Comprendre les archives : explorer et valoriser les documents historiques grâce à l'annotation sémantique

Nicolas Gutehrlé
C.R.I.T., Université de Franche-Comté
Printemps de la Données, 23/05/2024

Sommaire

Introduction

Contextualisation

Les campagnes de numérisation menées par les archives et bibliothèques ont permis :

- de préserver ces documents de manière plus sécurisée
- d'en faciliter l'accès au grand public
- l'émergence de nouveaux domaines d'études, tel que les humanités numériques ou l'histoire numérique

Projet EMONTAL

Numéro 70 | Le Numéro : 0.25 | Samedi 11 Avril 1936

LE FRONT OUVRIER

Organe hebdomadaire du prolétariat montcellien

Administrateur-délégué : VITTEAUT J.-B., 9, rue du Plessis - C. C. Postal Dijon 200 50

QUI HABILLE LE MIEUX ?

C'est

Paul MILLET

Les marchands de mort

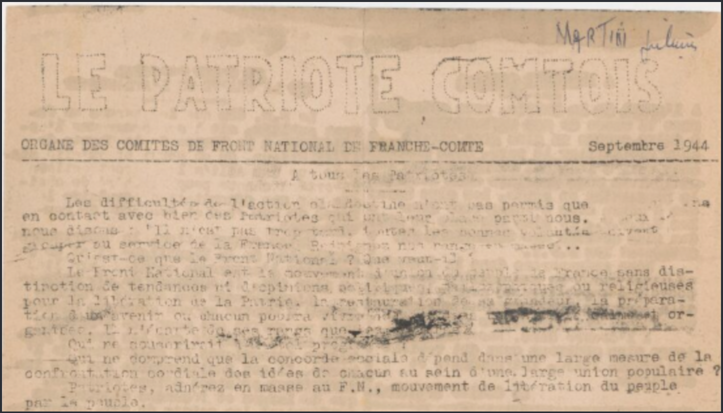
La Réunion de Mont-Saint-Vincent

Martin, candidat désigné par lui-même, et renié par ceux qui ont voté pour lui, nous donne enfin une réunion !

Billet de la semaine

A des figurants

Projet EMONTAL



LE PATRIOTE COMTOIS

ORGANE DES COMITES DE FRONT NATIONAL DE FRANCHE-COMTE Septembre 1944

A tous les Patriotes

Les difficultés de l'action clandestine n'ont pas permis que [illegible] en contact avec bien des Patriotes qui ont leur place parmi nous. [illegible] nous disons : "Il n'est pas trop tard. Toutes les bonnes volontés doivent [illegible] au service de la France. Rejoignez nos rangs [illegible]..

Qu'est-ce que le Front National ? Que veut-il ?

Le Front National est le mouvement d'union [illegible] la France sans distinction de tendances ni d'opinions [illegible] ou religieuses pour la libération de la Patrie, la restauration de sa grandeur, la préparation d'un avenir où chacun pourra vivre [illegible] or-ganisées. Il n'écarte de ses rangs que [illegible]

Qui ne souscrirait à [illegible] programme ?

Qui ne comprend que la concorde sociale dépend dans une large mesure de la confrontation cordiale des idées de chacun au sein d'une large union populaire ?

Patriotes, adhérez en masse au F.N., mouvement de libération du peuple par le peuple.

Contextualisation

Cependant, la découvrabilité, l'exploitation et la valorisation de ces documents restent des tâches difficiles :

- L'accès à ces collection se fait habituellement via des moteurs de recherche et des requêtes à base de mots-clés : le nombre de documents retournés pour une requête est souvent conséquent
- La lecture proche des documents n'est pas adaptée à cette surabondance
- Les méthodes de lecture distante (Moretti, 2013) ne permettent pas d'analyses fines

Contextualisation

Ainsi, il est nécessaire de structurer le contenu textuel de ces "Big Data of the Past" (Kaplan et di Lenardo, 2017) afin :

- d'améliorer les interfaces de recherches
- concevoir de nouvelles interfaces qui synthétiseraient les résultats obtenus et assisteraient dans leur compréhension
- faciliter l'exploration, l'exploitation et la valorisation de ces documents

Cela passe par l'annotation sémantique du contenu textuel des documents

Annotations sémantiques des documents et nouvelles interfaces de recherche

Annotation sémantique des documents d'archives

Structurer le contenu textuel des documents est réalisable par l'ajout d'annotations sémantiques

- Ces annotations s'obtiennent par l'emploi de méthodes de Traitement Automatique des Langues

Annotation sémantique des documents d'archives

Reconnaissance des entités nommées (REN) et extraction de relations (ER)

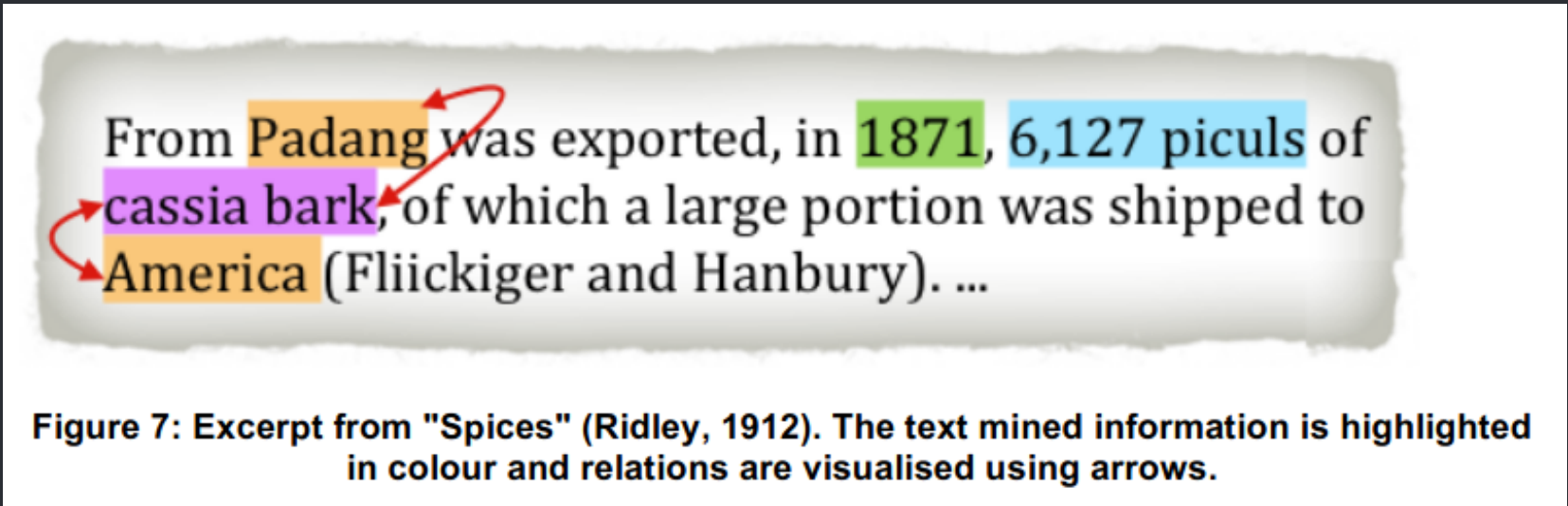


Figure 7: Excerpt from "Spices" (Ridley, 1912). The text mined information is highlighted in colour and relations are visualised using arrows.

