



**HAL**  
open science

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

Nicolas Gutehrlé

## ► To cite this version:

Nicolas Gutehrlé. Comprendre les archives : explorer et valoriser les documents historiques grâce à l'annotation sémantique. Printemps de la Donnée 2024, INRAE; Université Haute-Alsace; Université de Strasbourg; INSA; PNDB; AgroParisTech; Université de Lille; Sorbonne Université; Data Terra, May 2024, Besançon, France. 10.5281/zenodo.11609500 . hal-04683706

**HAL Id: hal-04683706**

**<https://hal.inrae.fr/hal-04683706v1>**

Submitted on 2 Sep 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# 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

1. Introduction
2. Annotation sémantique des documents d'archives
  - 2.1 Améliorer les interfaces de recherche actuelles
  - 2.2 Vers de nouvelles interfaces de recherche
3. Enjeux techniques et épistémologiques
  - 3.1 Défis techniques
  - 3.2 Interprétation des résultats
4. Conclusion

# 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

NUMERO 70La Nouvelle O.S.S.Mardi 11 Avril 1938

# LE FRONT OUVRIER

Organe hebdomadaire du prolétariat montcellien  
Administrateur-délégué: VITTEALAT A.-R., 5, rue de France - E. C. Postal N° 320 50

QUI HABILLE LE MIEUX ?

C'est Paul MILLET

Collection de vestes et de chemises  
drus au service grand du travail.

## Les marchands de mort

« Pour le pain, contre la guerre, à mort sur la guerre »  
Chuck HALL

(de la Nouvelle Montcellien, n. 67)

Une fois que l'Union Montcellienne a été créée, elle a commencé à fonctionner. Elle a commencé à fonctionner en organisant des manifestations contre la guerre, en organisant des manifestations contre la guerre, en organisant des manifestations contre la guerre...

## La Réunion de Mont-Saint-Vincent

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

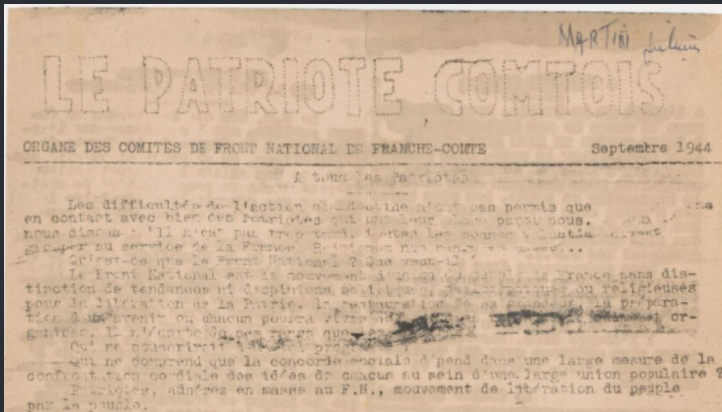
Il est certain que la réunion de Mont-Saint-Vincent a été une réussite. Elle a permis de réunir un grand nombre de camarades et de discuter de nos problèmes communs. Nous espérons que cela sera le début d'une série de réunions similaires...

## Billet de la semaine

A des signatures

Il est important de collecter des signatures pour notre campagne. Nous avons besoin de votre soutien et de votre voix. Merci de participer activement à cette tâche essentielle...

# Projet EMONTAL



## 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)

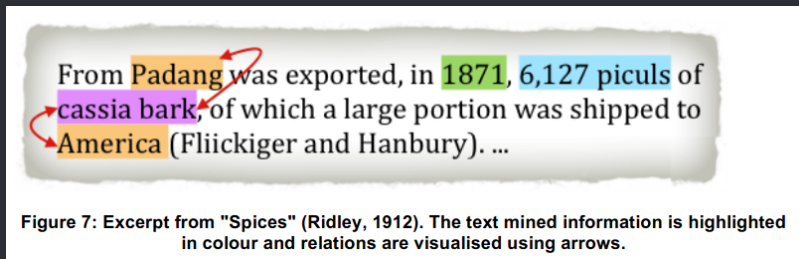


Figure – Exemple de reconnaissance d'entités nommées et d'extraction de relations dans le projet *Trading Consequences* (Hinrichs et al., 2015)

# Annotation sémantique des documents d'archives

Modélisation de sujet (*Topic Modelling, TM*)

