



**HAL**  
open science

## Analyser un processus mémoriel au travers des archives de presse numérisées et physiques

Malorie Guilbaud Perez

► **To cite this version:**

Malorie Guilbaud Perez. Analyser un processus mémoriel au travers des archives de presse numérisées et physiques : L'exemple du Triangle Fire. *Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*, 3, De Gruyter Oldenbourg, pp.335 - 358, 2022, *Studies in Digital History and Hermeneutics*, 9783110729719. 10.1515/9783110729214-015 . hal-04682821

**HAL Id: hal-04682821**

**<https://hal.science/hal-04682821>**

Submitted on 8 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 - NoDerivatives 4.0 International License

**Malorie Guilbaud-Perez**

**Analyser un processus mémoriel au travers des archives de presse numérisées et physiques : l'exemple du Triangle Fire.**

Le 25 mars 1911, au sein du Lower East Side de Manhattan, survient l'incendie de l'atelier de fabrication textile new-yorkais<sup>i</sup>, « Triangle Factory », installée aux trois derniers étages du Asch building, et qui, parmi les cinq cents employés de la compagnie, provoqua 146 victimes, quasi exclusivement de jeunes femmes migrantes, originaires d'Europe de l'Est ou du Sud. Cet accident industriel<sup>ii</sup> engendra des réactions multiples dont notamment une forte mobilisation des employés new-yorkais de l'industrie textile massivement présents à la suite du drame lors de cérémonies collectives comme à l'occasion des procès des propriétaires de la Triangle, Max Blanck et Isaac Harris. Preuve de l'actualité de cet événement, cette tragédie a été récemment rappelée par l'ancienne candidate à l'investiture démocrate, Elisabeth Warren<sup>iii</sup>, un exemple militant choisi par la sénatrice du Massachusetts pour dénoncer le « sweatshop system »<sup>iv</sup> d'hier à aujourd'hui. L'objectif global de notre étude dans lequel s'insère cet article est ainsi d'étudier les processus à l'œuvre dans l'inscription de ce drame dans la mémoire nationale américaine, comment l'accident industriel, le fait divers, a dépassé ce statut initial pour devenir un événement plus complexe.

Au début du XXème siècle, la presse écrite est un média de masse, la source d'information comme de formation de l'opinion la plus importante dans l'ensemble des territoires industrialisés<sup>v</sup>. La dynamique des publications est particulièrement importante aux Etats-Unis<sup>vi</sup>, accompagnant le tournant industriel et urbain de la nation américaine. Tout au long du siècle précédent, nombreux titres de presse ont pu acquérir de plus en plus d'autonomie, voire d'indépendance vis-à-vis des pouvoirs politiques<sup>vii</sup>, relayant les différents courants de réformes qui traversent alors la société américaine, mais également devenant des objets de consommation de plus en plus diffusés, de plus en plus populaire<sup>viii</sup>, entre les mains d'entrepreneurs aux activités économiques diversifiées<sup>ix</sup> mais conçus par des journalistes se professionnalisant. La presse américaine, anglophone ou non, même en se limitant ici à son contenu textuel<sup>x</sup>, a donc été l'un des socles du processus de fabrication de l'évènement et de sa circulation, vecteur de nouvelles autant que de mémoires<sup>xi</sup> ou de protestations<sup>xii</sup>. Ce média est ainsi tout autant témoin de l'évènement, que reflet, imparfait, de celui-ci, vecteur d'une vision, d'une compréhension, d'une sémantique spécifique, et qui évoluent selon les époques, les lieux, les communautés ou les titres de presse considérés<sup>xiii</sup>. Une presse donc qui ne fait pas que témoigner mais dont les choix d'édition sont révélateurs d'enjeux politiques et sociétaux qui traversent une nation tout entière et l'interroge sur les valeurs qui la fonde.

L'enjeu de l'étude en cours est ainsi de parvenir à construire, en partie grâce à la presse, la circulation diachronique et synchronique de l'évènement, à tracer des continuités et des ruptures en commençant par construire un corpus de documents pertinents, accessibles via une consultation in situ et/ou un accès en ligne, ce dernier étant de plus en plus facilité au travers de la multiplication d'initiatives de numérisation des publications des plus diverses<sup>xiv</sup>. Comment choisir et interroger les bases de données disponibles pour constituer un corpus adéquat aux objectifs de la recherche, qui puisse répondre à nos hypothèses.

**Constitution du corpus initial :**

Le choix initial a donc été de se focaliser sur un titre de presse new-yorkais, accessible en ligne et toujours publié depuis 1911, le *New-York Times*. A la date de l'incendie, le *Times* est l'un des huit quotidiens new-yorkais en circulation. Il n'est pas encore le plus important, mais a déjà imposé sa différence<sup>xv</sup> avec les deux journaux les plus vendus, *The World* acheté par Joseph Pulitzer en 1893, et *The Journal* détenu W. R. Hearst. Sa ligne éditoriale renouvelée par son acquéreur en 1896, Adolph

Osch, s'incarne au travers du refus du sensationnel, « *yellow journalism* », à l'inverse d'autres tirages<sup>xvi</sup>. Valorisant donc une presse de qualité, le *New York Times* vise un lectorat plutôt aisé, cultivé, sans perdre de vue la volonté de s'imposer dans le paysage médiatique new-yorkais, ce qui conduit son propriétaire à faire le choix de vendre son journal à un penny seulement. En 1911 la dynamique du *Times* est positive, toujours plus d'exemplaires sont vendus (de 9000 en 1896 à 352000 en 1921<sup>xvii</sup>), son lectorat s'agrandit, alors que ses concurrents directs eux stagnent ou déclinent. Ces choix permettent au *New York Times* d'acquérir une reconnaissance, « *the respectable New York Times* », à l'échelle nationale et internationale : « *It represents the best rather than the average of American journalism* »<sup>xviii</sup>. Et reste l'un des journaux les plus importants aujourd'hui malgré les difficultés rencontrées par la presse écrite traditionnelle.