Figure – Exemple de reconnaissance d'entités nommées et d'extraction de relations dans le projet Trading Consequences (Hinrichs et al., 2015)

Annotation sémantique des documents d'archives

Modélisation de sujet (*Topic Modelling*, TM)

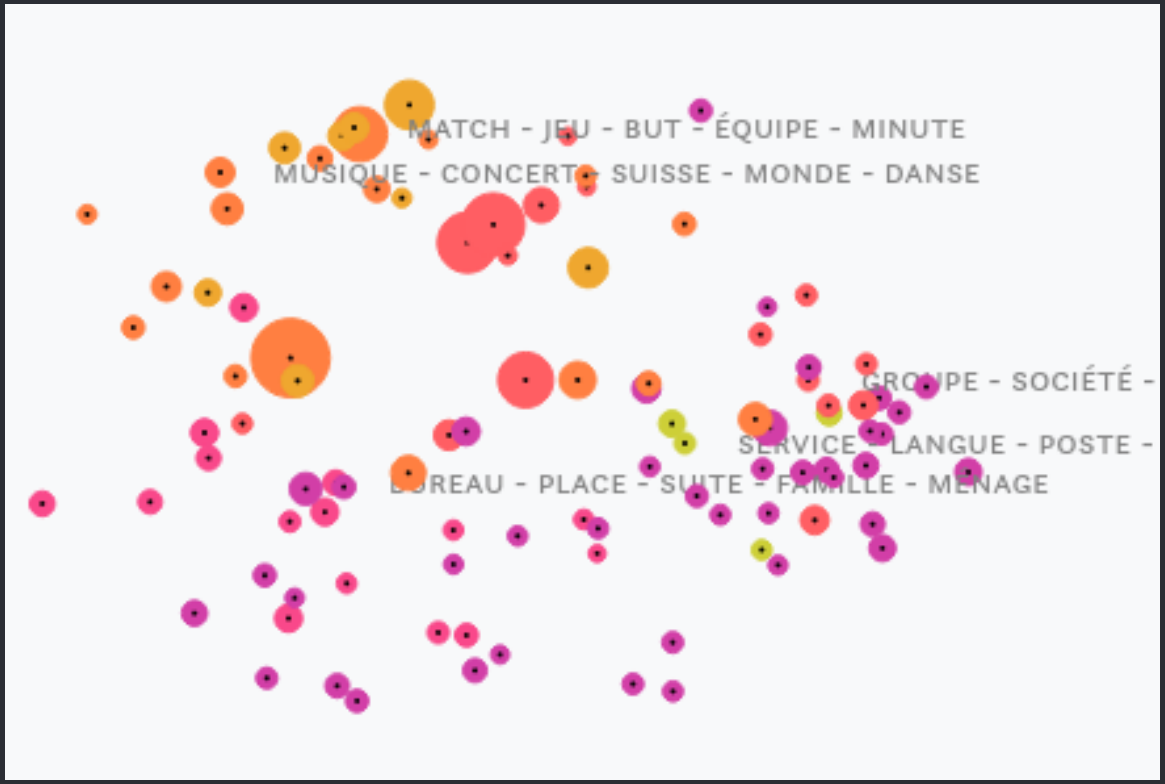


Figure – Exemple de thèmes identifiés dans le corpus de l'interface impresso

Annotation sémantique des documents d'archives

Analyse de la structure logique (*Logical Layout Analysis*, LLA) des documents

PREMIERE ANNEE. — N° 2. SAMEDI 23 AVRIL 1932.

RÉDACTION — ADMINISTRATION
1, rue des Prés
BELFORT
Directeur politique : Henri JACOB
Rédacteur en Chef : Paul RASSINIER
PUBLICITÉ
Petites Annonces, la ligne 1.25
Annonces Judiciaires — 1.75
Rubrique Locale . . . — 4 »
Pour les annonces périodiques ou de longues durées, traiter avec l'administrateur.

Le Semeur

JOURNAL COMMUNISTE

Territoire de Belfort, Doubs, Jura, Haute-Saône

ABONNEMENTS
6 mois..... 8.50
1 an..... 16 fr.
Envoyer les fonds au camarade R. Jacob, 1, rue des Prés, Belfort. Chèque Postal Dijon 292-93
Le Numéro : 30 CENTIMES
Paraissant le samedi

Contre la guerre et la misère Tous debout le Premier Mai!

Dans quelques jours, nous serons au 1er Mai. Malgré que le gouvernement de misère de Tardieu a fixé ce jour pour le premier tour, des élections, nous devons mobiliser les masses autour des mots d'ordre de la C. G. T. U.

Nous ne voulons pas faire du 1er Mai le « Fête du Travail » comme le veulent les réformistes. Au contraire, cette journée doit être considérée par tous les exploités comme une journée de lutte contre :

La guerre impérialiste ;
La répression ;
Le chômage et la misère ;
Les diminutions de salaire.

La classe ouvrière ne peut pas être en fête dans une telle situation.

tables assurances sociales, payées par la bourgeoisie, et dont les indemnités pour la maladie, devront correspondre aux salaires moyens et dans lesquelles devra être introduite l'assurance-chômage.

Avec nous, les chômeurs complets et partiels seront dans la rue pour arracher aux responsables de la crise économique actuelle, l'indemnité minimum de 12 francs par journée perdue.

Et au-dessus de tout en ce jour de 1er Mai 1932, la classe ouvrière signifiera à l'impérialisme français sa volonté de lutte contre la guerre.

C'est surtout dans notre pays que nous devons engager une bataille sans merci contre notre propre impéria-

Pour la vie et le développement du "Semeur"

Le premier numéro du *Semeur* a été accueilli avec joie par tous les travailleurs de notre région. Le titre même évoque la lutte qu'un journal du même nom, qui fut l'organe du Parti dont nous étions, mena contre la bourgeoisie régionale et ses soutiens, de 1922 à 1923.

Cette lutte, nous entendons la mener selon la tradition créée par le premier *Semeur*, même en dehors du Parti, nous restons ce que nous étions : des adversaires de tous les partis bourgeois, socialistes compris, des révolutionnaires fidèles à Lénine, à la révolution russe et à la troisième Internationale. Nous serons à la pointe du combat qui conduira la classe ouvrière mondiale à la dictature du prolétariat à l'échelle universelle et à son émancipation définitive.

Mais il faut que nous trouvions chez nos

Cette Semaine...

N n'oeuvrera de commerce encore cet article aux dimanche. Dans nos derniers numéro, nous avons attaché une certaine importance à ce genre de sport. Il faut y revenir : c'est la saison qui veut ça !

C'est qu'en effet, nous sommes en pleine période électorale, on essaie de semer dans les rangs du Prolétariat pas mal d'illusions que précisément ces discours viennent détruire.

Le Français moyen considère qu'en politique il y a trois grands maîtres de l'opinion : Tardieu, Herriot, Blum, qu'il faut prendre, Blum vaut mieux qu'Herriot, mais aussi, qu'à son tour Herriot vaut beaucoup mieux que Tardieu.

Le moindre mal.

C'est la théorie qu'encore samedi soir, le chef socialiste local, Narpon développait devant 200 personnes, à la Brasserie Georges.

Les trois grands maîtres de l'opinion publique, les trois énergies ont parlé en page — raisons dans le style du Français moyen — dimanche dernier.

