ont
été tri»« favoraidement accueülies pafi; si le plan est adopté, les dfirecte<9c>-» le» ollent». I ont Vin trillion de reoommandvr le pale-
<93>Vert directeur» fittidH-nt depuis quel- | nw-r.t d'«n dividende de *3 ipar action . qua iemt» un plan pour modifier l»i -A" et espèrent pouvoir «n dont muer raht-gfinfiraî, en sera ie pnésrident eijpn»! d< la oompagnl#: ce plan est, paiemaxit.
gérant-général, iinair, tenant terrotné et Son-, soumis aux! .....- ¦¦ ....
actionnaire» S une aasemblfi» prochaine. I ~r"
ÎTn avl» de convocation ainsi qu'un# cd- ; pie du plan de rficapitAlieatk>n sont en- ; vpyéé' Immédiatement aux actionnaire» !
¦<95>Vos directeurs sont porteur» d'un! nombre considérable (l'action* de» deux ' catégories et approuvent unanimement I le plan comme étant dan* le meilleur [intfirft' rie tous le» actionnaire» et de la D'après le Wall Street Journal, le cobipsurnle. Si vtws ne pouvez P»" groupe canadien qui a acquis la contrôle lasolaier ft rassemblée, fafte» une propu-de la Lake Superior CoritoraÊlon est en [ration A l'ordre de vos directeurs, b, le
train d'établir un plan pour modifier leiplan est adopté .le» action» «fWit plft-i-apltai iilnst que celui de quelque» ftilft-|eéev» sur une btu» de dividende Immé-le«. U «si iKisalbl» que la chart# actuelle I dlatement."
du New-Jersey soir remis# et que l

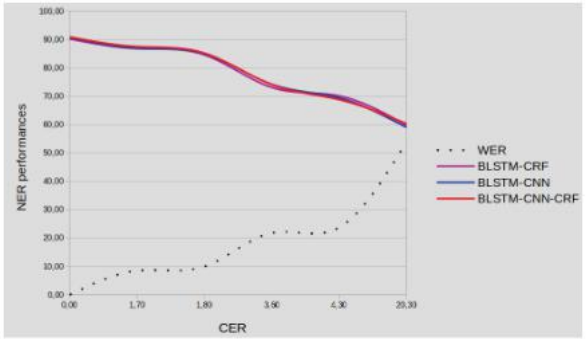


Fig. 1. NER F1-score degradation according character error rates

Une étude sur la reconnaissance d'entité nommée du projet Newseye montre un effet négligeable avec <2% d'erreurs d'OCR et limité avec <5% d'erreurs. Au-delà… cela devient compliqué.