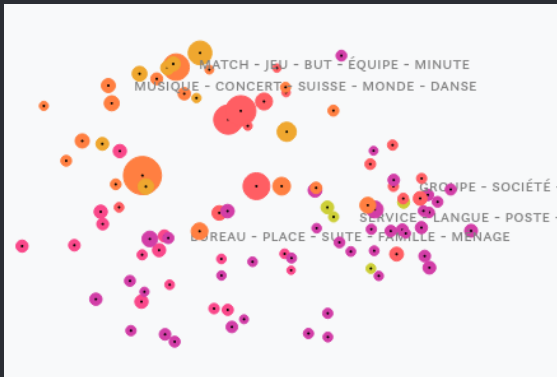


Figure – Exemple de thèmes identifiés dans le corpus de l'interface *impresso*

# Annotation sémantique des documents d'archives

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



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 annotations 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 découvrabilité des documents

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

# Annotation sémantique des documents d'archives

The screenshot displays the Impresso search interface. On the left, there are filters for 'PUBLICATION DATE' (a bar chart showing a peak in 1912) and 'FILTER BY CONTENT LENGTH' (a bar chart). Below these are filters for 'LANGUAGE OF ARTICLES' (French: 57 results, German: 8 results) and 'NEWSPAPER TITLES' (L'Express: 15 results, L'Impartial: 15 results). The main search area shows 'SEARCH ARTICLES' with 'titanic' and 'Belfast' entered. The results are grouped by 'ARTICLE' and show 65 articles found. Two articles are visible:

- Großbritannien.**  
Luxemburger Wort | WEDNESDAY, JULY 31, 1912 – p.2  
Personal use (no export)  
Großbritannien. Nachtlänge für Titanic-Katastrophe. In dem Urteil für Katastrophe bes „Titanic“ sagt Lord Merfy: Sie Umstände überzeugten ihn davon, LOCATIONS Belfast |  
Großbritannien. Nachtlänge für Titanic-Katastrophe. In dem Urteil für Katastrophe bes „Titanic“ ber „Titanic“ gewefen fel. Sie Nacht fel klar und die See ruhig gewesen. Wenn die „Californian“, die die Noträlten der „Titanic“ gesehen habe, burch das Eis hindurchfahren wäre, was sie ohne Gefahr hätte tun können, fo hätte sie twaifelhährlich viele, wenn nicht alle Leben ber „Titanic“ gerettet. Sas habe. — Sie Sarammtig für die Hinterbliebenen ber Opfer ber „Titanic“-Katastrophe hat im Ganzen die Summe von 8 250 000  
VIEW ADD TO COLLECTION ...
- AU BERCEAU DU «TITANIC»**  
L'Express | FRIDAY, MAY 23, 2014 – p.16  
Personal use  
AU BERCEAU DU « TITANIC » MÉMORIAL Une construction entièrement dédiée au « Titanic ». L'Irlande du Nord souffre encore de l'image véhiculée par ses LOCATIONS Nord (French department) | Belfast |, Florence |  
AU BERCEAU DU « TITANIC » MÉMORIAL Une construction entièrement dédiée au « Titanic ». L'Irlande universitaire et culturelle y sont pour beaucoup. Sans oublier les vastes chantiers du « Titanic », où  
VIEW ADD TO COLLECTION ...

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



# Annotation sémantique des documents d'archives

The screenshot displays the NewsEye search interface. At the top, there is a navigation bar with 'NEWS E E' and a 'META' badge. The main search area includes a search bar with the query 'titanic belfast', a search button, and a 'Suggest keywords' button. Below the search bar, there are options for 'Documents: Articles' and 'Search: Exact'. A 'Dates Frequencies' bar chart shows the distribution of search results by year from 1908 to 1914, with a significant peak in 1912. The search results are displayed in a list format, with the first result being '1. vastra\_finland\_791950\_article\_478' from 'Vastra Finland' dated '18/05/1909'. The second result is '2. neue\_freie\_presse\_nfp19120624\_article\_16' from 'Neue Freie Presse' dated '24/06/1912'. A sidebar on the left allows users to filter their search by 'Langue', 'Date', 'Month', 'Day', 'Journal', and 'Locations'. The 'Locations' filter is expanded, showing 'Londres' (5), 'Belfast' (2), 'Prague' (1), and 'peltun' (1). The main content area shows the search results with options to 'See extracts' for each result. A snippet of text from the second result is visible, discussing the Titanic disaster.