Le premier parlait à Grammont, le second

Coup d'œil sur la situation en Allemagne

L'élection de Hindenburg à la présidence du Reich est utilisée par la presse de « gauche » pour rassurer l'opinion de tous les pays sur l'éventualité d'une dictature fasciste en Allemagne. Les socialistes surtout, ont crié victoire et se sont vantés d'avoir infligé à Hitler une défaite qu'ils disent définitive.

Un bref examen de la situation suffit, pour se rendre compte qu'il s'agit là d'une nouvelle hypocrisie des chefs de social-démocratie de France et d'ailleurs, ils cherchent une fois de plus à tromper les masses, mais ils savent très bien qu'en portant le candidat Hindenburg à la Présidence, ils n'ont fait que trahir à nouveau la cause du prolétariat sans aucune solution au problème allemand.

Les causes profondes de la crise politaut aux candidats révolutionnaires, (communistes et hitlériens sans distinctions) prouvent une concordance exacte, entre la progression du chômage et la progression des voix révolutionnaires. Pendant la dernière période, celles-ci croissent même plus rapidement encore. Le 14 septembre 1930, elles s'montaient à 21,6 cent des suffrages exprimés ; le 13 mars 1932, à 43 pour cent ; le 10 avril dernier, elles atteignaient 49 pour cent.

On a d'autre part la certitude que le chiffre des exportations allemandes va baisser encore considérablement : un important client du Reich, l'U. R. S. S. étant ruiné l'achèvement de son industrialisation dans le cadre du plan quinquennal : ses importations dans la proportion de 60 à 70 pour cent.

Figure – Extrait de la 1ère page du second numéro du journal communiste *Le Semeur*, publié le 23 avril 1932

Annotation sémantique des documents d'archives

Ces annotation seraient exploitables par les moteurs de recherche pour indexer les documents, et permettraient :

- des requêtes plus fines
- de limiter le nombre de documents retournés
- de faciliter la découvrabiblité des documents

Les plateformes impresso et NewsEye sont des exemples de ces moteurs de recherche augmentés