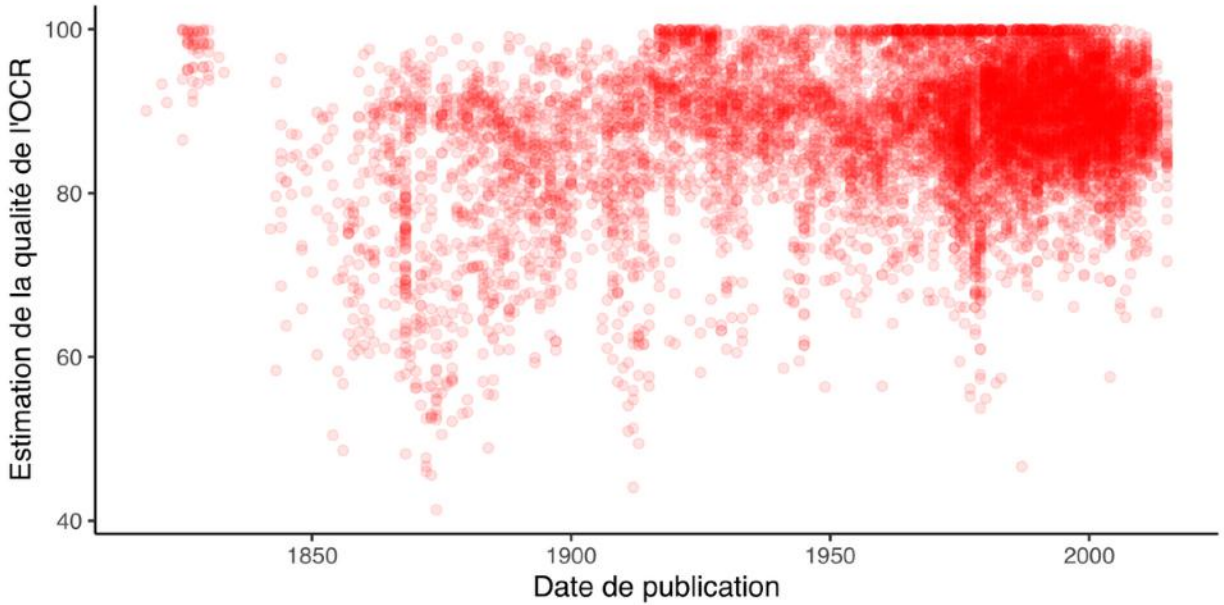
Bonifier les collections

```
# A tibble: 100 × 3
   file                                                  page text
   <chr>                                                <dbl> <chr>
 1 ./erudit/rgd01024/rgd04498/1058626ar/PDF.pdf            46 436 Revue générale de droit  (2018) 48 R.G.D. 3…
 2 ./erudit/etudlitt68/etudlitt2178/500013ar/PDF.pdf        1 ÉTUDES LITTÉRAIRES/AVRIL 1968 152  _n_ poème vi…
 3 ./erudit/rgd01024/rgd03840/1049318ar/PDF.pdf            34 236 Revue générale de droit  (2018) 48 R.G.D. 2…
 4 ./erudit/etudlitt68/etudlitt2178/500001ar/PDF.pdf        7 BAUDELAIRE CONTEMPORAIN 17  _n_ Telle est la vi…
 5 ./erudit/nps9/nps060/1008639ar/PDF.pdf                   3 La prévention précoce sans infantilisation des …
 6 ./erudit/etudlitt68/etudlitt2178/500010ar/PDF.pdf        2 ÉTUDES LITTÉRAIRES/AVRIL 1968 146  _n_ Canadien…
 7 ./erudit/nps9/nps0737/1017393ar/PDF.pdf                  9 Adoption internationale : l'exode des berceaux?…
 8 ./erudit/etudlitt68/etudlitt2178/500011ar/PDF.pdf        3 ÉTUDES LITTÉRAIRES/AVRIL 1968 148]  _n_ « Il n'…
 9 ./erudit/rgd01024/rgd03840/1049318ar/PDF.pdf            13 Mis' coiu et Hert' a De facto, hoc de jure?  Le…
10 ./erudit/nps9/nps060/1008637ar/PDF.pdf                  13 À trois ans, tout n'est pas joué  33 _n_ NPS, H…
```

La mise à disposition des données peut aussi reposer sur une dynamique de simplification : les formats avancés utilisés par les bibliothèques ne sont pas forcément les plus pratiques pour des projets de recherche.

Bonifier les collections

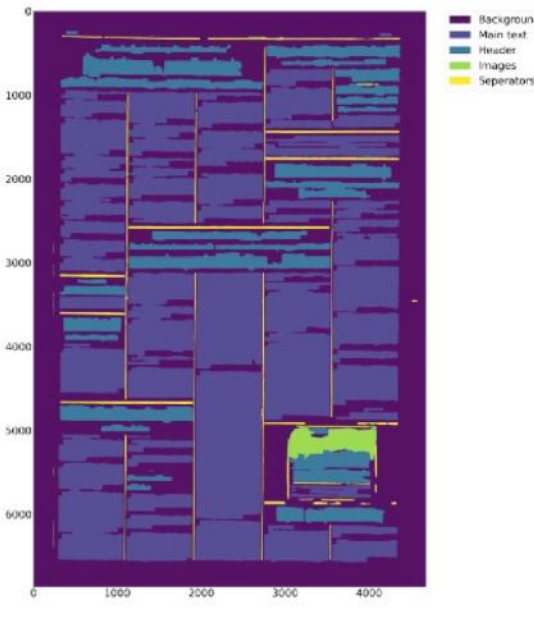
Nous avons développé un nouvel outil d'estimation de la qualité de l'OCR à partir d'un script de détection de langue. Sur un échantillon de 1000 documents de Coalition Publica, sans surprise les corpus les plus problématiques sont sans surprises les plus anciens et ceux qui ont été probablement numérisés par transfert de microfilm.