Le *New York Times* a ainsi développé assez tôt des outils numériques avec plus ou moins de succès : éphémère *New York Pulse* entre 1983 et 1986, *Continuous News Desk* en 1991, site internet dès 1996. Depuis le journal n'a cessé de diversifier son offre web notamment pour résister à la multiplication des numériques gratuits. En 2014, le *New York Times* lance pour ses abonnés la « TimesMachine »<sup>xix</sup>, un lecteur de microfilm virtuel, qui rend progressivement accessible les articles du journal depuis sa création: première vague numérisant les articles publiés entre 1851 et 1980 (11 millions d'articles, 2,5 million de pages), deuxième vague mettant à disposition ceux publiés entre 1981 et 2002 (8035 éditions, 1,4 million d'articles, 1,6 million de pages), les articles postérieurs au 31 décembre 2002 étant accessibles via le moteur de recherche du site internet.

L'utilisation de la « TimesMachine » a donc permis la constitution d'un premier corpus d'étude, restreint à un titre de presse new-yorkais, mais dont la circulation peut être analysée sur le siècle écoulé après l'évènement. La recherche d'articles traitant de l'évènement a été réalisée sur l'ensemble des pages de la publication sur une période allant du 25 mars 1911 au 31 décembre 2002<sup>xx</sup>. Nous recherchons alors tous les types d'articles liés directement ou indirectement à l'évènement, que celui-ci soit le sujet de l'article ou bien une référence évoquée au sein du texte. Sur ce point la numérisation des journaux permet de ne plus se restreindre à une recherche par index mais bien sur l'ensemble des contenus. Il a donc fallu d'abord déterminer quelle combinaison de mots clés utiliser afin de borner correctement la recherche sans qu'elle soit trop discriminante<sup>xxi</sup>.

*Figure 1. Tableau présentant les résultats synthétiques d'une recherche par mot clé simple sur TimesMachine (1911-2002), résultats bruts (décembre 2019) :*

mot clé possible pour définir l'évènement	nombre d'articles contenant le mot clé	mot clé possible pour définir l'évènement	nombre d'articles contenant le mot clé
MARCH	67 970 054	WAIST	3 362 180
FIRE	41 833 002	TRIANGLE	2 996 169
FACTORY	16 585 043	COMPANY	2 166 923
HARRIS	12 960 934	SHIRTWAIST	175 870
ASH	6 362 450	ASCH	106 538
1911	6 208 163	BLANCK	29 129
BLANK	4 793 726		

Cette première évaluation de la base de données nous conduit à ne pas retenir les termes les plus discriminants comme les noms propres (« Asch » building ou « Blanck » l'un des propriétaires de la compagnie) ou des noms trop spécifiques comme « shirtwaist ». Les combinaisons les plus adéquates semblent être « fire + factory + triangle » ou « fire + factory + waist », des combinaisons de trois mots d'usage courant, pouvant définir au mieux l'évènement, mais aussi être utilisés dans des

contextes très différents. Se pose alors la question de l'adjonction d'un quatrième mot clé qui permettrait de restreindre d'avantage les résultats obtenus.

*Figure 2. Tableau présentant les résultats synthétiques d'une recherche par mot clé sur TimesMachine (1911-2002) et sur le site du New York Times en ligne (2003-2019), résultats bruts (décembre 2019) :*

Période	Fire + Factory + Waist	Fire + Factory + Triangle	Pertinence de l'ajout qu'un quatrième mot-clé ?						
			+1911	+company	+ shirtwaist	+ waist	+ harris	+ blanck	+ asch
1911-2002	<b>8 140</b>	<b>1 279</b>	<b>287</b>	<b>664</b>	<b>230</b>	<b>321</b>	<b>182</b>	<b>497</b>	<b>117</b>
2003-2019	<b>7 101</b>	<b>814</b>	<b>230</b>	<b>410</b>	<b>219</b>	<b>254</b>	<b>131</b>	<b>403</b>	<b>66</b>

Il apparaît ainsi que la combinaison la plus pertinente soit « fire + factory + triangle », à laquelle l'adjonction d'un quatrième mot serait trop restrictive. En effet, après vérification de leur contenu, le nombre d'articles réellement pertinents pour la période 1911-2002 est de 376, et pour la période 2003-2019 il est de 89. Donc dans 5 cas sur 7 l'ajout d'un quatrième mot-clé restreint trop les résultats obtenus.