Annotation sémantique des documents d'archives

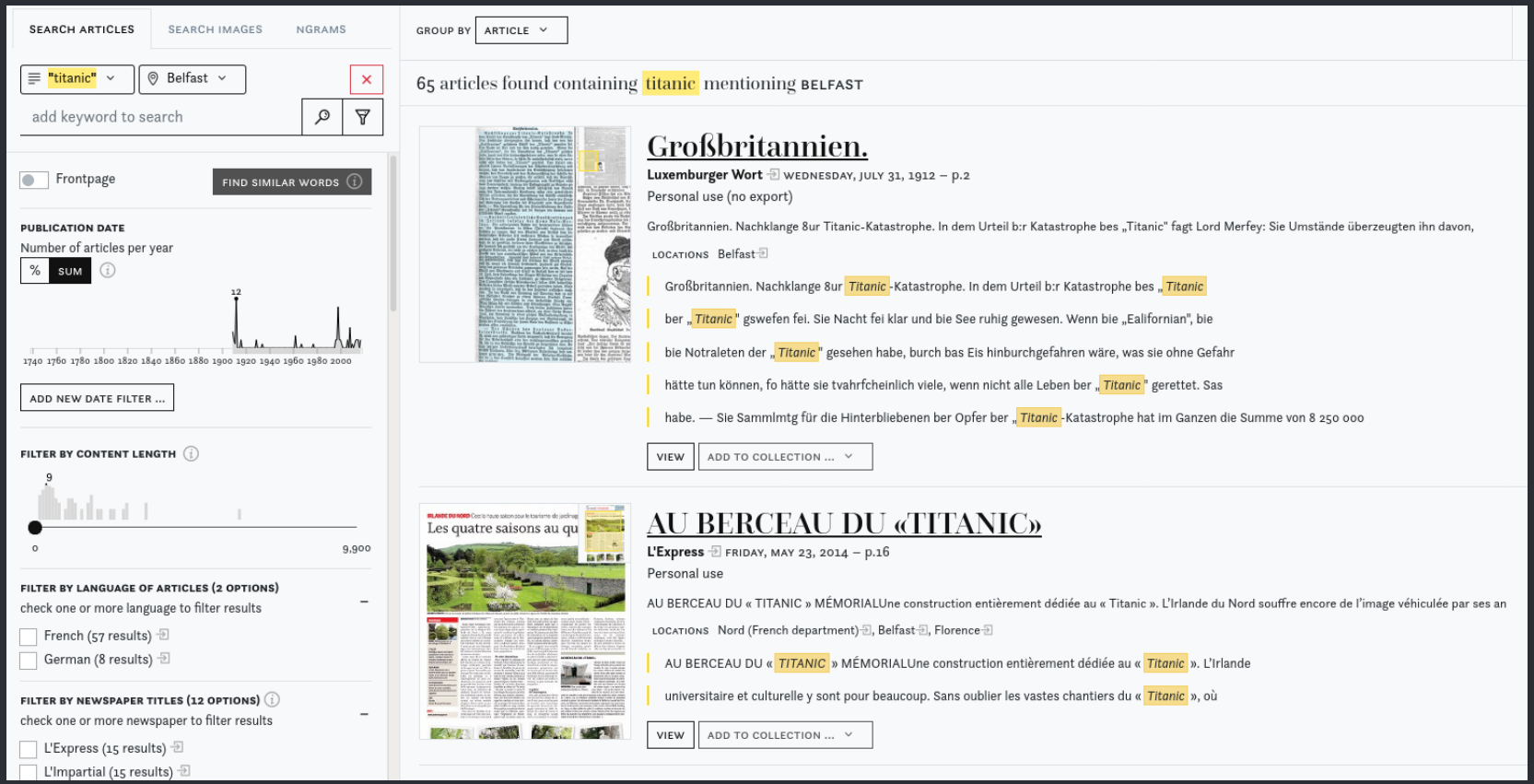


Figure – Résultats pour la requête "Titanic ;Belfast" dans impresso

Annotation sémantique des documents d'archives

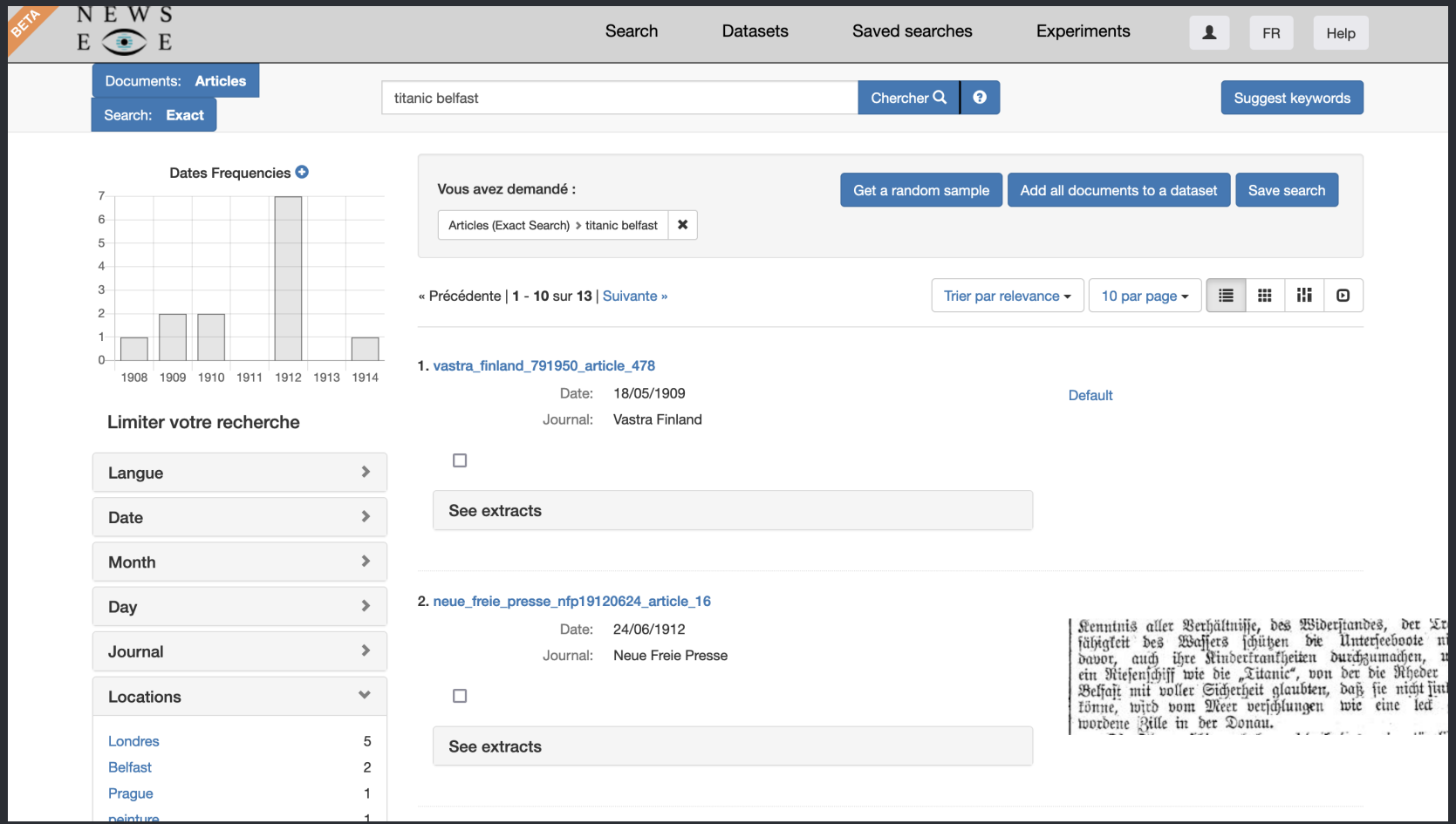


Figure – Résultats pour la requête "Titanic ;Belfast" dans NewsEye

Vers de nouvelles interfaces de recherche

L'ajout d'annotations sémantiques et les outils de TAL permettent d'imaginer de nouvelles interfaces qui :

- compléteraient les interfaces actuelles
- synthétiseraient les documents retournés par les requêtes et assisteraient dans leur compréhension (Jatowt, 2021)

Vers de nouvelles interfaces de recherche

Des graphiques permettent de voir l'emploi des termes dans le temps et de les étudier en diachronie, à la manière de Google NGrams Viewer (Mann et al., 2014)

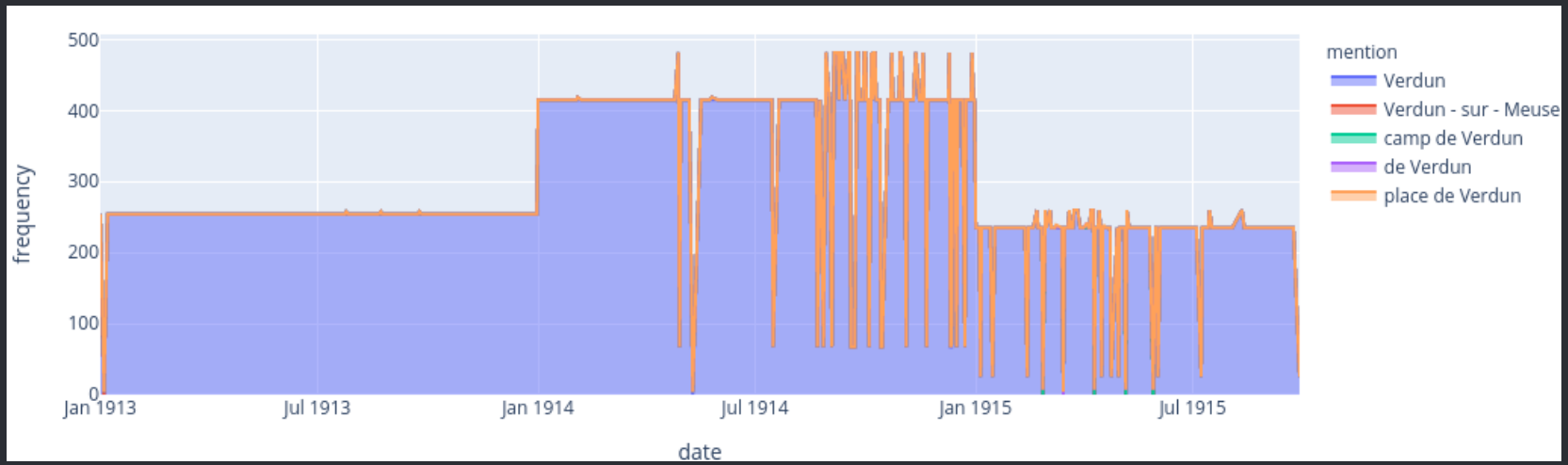
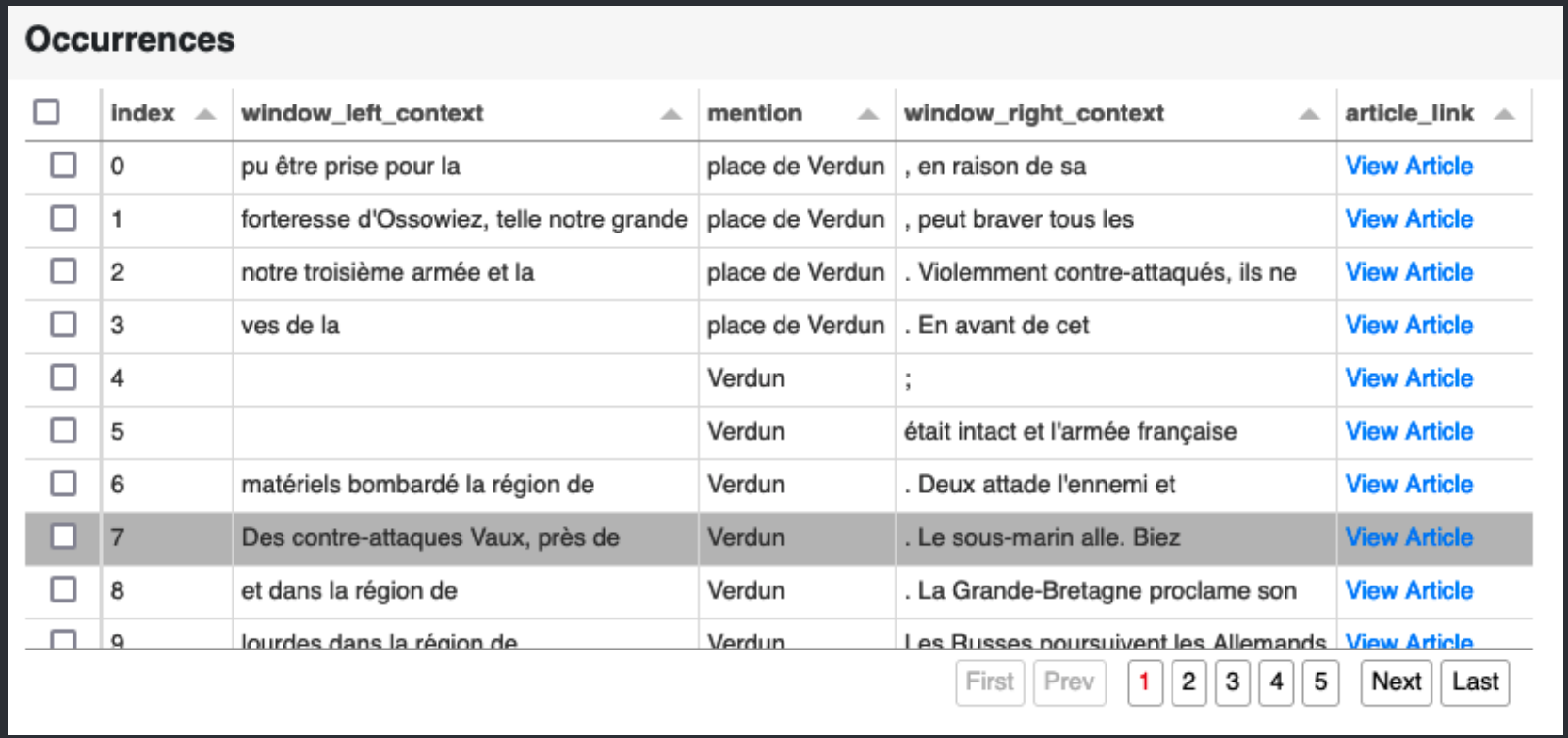


