Mémoires du Covid-19 et archives du Web. Preuve de concept pour une analyse quantitative des données du dépôt légal du web de la BnF

Roch Delannay, Marta Severo, Louis Gabrysiak

Humanités numériques

9 | 2024
Varia

Mémoires du Covid-19 et archives du Web : pour une analyse quantitative du dépôt légal de la BNF

Covid-19 Memories and Web Archives: Towards Quantitative Analysis of Legal Deposit at the BNF

Roch Delannay, Marta Severo et Louis Gabrysiak

Édition électronique
URL : https://journals.openedition.org/revuehn/3955
DOI : 10.4000/11wmx
ISSN : 2736-2337

Éditeur
Humanistica

Référence électronique
Roch Delannay, Marta Severo et Louis Gabrysiak, « Mémoires du Covid-19 et archives du Web : pour une analyse quantitative du dépôt légal de la BNF », *Humanités numériques* [En ligne], 9 | 2024, mis en ligne le 01 juin 2024, consulté le 01 juillet 2024. URL : http://journals.openedition.org/revuehn/3955 ; DOI : https://doi.org/10.4000/11wmx

humanités
numériques

9 | 2024
Varia

Mémoires du Covid-19 et archives du Web : pour une analyse quantitative du dépôt légal de la BNF

Covid-19 Memories and Web Archives: Towards Quantitative Analysis of Legal Deposit at the BNF

Roch Delannay, Marta Severo et Louis Gabrysiak

Résumés

Une impressionnante quantité de traces numériques ont été produites lors de la crise sanitaire du Covid-19, dont certaines ont fait l'objet de collectes institutionnelles. Cet article vise à étudier comment les archives du Web jouent un rôle fondamental dans la construction de la mémoire collective. D'un point de vue thématique, notre travail s'efforce de comprendre comment ces archives offrent une opportunité exceptionnelle d'étudier les phénomènes liés au Covid-19. Nous nous concentrerons en particulier sur les archives du Web du dépôt légal de la Bibliothèque nationale de France (BNF) que nous avons pu explorer dans le cadre du projet *Web-mémoires*, développé en collaboration avec le DataLab de la BNF et soutenu par le Labex *Les passés dans le présent*. D'un point de vue méthodologique, cet article traite de la problématique de l'analyse de ces archives du Web. En particulier, nous visons à proposer un *workflow* inédit pour l'analyse quantitative des archives du Web de la BNF à partir d'une thématique ciblée qui prend en compte les contraintes techniques et juridiques liées à ce type de données et à leur consultation dans le cadre du dépôt légal du Web.

An impressive quantity of digital traces was produced during the Covid-19 health crisis, some of which were the subject of institutional collections. The aim of this article is to explore how Web archives play a fundamental role in the construction of collective memory. From a thematic point of view, our work strives to understand how these archives offer an exceptional opportunity to study phenomena linked to Covid-19. We will

focus in particular on the BNF's legal deposit web archives, which we were able to explore as part of the *Web-mémoires* project, developed in collaboration with the BNF DataLab and supported by the Labex *Les passés dans le présent*. From a methodological point of view, this article confronts the problem of analysing these Web archives. In particular, we aim to propose a novel workflow for the quantitative analysis of the BNF Web archives, based on a targeted theme that takes into account the technical and legal constraints associated with this type of data and its consultation within the framework of the Web legal deposit.

Entrées d'index

MOTS-CLÉS : sciences de l'information et de la communication, Web, archives, chaîne de traitement, analyse textuelle, collecte de données

KEYWORDS: information and communication sciences, Web, archives, processing chain, textual analysis, data acquisition

Introduction

1 Au printemps 2020, le déclenchement de la crise sanitaire et l'expérience du confinement qui s'est ensuivie ont entraîné la mise en œuvre massive et généralisée de collectes de traces de cette période décrite d'emblée comme « historique » (Adams et Kopelman 2022). Cette démarche collective a engendré deux types d'initiatives distinctes. Tout d'abord, en accord avec sa mission, la Bibliothèque nationale de France (BNF) et son réseau de correspondants régionaux ont lancé des efforts de collecte d'activités numériques (Benoist *et al.* 2020). Cette collecte du Web, incluant sites Internet, réseaux sociaux et vidéos en ligne, s'est intensifiée pour conserver les réactions et les modes de vie de la société française face à la pandémie et au confinement. Simultanément, l'Institut national de l'audiovisuel (INA) a également entrepris une collecte ciblée, archivant une vaste quantité de documents audiovisuels, de tweets et de vidéos liés à la pandémie (Schafer 2021 ; Hervé 2022).