*Figure 3. Tableau évaluant la fiabilité de la recherche par mot clé sur TimesMachine (décembre 2019)*

Période entre 1911 et 2002	Fire + Factory + Triangle	Avec l'ajout d'un quatrième mot-clé						
		+1911	+company	+ shirtwaist	+ waist	+ harris	+ blanck	+ asch
Résultats bruts	1 279	287	664	230	321	182	497	117
Résultats filtrés	376	95	152	136	94	37	41	35
% réellement utilisable	<b>30%</b>	<b>33%</b>	<b>23%</b>	<b>60%</b>	<b>30%</b>	<b>20%</b>	<b>8%</b>	<b>30%</b>

Tous les articles obtenus par cette recherche ne sont donc pas pertinents, car bien que contenant les mots clés, ils ne traitent pas de l'évènement recherché. La pertinence de la recherche s'établit donc autour des 30%. Celle-ci est moindre quand des noms propres sont utilisés comme mots clés (les noms des propriétaires Harris et Blanck). Un écart entre les résultats et les articles effectivement utiles car les 3 mots clés retenus sont des mots d'usage courant que l'on retrouve dans des contextes très différents notamment des articles liés aux deux conflits mondiaux, au conflit vietnamien, ou encore des faits divers et des informations culturelles. Pour ces derniers aspects, le manque de fiabilité est en partie dû au processus de segmentation réalisé au sein des journaux numérisés. La recherche par mots clés s'effectue aussi au sein d'espaces non segmentés en articles comme par exemple des pages présentant essentiellement des listes : listes nominatives<sup>xxii</sup> ou encore listes d'activités culturelles possibles, de livres disponibles, de spectacles du moment (critiques littéraires, programmes télévisuels, actualités du *Princeton's Triangle Club*, attractions, sorties « *spare times* ou *leisure times* »). Ces résultats non pertinents viennent gonfler le nombre d'articles potentiels donnés par le moteur de recherche.

D'autre part on observe également que la fiabilité de la recherche par mots clés est plus forte, quasi excellente, sur les 36 mois qui suivent la tragédie. Par contre elle s'étioule plus on s'éloigne de la date de l'évènement. Un facteur permet en partie d'expliquer ce constat, c'est la forme des articles du Times. Ceux-ci sont en effet de plus en plus fournis au fil des années pouvant s'étendre sur une, deux, trois voire quatre pages en fonction des thématiques, des « dossiers » traités, donc la possibilité de trouver la combinaison des trois mots clés est plus forte, l'interface ne limitant pas les résultats à une seule page mais proposant l'intégralité de l'article même si celui-ci est filé sur plusieurs pages.

### **Constitution des corpus miroirs :**

A l'issue de l'exploitation des ressources du New-York Times nous avons donc un corpus de 435 articles (corpus 1) directement lié l'évènement ou faisant référence explicitement à celui-ci, publiés à New-York entre 1911 et 2019. Or pour pouvoir appréhender la circulation de l'évènement dans l'espace et dans le temps, l'analyse de ce seul titre de presse n'est pas suffisante même si sa diffusion ne se limite pas à New-York. Nous avons besoin d'autres collections numérisées accessibles, des publications locales, nationales et internationales, proposant une durée de parution suffisamment étendue qui puisse soutenir la comparaison avec le New York Times, afin d'affiner les mécanismes synchroniques et diachroniques de présentation de l'évènement, de circulation au sein de la société américaine et au-delà.

Les titres de presse sont particulièrement nombreux sur le territoire américain. La publication de journaux aux Etats-Unis a atteint son maximum en 1909 avec 2600 titres en circulation, pour un tirage moyen de 13531 exemplaires (un tirage moyen plus élevé pour les éditions du dimanche). En parallèle, le lectorat abonné augmente passant en 1870 de 11,5% de la population éduquée à 45% en 1930. Pour l'année 1991, on ne compte plus que 1610 quotidiens en circulation, représentant un tirage cumulé de 60 millions d'exemplaires par jour<sup>xxiii</sup>.

Pour avoir un accès libre cette fois à des collections numérisées, nous nous sommes tournés vers le site internet de la bibliothèque du congrès. En effet, le travail de numérisation des collections de journaux et périodiques américains fait l'objet d'un programme spécifique le *National Digital Newspaper Program (NDNP)* qui associe le *National Endowment for the Humanities (NEH)* et la *Bibliothèque du Congrès (LC)*, le premier assurant le financement des projets de numérisation soumis par différentes organisations des états, le second fixant les cadres chronologiques des archives à numériser puis centralisant ensuite l'ensemble des données au sein d'un seul serveur, *Chronicling America*, permettant d'effectuer une recherche par mots clés au sein des pages numérisées des différents supports transmis à la bibliothèque du congrès<sup>xxiv</sup>.

Le programme *Chronicling America* rend donc à priori facilement accessible une collection numérisée de plus de 3000 journaux<sup>xxv</sup>, sur une période allant de 1789 à 1963. Donc à ce stade, même si les collections ne couvrent pas l'ensemble de la chronologie souhaitée, elles peuvent permettre en théorie l'analyse de la circulation temporelle et spatiale de l'évènement sur une période de cinquante ans. Cependant, il nous apparaît rapidement que la collection proposée en ligne est assez inégale en termes de couverture du territoire américain. En effet, le programme lancé en 2005 a été alimenté de manière disparate par différentes organisations (sociétés historiques, universités...) ayant répondu à l'appel à projet : certaines ont soumis régulièrement des demandes de financement et ont rendu accessible une part conséquente de leurs journaux historiques, alors que d'autres états ne sont pas encore représentés dans la liste en 2019.