Figure – Occurrences de "Verdun" et ses variations dans *Le Matin* (1913-1915) (Gutehrlé et al., 2021)

Vers de nouvelles interfaces de recherche

Un concordancier permet d'observer l'emploi des termes dans leur contexte



Occurrences

index	window_left_context	mention	window_right_context	article_link
0	pu être prise pour la	place de Verdun	, en raison de sa	View Article
1	forteresse d'Ossowiez, telle notre grande	place de Verdun	, peut braver tous les	View Article
2	notre troisième armée et la	place de Verdun	. Violemment contre-attaqués, ils ne	View Article
3	ves de la	place de Verdun	. En avant de cet	View Article
4		Verdun	;	View Article
5		Verdun	était intact et l'armée française	View Article
6	matériels bombardé la région de	Verdun	. Deux attade l'ennemi et	View Article
7	Des contre-attaques Vaux, près de	Verdun	. Le sous-marin alle. Biez	View Article
8	et dans la région de	Verdun	. La Grande-Bretagne proclame son	View Article
9	lourdes dans la région de	Verdun	Les Russes poursuivent les Allemands	View Article

First Prev 1 2 3 4 5 Next Last

Figure – Occurrences des emplois de "Verdun" en contexte dans *Le Matin* (1913-1915) dans un concordancier (Gutehrlé et al., 2021)

Vers de nouvelles interfaces de recherche

Le géocodage (*geocoding*) pour obtenir les coordonnées des lieux mentionnés à partir de *gazeteer* (ex : Geonames) et générer des cartes automatiquement

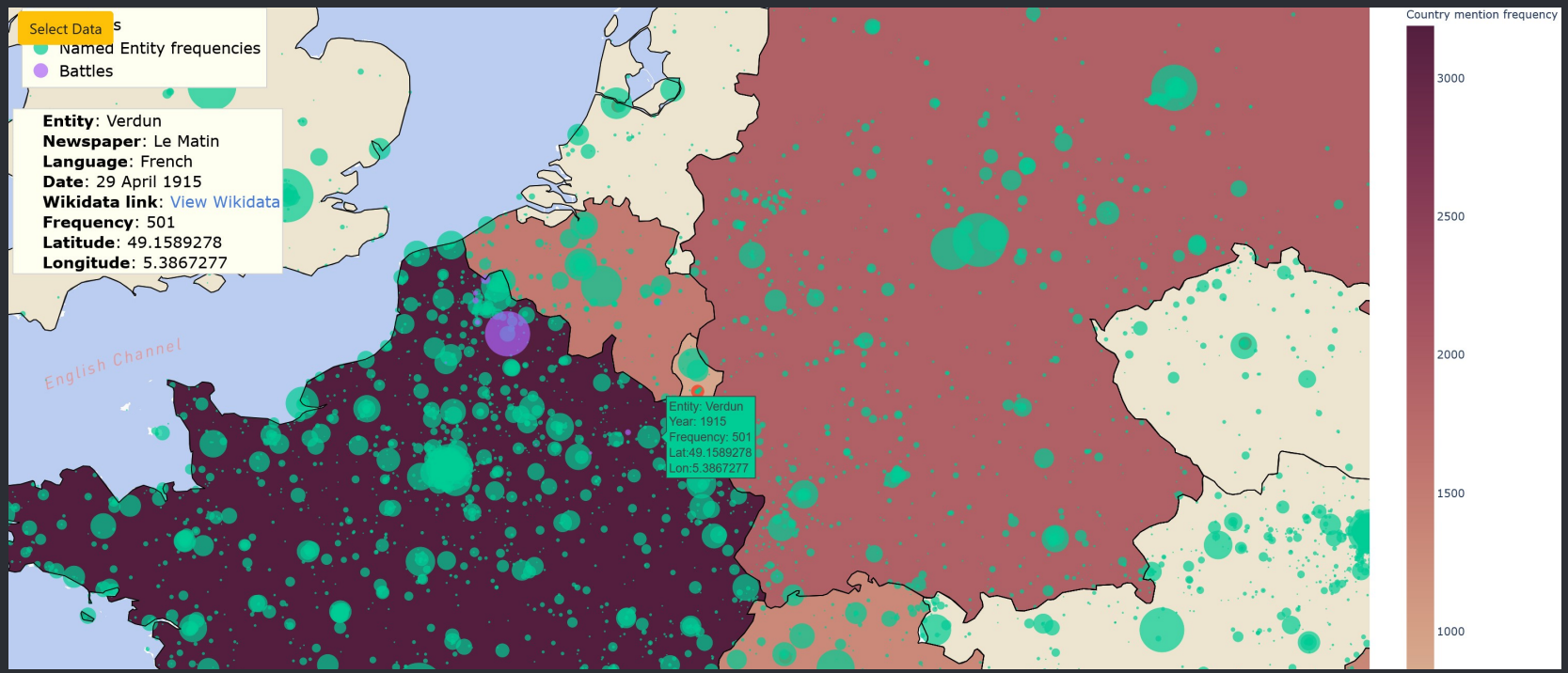


Figure – Cartographie des lieux mentionnés dans *Le Matin* (1913-1915), augmentée d'informations contextuelles (Gutehrlé et al., 2021)

Vers de nouvelles interfaces de recherche

Les entités nommées et leurs relations peuvent être visualisées sous forme de réseau