NEWS E E

Search Datasets Saved searches Experiments

Documents: Articles

Search: Exact

titanic belfast

Chercher

Suggest keywords

Dates Frequencies

7  
6  
5  
4  
3  
2  
1  
0

1908 1909 1910 1911 1912 1913 1914

Vous avez demandé :

Articles (Exact Search) > titanic belfast

Get a random sample Add all documents to a dataset Save search

« Précédente | 1 - 10 sur 13 | Suivante »

Trier par relevance 10 par page

1. vastra\_finland\_791950\_article\_478

Date: 18/05/1909

Journal: Vastra Finland

See extracts

2. neue\_freie\_presse\_nfp19120624\_article\_16

Date: 24/06/1912

Journal: Neue Freie Presse

See extracts

Langue >

Date >

Month >

Day >

Journal >

Locations v

Londres 5

Belfast 2

Prague 1

peltun 1

Renntnis aller Verhältnisse, des Widerstandes, der Unfähigkeit des Schiffes, die Unversehrtheit zu bewahren, und ihre Kinderfranzösischen Durchschnittsmänner, in ein Schiff wie die „Titanic“, von der die Schreiber Befehl mit voller Sicherheit gaben, daß sie nicht nur könne, wird vom Meer verschlungen wie eine led' werdende Nille in der Donau.

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)

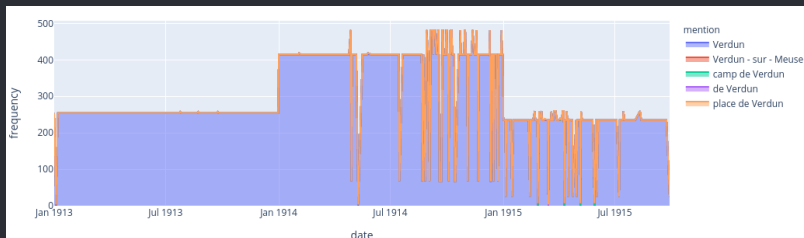


Figure – Occurrences de "Verdun" et ses variations dans *Le Matin* (1913-1915) (Gutehrlé et al., 2021)

## Vers de nouvelles interfaces de recherche

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

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

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 (Gutehrlié 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

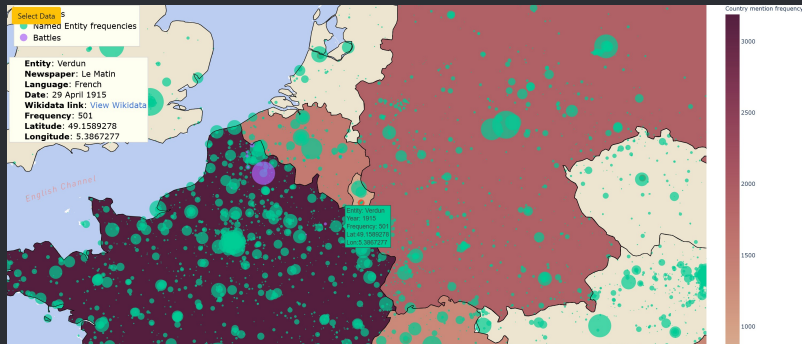


Figure – Cartographie des lieux mentionnés dans *Le Matin* (1913-1915), augmentée d'informations contextuelles (Gutehrle et al., 2021)

## Vers de nouvelles interfaces de recherche

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

- François Pompon (PERS), né à Saulieu (LIEU) le 9 mai 1855 (DATE) et mort le 6 mai 1933 (DATE).
- François Pompon (PERS) fut l'élève du sculpteur (PROF) dijonnais Sameron (PERS), puis d'Aimé Millet (PERS)
- Le grand sculpteur animalier (PROF) et médailleur (PROF), François Pompon (PERS)

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

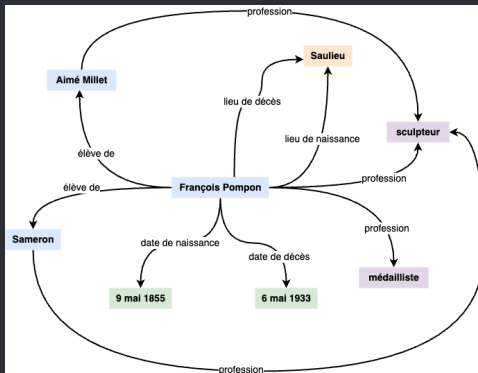


Figure – Exemple d'entités nommées et de leurs relations extraites du corpus EMONTAL