Figure 4. Tableau synthétisant l'hétérogénéité des collections de journaux numérisés disponibles par « Chronicling America » (Library of Congress, décembre 2019)

Les 5 états les plus présents dans les collections de la LC			Les états les moins présents dans les collections numérisées de la LC		
	<i>Nombre de publications accessibles</i>	<i>Nombre de financements reçus depuis 2005</i>		<i>Nombre de publications accessibles</i>	<i>Nombre de financements reçus depuis 2005</i>
Mississippi	254	3	Wyoming	0	1 (2019)
Ohio	160	5	Virgin Islands	0	1 (2019)
Tennessee	126	3	Rhode Island	0	1 (2019)
South Carolina	125	3	Massachusetts	2	0
Louisiana	122	3	Puerto Rico	4	4
North Carolina	108	4	New Jersey	10	2
Delaware	106	3	Maine	13	2

De plus, une étude de la composition de la collection fait apparaître que les journaux numérisés sont en nombre important pour une période précédant notre événement. Le cœur de la collection est ainsi composé de journaux du XIX<sup>ème</sup> siècle, qui pour un grand nombre ne sont plus édités en 1911.

*Figure 5. Tableau synthétisant le nombre de publications numérisées disponibles sur Chronicling America correspondant à notre période d'étude (décembre 2019)*

<i>Nombre total de publications sur l'ensemble du territoire</i>	<i>Dont publications encore éditées entre 1911 et 1927</i>	<i>Dont publications encore éditées entre 1928 et 1945</i>	<i>Dont publications encore éditées entre 1946 et 1962</i>	<i>Dont publications encore éditées après 1963</i>	<i>Dont publications éditées en continu entre 1911 et au moins 1963</i>	<i>Publications éditées entre 1911 et au moins 1963 entièrement disponibles</i>
<b>3157</b>	<b>1144</b>	<b>560</b>	<b>399</b>	<b>288</b>	<b>210</b>	<b>0</b>

Très peu de titres couvrent en théorie une période adéquate pour notre recherche, d'autant plus que les collections par titre ne sont pas complètes. Les numérisations par titre ne sont pas exhaustives et couvrent une période bien plus courte. Par exemple, le quotidien le *Dermott News* publié en Arkansas entre 1910 à 1977 n'est accessible via la bibliothèque du congrès que pour les numéros couvrant la période de 1913 à 1919. Ces carences dans les collections numérisées sont inhérentes aux modalités de constitution de la base de données qui a été réalisée par strate chronologique : les projets soumis pour le premier cycle devaient ainsi proposer des numérisations de publications éditées entre 1900 et 1910, le cycle suivant s'est focalisé sur la période 1880-1910, puis 1880 – 1922, puis 1860 – 1922, puis 1836 – 1922, pour enfin recouvrir la période 1690 – 1963 pour le dernier cycle proposé en 2019.

La base de données de la bibliothèque du congrès ne pouvait donc que partiellement répondre à la constitution d'un corpus miroir sur le temps long, les possibilités étant majoritairement restreintes à la période 1911-1922, soit la diffusion de l'évènement uniquement sur la décennie lui succédant et non sur le siècle ou le demi-siècle écoulé. Mais elle présente l'avantage d'avoir une emprise territoriale assez importante bien qu'inégale. L'exploitation de la base de données de *Chronicling America* nous donne ainsi accès en théorie à 14537 pages de journaux comportant les mots clés ciblés. Si l'on réduit

ce nombre à la présence de ces mots clés uniquement en première page des publications seulement, on tombe à 1262 pages théoriquement pertinentes. Une fois étudiées individuellement nous obtenons un corpus effectif de 307 unes de presse (corpus 2) mettant réellement en avant le Triangle Fire, publiées entre mars 1911 et novembre 1915, par une diversité de 124 titres de presse différents.

*Figure 6. Tableau recensant le nombre de pages de journaux numérisés, dont les premières pages, répondant à notre recherche par mot clé, publiées entre 1911 et 1963 (date limite imposée par le moteur de recherche) pour chaque état ayant participé avant l'année 2019 au projet « Chronicling America » (Library of Congress – décembre 2019)*

Territoires proposant des publications entre 1911 et 1963	Potentiel ciblé par mots clés		Unes « réelles »	nb titres de presse		Territoires proposant des publications entre 1911 et 1963	Potentiel ciblé par mots clés		Unes « réelles »	nb titres de presse
	Toutes les pages	Unes seulement					Toutes les pages	Unes seulement		
CONNECTICUT	1006	100	30	2		CALIFORNIA	43	5	4	2
NEW YORK	813	27	23	3		LOUISIANA	136	21	4	4
WASHINGTON	190	32	17	6		NORTH CAROLINA	151	29	4	4
D. OF COLUMBIA	5501	124	14	3		OHIO	124	6	4	2
IOWA	281	42	14	3		TEXAS	207	12	4	1
OREGON	241	27	13	4		ALASKA	81	19	3	3
ARIZONA	257	32	12	4		ARKANSAS	38	6	3	2
VERMONT	255	45	10	4		DELAWARE	135	10	3	1
MICHIGAN	154	17	9	2		HAWAI	101	4	3	2
NEBRASKA	235	21	9	4		KENTUCKY	75	10	3	3
SOUTH CAROLINA	91	13	9	5		MINNESOTA	153	14	3	1
WEST VIRGINIA	177	13	9	5		IDAHO	200	21	2	2
NEW JERSEY	309	32	8	2		KANSAS	248	11	2	2
NORTH DAKOTA	343	28	8	4		PENNSYLVANIA	306	11	2	2
OKLAHOMA	128	21	8	4		COLORADO	60	3	1	1
SOUTH DAKOTA	136	15	8	4		GEORGIA	14	8	1	1
FLORIDA	157	25	7	3		MARYLAND	137	16	1	1
ILLINOIS	155	21	7	3		NEVADA	66	11	1	1
NEW MEXICO	175	17	7	3		TENNESSEE	99	9	1	1
UTAH	178	11	7	3		ALABAMA	1	0	0	0
INDIANA	484	43	6	2		MAINE	45	0	0	0
MISSOURI	114	13	6	3		MASSACHUSETTS	0	0	0	0
MONTANA	132	11	6	4		PIEDMONT	0	0	0	0
VIRGINIA	221	33	6	4		PUERTO RICO	0	0	0	0
MISSISSIPPI	121	8	5	4		WISCONSIN	263	22	0	0