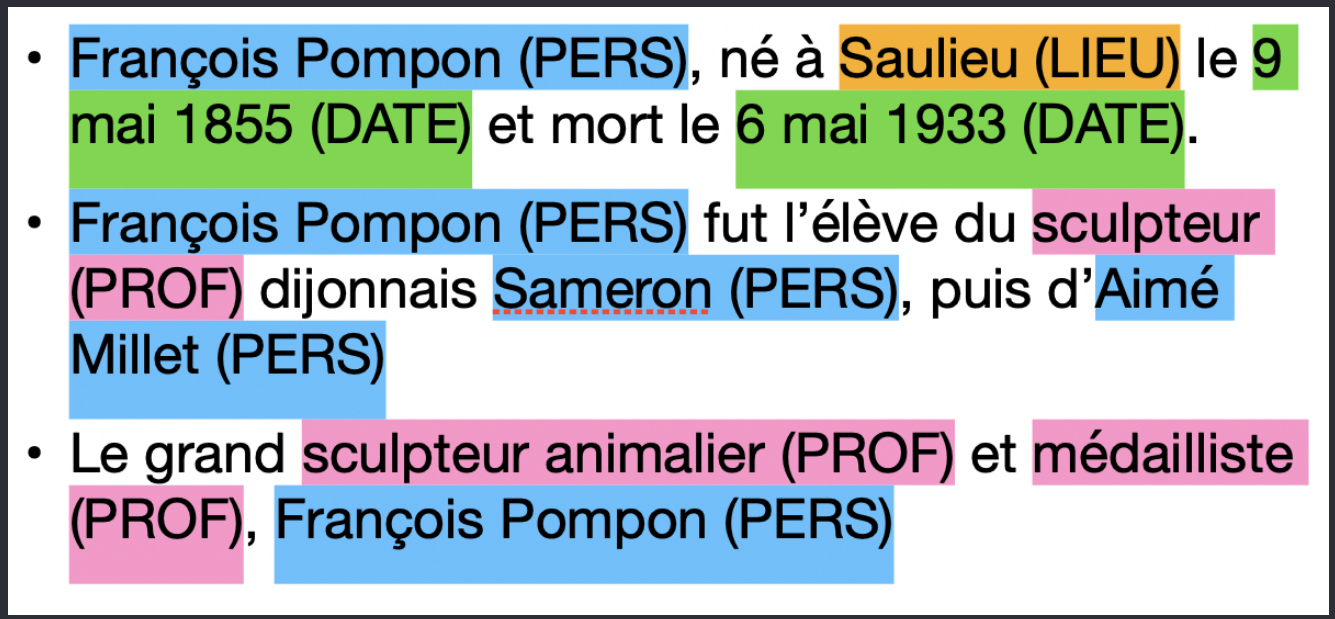


Figure – Exemple d'entités nommées identifiées dans le corpus EMONTAL

Vers de nouvelles interfaces de recherche

Les entités nommées et leurs relations peuvent être visualisées sous forme de réseau

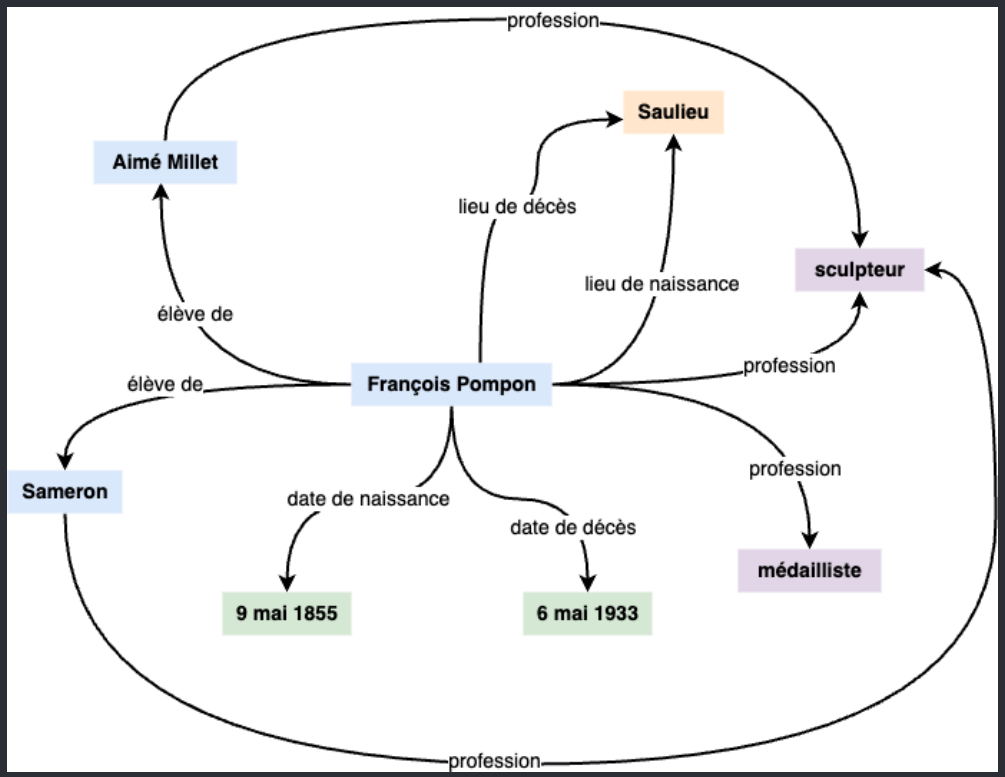


Figure – Exemple d'entités nommées et de leurs relations extraites du corpus EMONTAL

Vers de nouvelles interfaces de recherche

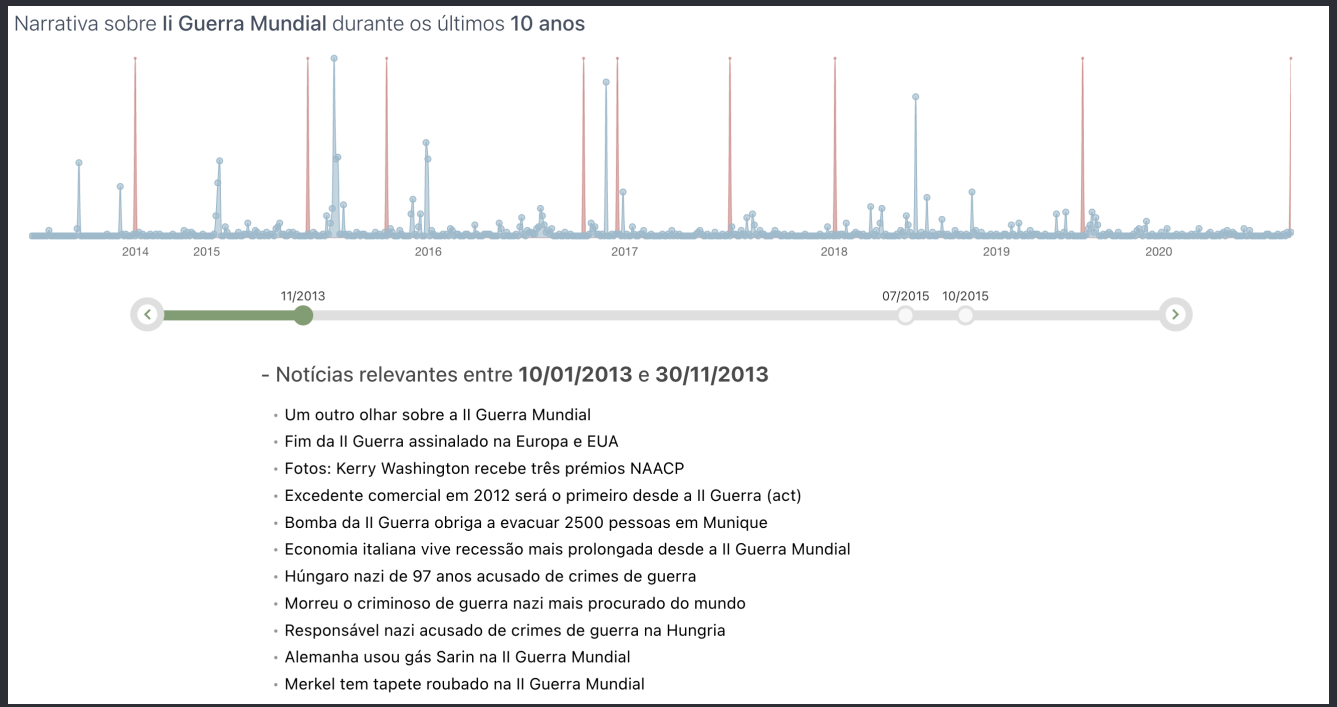


Figure – Frise chronologique générée par Conta-me Histórias (*Tell Me Stories*) (Pasquali et al., 2019) à partir des événements liés à la Seconde Guerre Mondiale mentionnés dans la presse des dix dernières années