2 La crise sanitaire a également généré un autre type d'archives, géré par des acteurs institutionnels, principalement des centres d'archives municipales ou départementales. Dans ce contexte, les traces éphémères sont devenues des archives, souvent sous forme de récits d'expériences vécues. Ces récits ont été largement produits en réponse à des « appels à témoignages » lancés par les institutions culturelles elles-mêmes. Ces appels ont été diffusés par des sites Internet et les réseaux sociaux, suscitant une réponse significative de la part des citoyens. Des entreprises, journalistes, chercheurs, associations et regroupements informels ont également participé, créant un éventail varié de contenus, allant des objets personnels aux photographies de l'espace public, en passant par des œuvres d'art, des dessins, des poèmes et bien d'autres formes d'expression (Ducas, De Angelis et Cormier 2022 ; Gensburger et Severo 2021).

3 Ces deux groupes d'initiatives, bien qu'ayant des origines différentes, ont contribué à la construction d'archives numériques riches et diversifiées, capturant les multiples facettes de la réaction de la société à la crise sanitaire. Les archives du Web de la BNF et de l'INA, en particulier, sont devenues des trésors d'informations permettant de reconstituer l'évolution de la pandémie, d'étudier le langage associé au Covid-19, les registres de discours mobilisés et leurs réceptions.

4 Face à ces initiatives de collecte, cet article vise à étudier comment les archives du Web, en dépit de leur nature immatérielle, jouent un rôle fondamental dans la construction de la mémoire collective. D'un point de vue thématique, notre travail s'efforce de comprendre comment ces archives, collectées dans un présent jugé historique, offrent une opportunité exceptionnelle d'étudier les phénomènes liés au Covid-19. Nous nous concentrerons notamment sur les archives du Web du dépôt légal de la BNF que nous avons pu explorer dans le cadre du projet *Web-mémoires* [1]. D'un point de vue méthodologique, cet article traite de la problématique de l'analyse de ces archives du Web. En particulier, nous visons à proposer un *workflow* inédit pour l'analyse quantitative des archives du Web de la BNF à partir d'une thématique ciblée qui prend en compte les contraintes techniques et juridiques liées à ce type de données et à leur consultation dans le cadre du dépôt légal du Web.

5 Si les archives du Web sont une source de plus en plus exploitée et considérée comme légitime dans des études historiques, sociologiques et infocommunicationnelles, les travaux existants s'appuient principalement sur des méthodes qualitatives d'exploration, comme des techniques ethnographiques, sémiotiques ou d'analyse de discours (Gebeil 2017 ; Schafer et Winters 2021). Cela n'est pas toujours un choix assumé du chercheur mais c'est plutôt une conséquence des contraintes de consultation du dépôt légal des archives du Web. Le récent cadre des *datalabs* (laboratoires de données) qui visent à créer des conditions plus favorables de consultation de ces données pour des objectifs de recherche ouvre aujourd'hui de nouvelles perspectives d'accès à ces données (Segault et Severo 2023). Dans ce texte, nous restituons une expérience pilote d'analyse quantitative d'un corpus tiré des archives du Web de la BNF que nous avons pu réaliser grâce à la collaboration construite avec le BNF DataLab. Ce type de collaboration étant encore rare, le focus de notre analyse portera principalement sur le protocole mis en place et les choix techniques adoptés. L'exploration thématique proposée ne prétend pas à être exhaustive mais veut être une preuve de concept visant à faciliter de travaux futurs dans ce domaine.