Ce premier corpus complémentaire nous permet d'envisager la circulation de l'évènement entre 1911 et 1915 seulement mais au sein d'une majorité d'Etats du territoire américain. Il n'est cependant pas suffisant pour répondre à nos hypothèses de travail. Nous avons donc essayé de compléter le corpus obtenu par les ressources numérisées mises à disposition gratuitement par la bibliothèque du Congrès en utilisant le moteur de recherche ProQuest afin d'évaluer si la circulation temporelle de l'évènement était réalisable sur d'autres séries de publications que celles du new York Times.

Les collections réunies par cet agrégateur de base de données s'élèvent, d'après son interface, à plus de six milliards de pages numérisées, dont vingt millions de pages de journaux (donc proche des 16 millions proposés par la bibliothèque du Congrès), comprenant à la fois des pages numérisées pour les journaux historiques mais aussi des articles de publications récentes (de plus de trois mois à la date

de la consultation). La recherche effectuée via Proquest nous permet d'interroger simultanément 122 bases de données sur une plage de recherche démarrant au 25 mars 1911. En ne demandant que les résultats liés aux journaux et périodiques dont sont exclues les annonces et les publicités, le nombre d'articles à priori correspondant à l'évènement est conséquent.

*Figure 7. Tableau recensant le nombre de sources potentielles, articles ou pages de journaux, éditées depuis le 25 mars 1911 et correspondant à différentes requêtes par mots clés, disponibles sur ProQuest (décembre 2019).*

<i>Mots clés utilisés</i>	<i>Résultats bruts sans restriction</i>	<i>Résultats avec sources filtrées : quotidiens, journaux, dépêches, magazines, revues, périodiques historiques</i>	<i>Résultats avec type de document filtré : Exclusion des publicités et des petites annonces,</i>
FACTORY + FIRE	<b>2 774 761</b>	<b>2 618 715</b>	<b>2 040 713</b>
TRIANGLE + FIRE	<b>555 692</b>	<b>458 060</b>	<b>391 284</b>
TRIANGLE + FACTORY	<b>331 437</b>	<b>283 297</b>	<b>192 735</b>
TRIANGLE + FACTORY + FIRE	<b>136 974</b>	<b>105 603</b>	<b>70 842</b>
TRIANGLE + FACTORY + FIRE + 1911	<b>24 188</b>	<b>11 649</b>	<b>9 192</b>
TRIANGLE + FACTORY + FIRE + WAIST	<b>11 890</b>	<b>5 237</b>	<b>3 747</b>

Ce programme nous autorise donc une recherche simultanée sur une centaine de publications éditées entre 1911 et aujourd'hui, soit plus de 70 000 articles ou pages correspondant aux mots clés déjà utilisés pour la requête du New York Times. Environ un tiers seulement de ces publications proposent des durées d'édition après 1911 et des disponibilités supérieures à celles de *Chronicling America*, le reste de la base de données étant constitué soit par des journaux historiques antérieurs à 1911 ou par des publications récentes accessibles uniquement pour les numéros des vingt ou trente dernières années.

Figure 8. Tableau recensant par lieu de publication les titres de presse numérisés accessibles sur Proquest proposant une période d'édition étendue (décembre 2019).