Vers de nouvelles interfaces de recherche

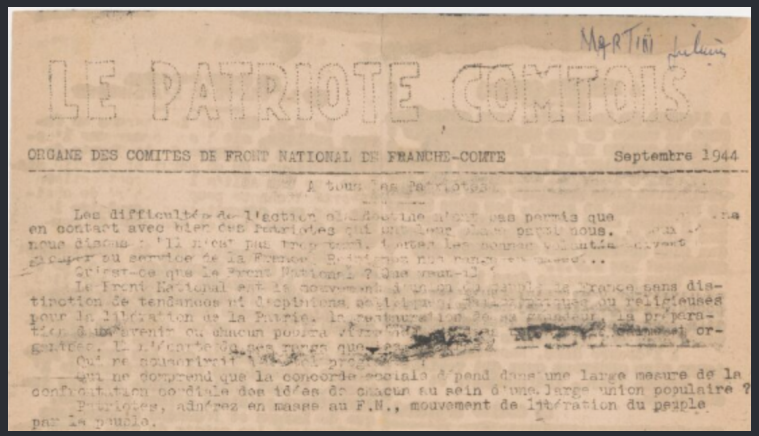
L'amélioration des interfaces existantes et la création de nouvelles interfaces faciliteraient l'ouverture et l'accessibilité des documents d'archives :

- les annotations sémantiques permettraient d'indexer les documents à des niveaux plus fins, et donc de soumettre des requêtes plus précises qui limiteraient le nombre de documents retournés
- les nouvelles interfaces viendraient synthétiser les résultats d'une requête et assisteraient dans la lecture distante
- les annotations et les nouvelles interfaces peuvent s'associer : par exemple, générer des frises chronologiques propres à chaque entité nommée identifiée

Enjeux techniques et épistémologiques du traitement des documents d’archives

Défis techniques

L'état ancien et historique des documents d'archives soulève de nombreuses difficultés quant à leur traitement automatique

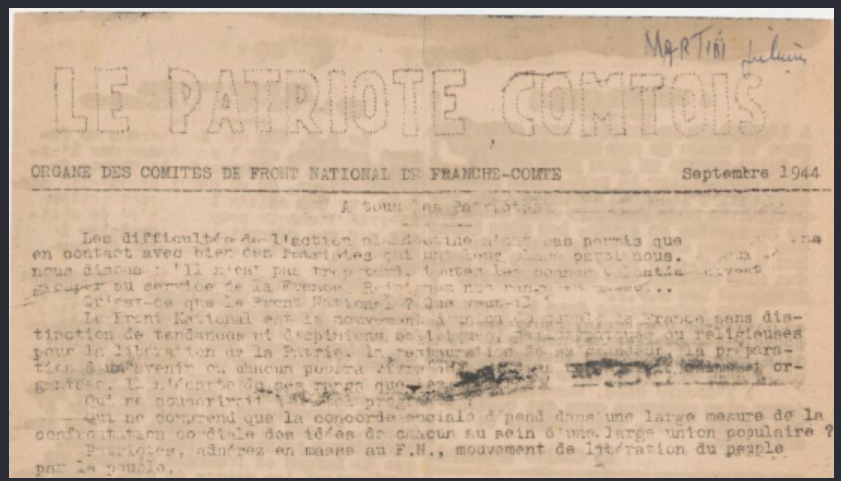

LE PATRIOTE COMTOIS

ORGANE DES COMITES DE FRONT NATIONAL DE FRANCHE-COMTE Septembre 1944

Défis techniques

Le contenu textuel des documents est habituellement obtenu à l'aide de méthode de reconnaissance optique de caractères (*Optical Character Recognition*, OCR)

- La qualité des transcriptions dépend de l'état du document d'origine et sa version numérique
- Impacte les résultats des méthodes qui y seront employées

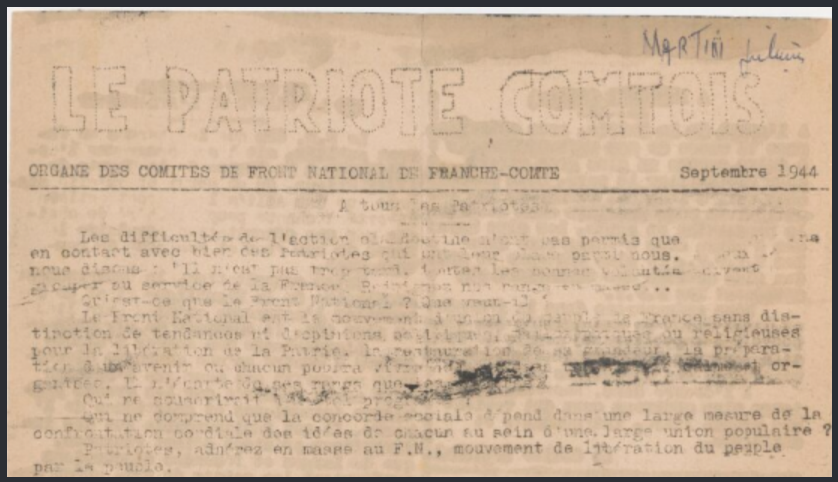


LE PATRIOTE COMTOIS

ORGANE DES COMITES DE FRONT NATIONAL DE FRANCHE-COMTE Septembre 1944

Défis techniques

Scan d'une page du *Patriote Comtois* (septembre 1944), accompagné de sa transcription obtenue par OCR



LE PATRIOTE COMTOIS

ORGANE DES COMITES DE FRONT NATIONAL DE FRANCHE-COMTE Septembre 1944

```
- ---"

- IJb-'

i

ORGANE DES COMITES DE FRONT NATIONAL DP FRANCHE-COMTE
 Septembre 1.944

--:---------- A tour; ^trioty.3 , 1 , y.: ;

'i

1' ~; ,él,":'i,;,'.,'l 'lA :L y ,-:'-,i~, H.:\ S n~-r"r.i q quP ,', :.° 1

Les dif fleixljb<5 , rr>- yu>- x' action olcw.d-'^a^uie n* ont oas permis ; que n' ♦ \I
en contact avec - bien des J^c-rtrlo-tes qui cni il^ur. .nous disons : '"Il î; ▪ cb1 »
 pat; tjk> iXTr^ - i /ites 1er-: 1aonn^o T>» :1 jjitcCs. :-c.uv^>i"C ;
au s-^T,ricfî"' ô.e j.a pt--i.-nr. no-"5 r,->.r.urî-- -«•->- »



)..J.J.u '-,~, v,L _o"-,--,,.).J.1.-" U' .,. ,J ;1,_,-8

qf.i¿ la -viotoi! re totale le r-0(i.la.t 'Uï!

la lil-t-rte Pour cela la bataille oon—i
ijnuej pour nous 1? action civxqua--ccm-?•
me o oe ,, i *
```