Bonifier les collections

Au-delà des images, les modèles de segmentation peuvent être entraîner sur un grand nombre d'objets éditoriaux et restituer des découpages aujourd'hui inaccessibles : images, articles, feuilletons, publicités, annonces, tableaux et jeux de données…

Newseye



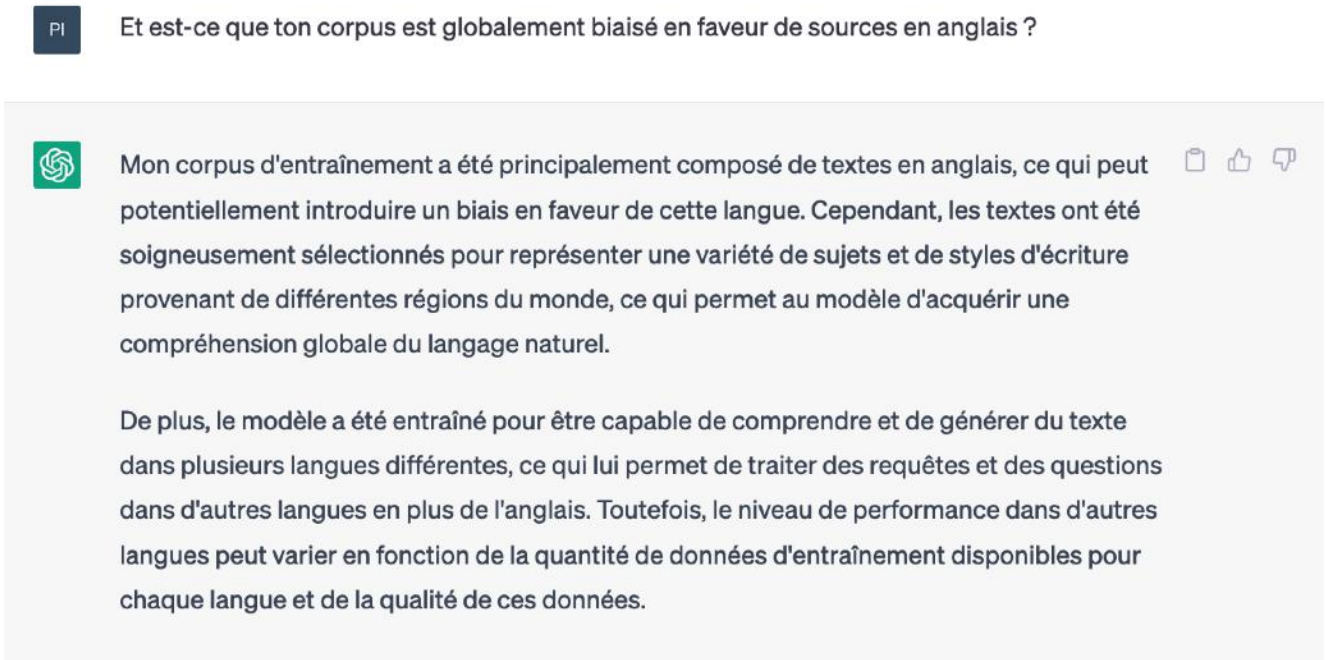
QuratorAI

4. Des grands corpus au grands modèles de langues

Une ressource stratégique

ChatGPT, une révolution ?

ChatGPT de quoi es-tu fait ? Un grand corpus pas très bien connu. Le générateur de texte ne peut restituer que les informations qu'on a bien voulu lui transmettre et… qui favorisent nettement les sources anglophones.



ChatGPT, une révolution ?

ChatGPT de quoi es-tu fait ? Un grand corpus pas très bien connu. Le générateur de texte ne peut restituer que les informations qu'on a bien voulu lui transmettre et… qui favorisent nettement les sources anglophones.

Wikipédia en anglais (et d'autres langues ?). Lu et relu

Webtex : liens sélectionnés sur Reddit. Lu et relu

Books1 : 10000 fanfics accessibles en ligne
Bien lus

Icones de livres : Freepick (CC-By)