	VILLE	ETAT	TITRE de la publication	période disponible			VILLE	ETAT	TITRE de la publication	période disponible	
				début	fin					début	fin
1	LOS ANGELES	CA	LOS ANGELES TIMES	1881	RECENT	16	NYACK	NY	JOURNAL-NEWS	1889	1990
2	HARTFORD	CO	HARTFORD COURANT	1764	RECENT	17	NEW YORK	NY	NEW YORK TIMES	1857	RECENT
3	WASHINGTON	DC	WASHINGTON POST	1877	CURRENT	18	NEW YORK	NY	NEW YORK TRIBUNE	1842	1962
4	ATLANTA	GE	ATLANTA CONSTITUTION	1946	1984	19	LONG ISLAND	NY	NEWSDAY	1940	1991
5	CHICAGO	IL	CHICAGO TRIBUNE	1872	RECENT	20	ALLENTOWN	PE	MORNING CALL	1939	RECENT
6	PLAINFIELD	IL	PLAINFIELD COURIER NEWS	1894	1962	21	PHILADELPHIA	PE	JEWISH EXPONENT	1897	present
7	BALTIMORE	MD	SUN BALTIMORE	1837	1994	22	PHILADELPHIA	PE	PHILADELPHIA INQUIRER	1860	2001
8	BOSTON	MA	BOSTON GLOBE	1872	2019	23	PHILADELPHIA	PE	PHILADELPHIA TRIBUNE	1912	2001
9	BOSTON	MA	CHRISTIAN SCIENCE MONITOR	1908	CURRENT	24	PITTSBURGH	PE	PITTSBURGH POST GAZETTE	1927	RECENT
10	BOSTON	MA	JEWISH ADVOCATE BOSTON	1909	2018	25	PITTSBURGH	PE	PITTSBURGH PRESS	1887	1992
11	ASBURY PARK	NJ	ASBURY PARK PRESS	1905	2011	26	AUSTIN	TE	AUSTIN STATESMAN	1923	1971
12	BINGHAMTON	NY	BINGHAMTON	1905	1960	27	BURLINGTON	VE	BURLINGTON FREE PRESS	1885	2007
13	NEW YORK	NY	NEW YORK AMSTERDAM NEWS	1922	recent	28		CA-NADA	GLOBE AND MAIL	1936	CURRENT
14	NEW YORK	NY	DAILY NEWS	1920	2009	29		CA-NADA	TORONTO DAILY STAR	1900	RECENT
15	ROCHESTER	NY	DEMOCRAT AND CHRONICLE	1884	RECENT	30		IRELAND	IRISH TIMES	1921	CURRENT
16	NEW YORK	NY	FORWARD	1897	RECENT	31		ISRAEL	JERUSALEM POST	1950	1988
17	ITHACA	NY	ITHACA JOURNAL	1914	2012	32		UK	GUARDIAN	1959	2003

Si l'on affine encore ces résultats aux seuls titres dont la période de publication inclut l'année 1911 et se prolonge sur le siècle, nous obtenons 18 titres de presse, soit un corpus de 926 articles, pouvant soutenir la comparaison avec le New York Times, qui peuvent nous permettre d'étudier la circulation de l'évènement sur le siècle, au sein d'une partie du territoire américain (11 états représentés) comme au sein de l'état ou de la ville de New York uniquement, mais aussi au sein de publications communautaires notamment trois issues de la communauté juive.

*Figure 9. Tableau recensant par titre de publication les articles ou pages de presse numérisés accessibles sur Proquest effectivement pertinents (décembre 2019)*

VILLE	ETAT	TITRE de la publication	période disponible		POTENTIEL	nombre d'articles pertinents publiés à partir du 25 mars 1911
			début	fin		
LOS ANGELES	CA	LOS ANGELES TIMES	1881	RECENT	664	<b>70</b>
HARTFORD	CO	HARTFORD COURANT	1887	RECENT	531	<b>77</b>
WASHINGTON	DC	WASHINGTON POST	1877	CURRENT	1432	<b>65</b>
CHICAGO	IL	CHICAGO TRIBUNE	1872	RECENT	564	<b>61</b>
BALTIMORE	MD	SUN BALTIMORE	1837	RECENT	317	<b>44</b>
BOSTON	MA	DAILY GLOBE	1908	CURRENT	283	<b>40</b>
BOSTON	MA	CHRISTIAN SCIENCE MONITOR	1908	CURRENT	107	<b>31</b>
BOSTON	MA	JEWISH ADVOCATE BOSTON	1909	2018	65	<b>14</b>
ASBURY PARK	NJ	ASBURY PARK PRESS	1905	2011	2500	<b>51</b>
ROCHESTER	NY	DEMOCRAT AND CHRONICLE	1884	RECENT	4984	<b>76</b>
NEW YORK	NY	FORWARD	1994	RECENT	71	<b>71</b>
NEW YORK	NY	NEW YORK TRIBUNE	1842	1962	323	<b>160</b>
PHILADELPHIA	PE	JEWISH EXPONENT	1897	PRESENT	38	<b>24</b>
PHILADELPHIA	PE	PHILADELPHIA INQUIRER	1860	2001	10 326	<b>54</b>
PITTSBURGH	PE	PITTSBURGH PRESS	1887	1992	6 370	<b>47</b>
BURLINGTON	VE	BURLINGTON FREE PRESS	1885	2007	1051	<b>26</b>
	CANADA	TORONTO DAILY STAR	1900	RECENT	668	<b>15</b>

Au final la constitution de ces corpus miroirs est donc très contrainte par l'indisponibilité, ou la difficulté d'accès, aux sources numérisées. Le processus de numérisation de la presse écrite américaine est encore très largement inachevé. Les bases de données mettent en avant des quantités importantes de documents, mais cette apparente abondance est trompeuse à plus d'un titre. Les priorités données par les programmes qui financent le processus relèvent de logiques s'accordant parfois mal à celles des chercheurs. Les collections disponibles ne nous permettent donc d'étudier la circulation de l'évènement au sein du territoire américain sur le siècle écoulé que de manière partielle. D'autre part les publications accessibles sont majoritairement des publications anglophones d'information générale qu'il faut mettre en regard aussi de publications plus spécialisées comme les publications syndicales ou communautaires. Grâce à la consultation des archives du syndicat américain ILGWU, *International Ladies Garment Workers Union*, acteur important de la mémoire de l'évènement, nous avons pu collecter au sein du Kheel Center de l'université de Cornell 142 articles de presse spécialisés, notamment les articles du magazine *Justice* édité par le syndicat ouvrier. Des corpus, enfin, limités, pour l'instant, aux articles en langue anglaise, à compléter par d'autres issus de journaux communautaires dont les publications sont encore nombreuses sur le territoire américain au début du XXème siècle ou encore par ceux issus de la presse internationale non-anglophone dans le but d'élargir la circulation de l'évènement.