Défis techniques

La majorités des outils OCR et TAL disponibles aujourd'hui sont dédiés au traitement de la langue moderne. Or, la langue des documents historiques peut différer plus moins fortement avec la langue moderne


A VENDRE
Petit Cheval
7 ans, harnaché, avec ou sans voiture.
S'ad. au bur. du journal. 51

Figure – Extrait d'une annonce extraite du *Salinois* (1931, n°4), écrit dans un style ancien et télégraphique (Gutehrlé et Lethier, 2019)

Défis techniques

Figure – Exemple de l’emploi du "s" long dans le mot "congress" dans la Déclaration des droits des Etats-Unis (1789) (Wikipedia, 2023)

Défis techniques

Il est donc nécessaire d'adapter les outils existants ou d'en créer de nouveaux :

- l'emploi de ressources non adaptées (ex : Wikidata, geonames) peut fausser les résultats
- les ressources adaptées (ex : Hamdi et al., 2021 pour la REN, Gutehrlé et Atanassova, 2021 pour la LLA) sont lacunaires

Interprétation des résultats

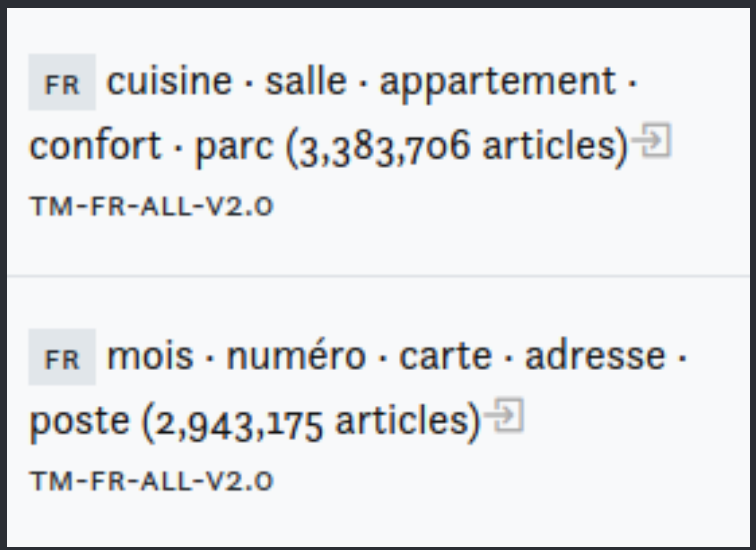
Les annotations et nouvelles interfaces doivent assister dans la recherche :

- en permettant des recherches plus fines
- en synthétisant les résultats obtenus

Pour garantir leur validité, les annotations et les processus qui y mènent doivent être interprétables.

Interprétation des résultats

L'interprétation des résultats nécessite parfois une expertise, ce qui limite l'utilité de certaines méthodes pour un public large



Exemple de thèmes identifiés Topic Modelling dans la plateforme impresso. Les thèmes sont représentés sous forme de nuage de mots

Interprétation des résultats

Certaines méthodes génèrent du texte à partir des données d'origine. Cependant, ces méthodes tendent à inventer des informations (hallucinations), ce qui déconnecte des documents d'origine et peut en fausser l'étude

PTGEN	UKIP leader Nigel Goldsmith has been elected as the new mayor of London to elect a new Conservative MP.	[45.7, 6.1, 28.6]
TCONVS2S	Former London mayoral candidate Zac Goldsmith has been chosen to stand in the London mayoral election.	[50.0, 26.7, 37.5]
TRANS2S	Former London mayor Sadiq Khan has been chosen as the candidate to be the next mayor of London.	[35.3, 12.5, 23.5]
GPT-TUNED	Conservative MP Zac Goldwin's bid to become Labour's candidate in the 2016 London mayoral election.	[42.4, 25.8, 36.4]
BERTS2S	Zac Goldsmith has been chosen to contest the London mayoral election.	[66.7, 40.0, 51.9]

Figure – Exemples de textes générés par différentes méthodes résumant un article de presse (Maynez et al., 2020)

Conclusion

Conclusion

L'exploitation et la valorisation des collections de documents d'archive requiert de nouveaux outils :

- les interfaces actuelles peuvent être augmentées et exploiter les annotations sémantiques ajoutées aux documents
- de nouvelles interfaces exploitant les résultats de ces traitements peuvent assister dans la compréhension des documents obtenus

Conclusion

L'ajout de ces annotations sémantiques implique de nombreux défis techniques :

- correction de l'OCR
- nettoyage du texte (césure…)
- emploi d'outils adaptés à langue des documents
- besoin d'outils explicites pour les utilisateurs

Explicité des résultats

Et soulève plusieurs questions :

- Comment peut-on faciliter la création de ressources adaptées aux documents historiques ? Devrait-on se concentrer sur la création de ces ressources ?
- Est-il préférable d'employer des systèmes moins performants mais plus explicites et moins gourmands en ressources ?
- Devrait-on évaluer les outils en TAL selon leur utilité et explicité, et non pas seulement selon leurs performances (Précision, Rappel, F1) ?
- Quand est-il de l'accessibilité à l'état de l'art (ex : BERT, GPT) ?

Merci pour votre attention !

Bibliographie

Gutehrlé, N. & Atanassova, I. (2021). Dataset for Logical-layout analysis on French historical newspapers.

Gutehrlé, N., Harlamov, O., Karimi, F., Wei, H., Jean-Caurant, A. & Pivovarova, L. (2021). SpaceWars : A Web Interface for Exploring the Spatio-temporal Dimensions of WWI Newspaper Reporting. *CEUR Workshop Proceedings*.

Gutehrlé, N. & Lethier, V. (2019). Cartographier les données textuelles des petites annonces du Salinois (1840-1939). *Journée d'études Humaspatia*.

Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G. & Doucet, A. (2021). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2328-2334.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E. & Coates, C. M. (2015). Trading consequences : A case study of combining text mining and visualization to facilitate document exploration. *Digital Scholarship in the Humanities*, *30*(suppl_1), i50-i75.

Jatowt, A. (2021). Timeline as Information Retrieval and Ranking Unit in News Search. *DESIRES*.

Kaplan, F. & di Lenardo, I. (2017). Big Data of the Past. *Frontiers in Digital Humanities*, *4*, 12. https://doi.org/10.3389/fdigh.2017.00012

Mann, J., Zhang, D., Yang, L., Das, D. & Petrov, S. (2014). Enhanced search with wildcards and morphological inflections in the Google Books Ngram Viewer. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, 115-120.

Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. https://doi.org/10.48550/ARXIV.2005.00661

Moretti, F. (2013). *Distant reading*. Verso Books.

Pasquali, A., Mangaravite, V., Campos, R., Jorge, A. M. & Jatowt, A. (2019). Interactive System for Automatically Generating Temporal Narratives. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff & D. Hiemstra (Éd.), *Advances in Information Retrieval* (p. 251-255). Springer International Publishing.

Wikipedia. (2023). Long s — Wikipedia, The Free Encyclopedia [[Online ; accessed 25-January-2023]].