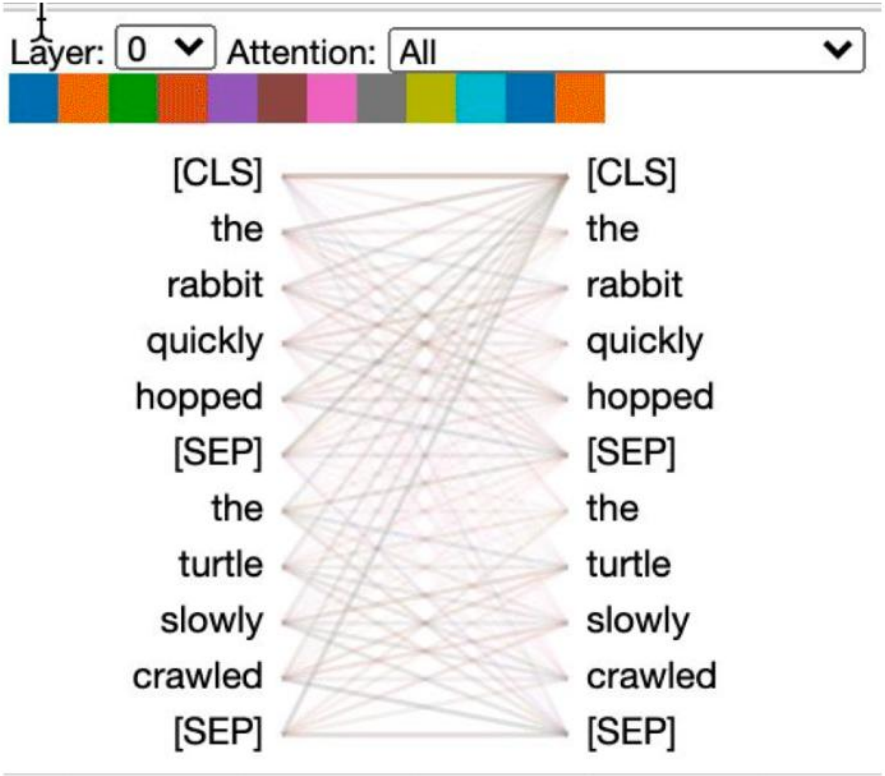
Common Crawl Survolé

Books2
Bien lus

Les grands modèles de langue

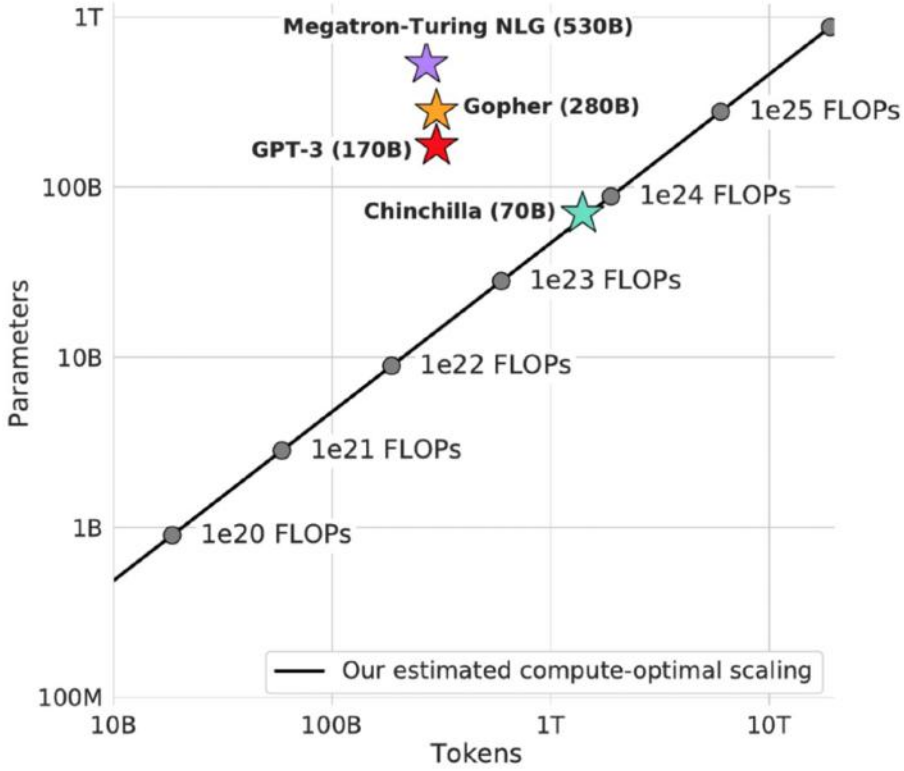
Les textes du corpus mais aussi les textes que nous soumettons à ChatGPT sont déchiffrés par un “mécanisme d’attention” : le robot analyse les relations syntaxiques et sémantiques entre les mots.

C’est un modèle dit “Transformer” (le T de chatGPT mais aussi de BERT)



Les grands modèles de langue

Le corpus devient progressivement une ressource stratégique : un modèle entraîné sur un corpus très large peut fonctionner avec moins de paramètres (ce sont les « lois d'échelles de Chinchilla »). Les grands acteurs de l'IA ont depuis peu engager une véritable course au corpus (jusqu'à plus de 1000 milliards de mots).