# Vers de nouvelles interfaces de recherche

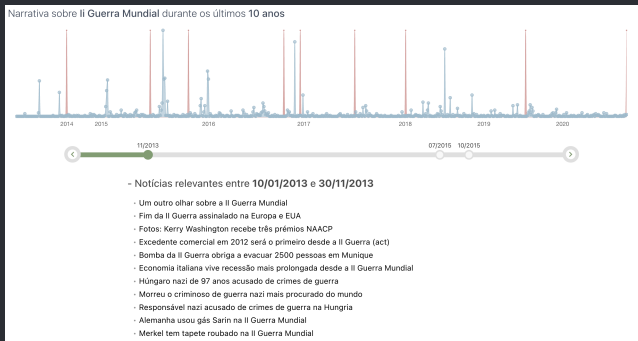


Figure – Frise chronologique générée par *Conta-me Histórias (Tell Me Stories)* (Pasquali et al., 2019) à partir des événements liés à la Seconde Guerre Mondiale mentionnés dans la presse des dix dernières années



## Vers de nouvelles interfaces de recherche

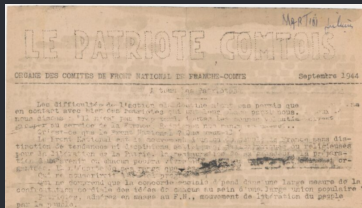
L'amélioration des interfaces existantes et la création de nouvelles interfaces faciliteraient l'ouverture et l'accessibilité des documents d'archives :

- 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

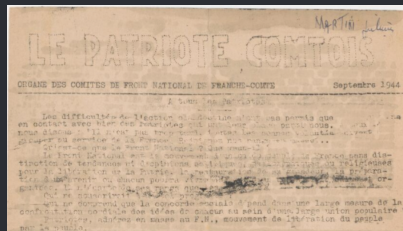
L'état ancien et historique des documents d'archives soulève de nombreuses difficultés quant à leur traitement automatique



# Défis techniques

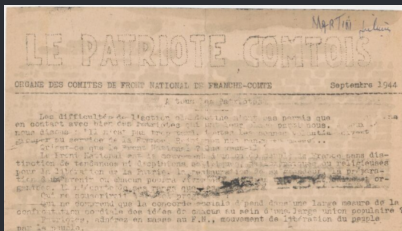
Le contenu textuel des documents est habituellement obtenu à l'aide de méthode de **reconnaissance optique de caractères (Optical Character Recognition, OCR)**

- 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



# Défis techniques

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



```
- ---*
- lJb'
i
ORGANE DES COMITES DE FRONT NATIONAL DP FRANCHE-COMTE
Septembre 1.944
----- A tour; ^trioty.3 , 1 , y.: ;
i
l' ~; ,él, "i; ; ,l 'lA :L y ,-'-i-, H.A.S n--r"ri q quP , ; :° 1
Les dif fleixljb<5 , r>- yu>- x' action olcw.d'^a^uie n* ont oas permis ; que n' ♦ V
en contact avec - bien des J^c-rtrlo-tes qui cni il'ur. nous disons : "Il i; ■ cb1 »
pat; tjk> iXTr^ - i /ites ler- : laonn^o T>» :1 jjitcCs. :-c.uv^>i^C ;
au s-^T,ricf" δ.e j.a pt--i.-nr. no-"5 r,-> r.urf-- «>-> »
).J.J.u '-,-, vL _o"-,-,).J.l.-" U" .. J :l _,-8
qf.i,é la -viotoi! re totale le r-0(i.l.at 'U!
la lil-t-rte Pour cela la bataille oon—i
ijnuej pour nous !? action civxqua--cem-?•
me o oe ,, i *
```

## Défis techniques

La majorité 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 ou moins fortement avec la langue moderne



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

FR cuisine · salle · appartement · confort · parc (3,383,706 articles)   
TM-FR-ALL-V2.0

FR mois · numéro · carte · adresse · poste (2,943,175 articles)   
TM-FR-ALL-V2.0

FR maison · vente · choix · magasin · qualité (3,962,713 articles)   
TM-FR-ALL-V2.0

FR bureau · place · suite · famille · ménage (3,783,673 articles)   
TM-FR-ALL-V2.0

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

<b>PTGEN</b>	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]
<b>TCONVS2S</b>	Former London mayoral candidate Zac Goldsmith has been chosen to stand in the London mayoral election.	[50.0, 26.7, 37.5]
<b>TRANS2S</b>	Former London mayor Sadiq Khan has been chosen as the candidate to be the next mayor of London.	[35.3, 12.5, 23.5]
<b>GPT-TUNED</b>	Conservative MP Zac Goldwin's bid to become Labour's candidate in the 2016 London mayoral election.	[42.4, 25.8, 36.4]
<b>BERTS2S</b>	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]].