### Les exploitations possibles :

La constitution de corpus exploitables pour répondre aux objectifs de l'étude a donc été particulièrement complexe et reste encore inachevée à ce stade. Son exploitation comporte elle aussi de nombreuses difficultés et ce à plusieurs niveaux. D'une part car la qualité des documents proposés est d'une grande hétérogénéité. Les moteurs de recherche donnent accès à des images plus ou moins précises des articles en fonction du document source. Pour certaines collections, les documents proposés ne sont pas issus de prise de vue directe des pages de journaux, mais utilisent les images réalisées pour la conservation sur micro-film sans que celles-ci aient été améliorées. Or la faible qualité de l'image d'un article complexifie son exploitation. Les moteurs de recherche rendent accessibles les articles sous plusieurs formats, image ou texte, or pour beaucoup le texte issu du processus d'OCRisation utilisé est d'assez mauvaise qualité<sup>xxvi</sup>, au point de compromettre le processus d'exploitation sémantique par exemple<sup>xxvii</sup>, sa date ou le processus utilisés ne sont pas précisés. Nous avons donc préféré collecter des images. Ce sont pour la majorité des publications des images de pages entières au sein desquelles il faut découper l'article cible afin de constituer un corpus uniquement basé sur des images d'articles à OCRiser selon une méthodologie uniforme pour tous.

La collecte et la structure des corpus ont été organisées au sein d'une base de données (Zotero), au sein de laquelle chaque article possède une fiche d'identité propre permettant de préciser des informations concernant sa source, sa forme, son contenu... A partir donc des articles et de leurs fiches informatives, il nous semble possible grâce à des protocoles spécifiques dont nous maîtrisons les paramétrages, tout en prenant toutes les précautions nécessaires en matière d'exploitation de ces données<sup>xxviii</sup>, d'éprouver nos différentes hypothèses de travail<sup>xxix</sup> concernant tout d'abord la circulation, ou non, de l'évènement dans l'espace à différentes échelles par rapport à son épiscentre new yorkais, sa circulation dans le temps pour ancrer l'évènement dans la mémoire collective nationale, notamment au travers des biais éditoriaux de traitement de l'évènement, sa circulation au sein des différents groupes composant la, société américaine. Les corpus de travail obtenus nous permettent également de mettre en évidence des évolutions, des spécificités ou des différences sémantiques pour qualifier l'évènement<sup>xxx</sup> au sein de chaînes de valeurs caractérisées. Enfin ils nous permettent d'appréhender les multiples dimensions de l'évènement attribuées par les titres de presse et qui lui donnent ses caractéristiques propres par association, ou exclusion.

Au final, l'accès à des titres de presse numérisés nous a donc permis de collecter un nombre d'articles important et sur une période chronologique étendue, autorisant un renouvellement des questionnements concernant l'étude du drame du Triangle Fire. Néanmoins la constitution d'un corpus d'étude opérationnel s'est révélée complexe, en totale antinomie avec l'apparente abondance de la ressource vantée par les plateformes de mise à disposition. Si la numérisation des journaux permet, depuis son ordinateur personnel, un accès à priori facile, ou du moins facilité, à un nombre toujours croissant de ressources physiquement disséminées sur plusieurs sites, l'accès concret à l'exemplaire numérisé d'un périodique sélectionné se dérobe encore souvent et quand bien même il est atteint, ce double numérique se révèle souvent difficilement exploitable et suscite à son tour de nombreuses interrogations quant à sa réalisation. Pour notre étude, l'utilisation d'archives numérisées doit donc être complétée par une consultation sur site.

---

<sup>i</sup> GREEN Nancy L., *Ready-to-wear and ready-to-work: a century of industry and immigrants in Paris and New York*, Duke University Press, Durham (1997)

<sup>ii</sup> LE ROUX Thomas (dir.), *Risques industriels : Savoirs, régulations, politiques d'assistance, fin XVIIe-début XXe siècle*, Presses universitaires de Rennes, 2016

<sup>iii</sup> KAPLAN Thomas, "Elizabeth Warren, at Washington Sq. Park Rally, Promises to Take On Corruption", *New York Times*, 16 septembre 2019; « She spoke near the site of the Triangle Factory Fire of 1911, which killed

---

146 garment workers, most of them women. The fire spurred a push to improve workplace safety, which Ms. Warren harnessed as a parallel for the far-reaching change she wants to pursue as president”, érigeant la tragédie de la Triangle comme le pire du swaetshop system.