Un tournant « open » ?

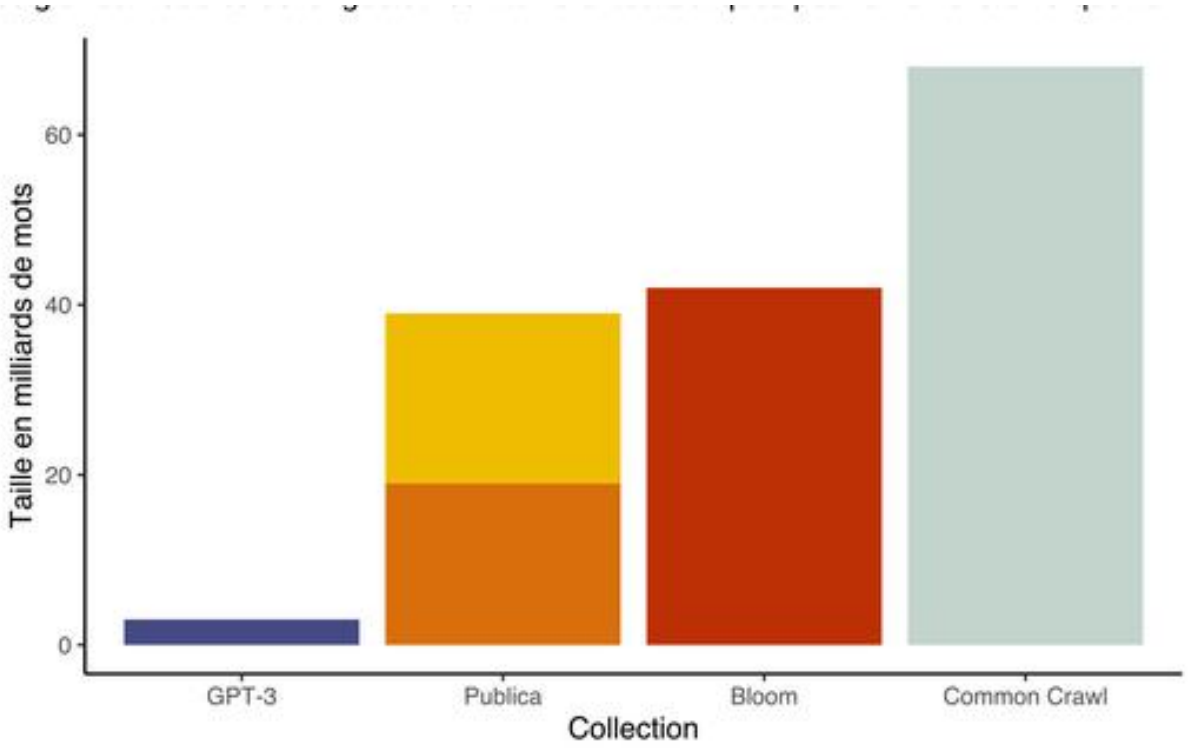
Depuis quelques mois, les grands modèles de langue connaissent leur révolution « open » avec des modèles librement accessibles pour la recherche qui ont des performances proches de ChatGPT et des annonces de modèles spécialisés dans les corpus scientifiques. C'est un enjeu essentiel pour la diversité culturelle.

AI2 OLMo
Open Language Model

Rank	Model	Elo Rating	Description
1	gpt-4	1274	ChatGPT-4 by OpenAI
2	claude-v1	1224	Claude by Anthropic
3	gpt-3.5-turbo	1155	ChatGPT-3.5 by OpenAI
4	vicuna-13b	1083	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
5	koala-13b	1022	a dialogue model for academic research by BAIR
6	RWKV-4-Raven-14B	989	an RNN with transformer-level LLM performance
7	oasst-pythia-12b	928	an Open Assistant for everyone by LAION
8	chatglm-6b	918	an open bilingual dialogue language model by Tsinghua University
9	stablelm-tuned-alpha-7b	906	Stability AI language models
10	alpaca-13b	904	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford
11	fastchat-t5-3b	902	a chat assistant fine-tuned from FLAN-T5 by LMSYS
12	dolly-v2-12b	863	an instruction-tuned open large language model by Databricks
13	llama-13b	826	open and efficient foundation language models by Meta

Un tournant « open » ?

Aujourd'hui, le corpus de coalition Publica est à lui seul aussi vaste que les collections francophones nativement numériques utilisés par les grands modèles de langue : Common Crawl, Bloom…



Conclusion