<sup>iv</sup> BARRAUD DE LAGERIE P., “The wages of sweat: a social history prespective on the foight against sweatshops”, *Sociologie du Travail*, volume 55, supplément 1, novembre 2013, pages e1-e23

<sup>v</sup> KALIFA Dominique, REGNIER Philippe, THERENTY Marie-Ève, VAILLANT Alain [dir.], *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Nouveau Monde éditions, 2011

<sup>vi</sup> KAPLAN Richard L., "From Partisanship to Professionalism: The Transformation of the Daily Press," in KAESTLE Carl F. and RADWAV Janice A., *A History of the hook in .America, Volume Four. Print in Motion: The Expansion of Publishing and Reading in the United States, 1880-1940*, Chapel Hill UNC Press, 2009

<sup>vii</sup> RISLEY Ford, “ Politics and Partisanship”, in SLOAN David, MULLIKIN PARCELL Lisa, *American journalism: history, principles, practices*, Jefferson NC, Mc Farland, 2002

<sup>viii</sup> Exemple de la « penny press » comme le New York Sun de Benjamin Day, voir WHITBY Gary, SLOAN David, “The purposes of the press: a reinterpretation of American journalism history”, in *Annual meeting of the association for education in journalism*, 1981

<sup>ix</sup> Exemple de Pulitzer voir BLEVENS Fred, « Publishers », in SLOAN David, MULLIKIN PARCELL Lisa, *American journalism: history, principles, practices*, Jefferson NC, Mc Farland, 2002

<sup>x</sup> L’analyse de nombreux dessins de presse illustrant ces articles fait l’objet d’une étude séparée

<sup>xi</sup> LENOBLE Benoît, « Célébrations », in *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Nouveau Monde éditions, 2011

<sup>xii</sup> WRONA Adeline, « La presse en son miroir, dénonciations », in *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Nouveau Monde éditions, 2011

<sup>xiii</sup> BOUCHET Thomas, « Evènement, opinion et politique », in *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Nouveau Monde éditions, 2011

<sup>xiv</sup> ACLAND Charles R., HOYT Eric, *The Arclight guidebook to media history and the digital humanities*”, Sussex, Reframe Books, 2016

<sup>xv</sup> ROYOT Daniel, RUEL Susan, *Médias, société et culture aux Etats-Unis*, Orphys, Ploton, 1996, p 16-17

<sup>xvi</sup> FOLKERTS Jean, TEETER Dwight, “chapitre 9: Mass Markets and Mass culture”, in *Voices of a nation : a history of mass media in the US*, MacMillan, 1994

<sup>xvii</sup> DAVIS Elmer, “History of the New York Times 1851-1921”, *New York Times*, 26 septembre 1921, p 12

<sup>xviii</sup> SCHUDSON Michael, *Discovering the news, a social history of American Newspapers*, New York, Basic Books, 1978

<sup>xix</sup> COTLER Jane, Sandhaus Evan, « How to build a TimesMachine”, *New York Times*, 01 février 2016

<sup>xx</sup> Borne chronologique imposée par le lecteur virtuel, les articles plus récents sont accessibles directement via le moteur de recherche du site internet du journal

<sup>xxi</sup> WALLER Gregory A., « Search and re-search: digital print archives and the history of multi-sited cinema”, in ACLAND Charles R., HOYT Eric, *The Arclight guidebook to media history and the digital humanities*”, Sussex, Reframe Books, 2016

<sup>xxii</sup> Par exemple les listes des nécessiteux « neediest » new-yorkais qui appellent à la générosité des lecteurs du *Times* en présentant précisément le cas de chaque famille

<sup>xxiii</sup> FOLKERTS Jean, TEETER Dwight, “chapitre 9: Mass Markets and Mass culture”, in *Voices of a nation : a history of mass media in the US*, MacMillan, 1994

<sup>xxiv</sup> Une dynamique de numérisation observable des deux côtés de l’Atlantique, voir par exemple RYGIEL Philippe, « Nouvelles frontières de l’historien », *Historien à l’âge numérique*, Villeurbanne, Presses de l’enssib (nouvelle édition en ligne), 2017

<sup>xxv</sup> Un des aspects donc de cet âge de l’abondance précisé par ROSENZWEIG Roy, « Scarcity or abundance ? Preserving the past in a digital era », *American Historical Review*, 108-3, juin 2003, p. 735-762.

<sup>xxvi</sup> MUTUVI Stephen, DOUCET Antoine, ODEO Moses, JATOWT Adam, « Evaluating the Impact of OCR Errors on Topic Modeling”, *Lecture Notes in Computer Science*, vol 11279, Springer, Cham, 2018, pp. 3-14

<sup>xxvii</sup> HAMDIA A., JEAN-CAURANT A., SIDERE N., COUSTATY M., and DOUCET A., “An Analysis of the Performance of Named Entity Recognition over OCRed Documents”, in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2019, pages 333–334

<sup>xxviii</sup> CARBOU, Guillaume. « Analyser les textes à l’ère des humanités numériques. Quelques questions pour l’analyse statistique des données textuelles », *Les Cahiers du numérique*, vol. vol. 13, no. 3, 2017, pp. 91-114.

<sup>xxix</sup> BARATS Christine, LEBLANC Jean-Marc, et FIALA Pierre, « Chapitre 5 - Approches textométriques du web : corpus et outils », *Manuel d’analyse du web en Sciences Humaines et Sociales*. sous la direction de BARATS Christine, Armand Colin, 2013, pp. 99-124.

---

<sup>xxx</sup> CHÂTEAURAYNAUD Francis, DEBAZ Josquin, « Prodiges et vertiges de la lexicométrie », *Socio-informatique et argumentation*, 23 décembre 2010