Les archives Web de la BNF face au Covid-19

6 Les dépôts légaux du Web de la BNF et de l'INA ont été construits dans le cadre légal de la loi relative au droit d'auteur et aux droits voisins dans la société de l'information, dite loi DADVSI, de 2006 suivi du décret d'application de 2011. Celle-ci élargit le champ du dépôt légal, en disposant que « sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique ». L'INA est chargée d'étendre sa conservation de l'audiovisuel français en dehors de ses supports traditionnels et en suivant ses mutations (sites de chaînes, Web TV…). La BNF, elle, doit s'occuper du reste de l'Internet français (extension des noms de domaine en .fr ou extensions régionales, sites créés en France, presse, etc.). Dès le départ, l'INA a un périmètre plus défini, plus balisé. Celui de la BNF se définit davantage en négatif : tout ce que ne couvre pas l'INA (Bachimont 2023).

7 Il existe une différence majeure entre le dépôt légal du Web et le dépôt légal traditionnel : il n'y a aucune démarche active de la part des éditeurs et fournisseurs de contenus, la collecte se fait sans eux. La loi dispose que tout contenu public sur Internet doit être conservé mais, dans les faits, c'est impossible. Pour les documentalistes qui s'occupent de l'archivage, il s'agit non pas simplement de « lancer » un robot, mais bien d'effectuer en amont un travail de sélection, de mener une réflexion sur ce qui peut et doit être collecté.

8 La tâche d'archiver Internet revient donc, en France, à deux institutions patrimoniales, préexistantes à cette mission. Toutes deux font partie de l'International Internet Preservation Consortium (IIPC), organisation internationale fondée par la bibliothèque du Congrès et diverses bibliothèques nationales. Leur action s'inscrit dans un mouvement global de conservation des archives d'Internet, bien qu'elles disposent toujours de leur autonomie. L'objectif de l'IIPC, dès sa création en 2004, est de favoriser la collaboration internationale. Elle a notamment, pour ce faire, mené une politique de normalisation des formats : le WARC à la fin des années 2000, aujourd'hui largement utilisé (Maemura 2023).

9 À la BNF, le dépôt légal du Web est assuré par un département créé pour l'occasion comptant cinq à six membres. L'équipe du dépôt légal est renforcée par certains membres de la direction des systèmes d'information, qui jouent, depuis le début des actions, un rôle primordial dans la constitution, la diffusion et la valorisation des archives. Du point de vue technique, la BNF pratique un archivage qui permet la restitution la plus « fidèle » possible du matériau original. L'utilisateur est mis face à une page Web telle qu'elle était au moment de son archivage, avec ses images, son organisation… L'ambition est de conserver l'éditorialisation des pages affichées. Du côté de l'archivage, il s'agit d'une approche par URL. À partir d'une URL, on va déterminer une « profondeur » de « crawl » pour le robot et l'archivage. Par exemple, + 1 clic. Cela signifie que seront archivées toutes les pages auxquelles on aura eu accès à partir de la page source en cliquant une fois sur chaque lien. Bien que l'archivage parte d'une URL, les différents éléments d'une page sont archivés séparément (la page est « découpée », en images, blocs de texte, etc.). Ce qui explique que, lors de la vi-

sualisation, il manque parfois des éléments sur une page, remplacés par des bandeaux BNF. La même page est très souvent archivée plusieurs fois, à des dates et heures différentes. Lors de la visualisation d'une URL, la date et l'heure de l'enregistrement sont systématiquement présentes. Étant donné que la logique d'archivage est celle de l'URL, il est fréquent, en naviguant sur différentes pages d'un même site [2], d'effectuer des sauts temporels sans nécessairement en avoir conscience si l'on n'y est pas attentif.

10 La BNF conduit plusieurs types de campagnes de collecte du Web. La « collecte large » a lieu tous les ans, autour d'octobre-novembre. Il s'agit de collecter l'ensemble du « Web français » (les sites en .fr, les sites régionaux, les sites créés en France…), avec une faible profondeur (page + 1 clic). Cette collecte a couvert 5,8 millions de sites en 2022. Elle est assurée par le département du dépôt légal du Web (DLWeb). À ces collectes larges annuelles, il faut ajouter des collectes ciblées, de plusieurs types : des collectes « courantes », dont les sélections sont opérées par les différents départements (disciplinaires) de la BNF et des collectes « projets », sur des thématiques précises. Il existe également une procédure de collecte « d'urgence », qu'il est possible de déclencher pour collecter rapidement des sites amenés à disparaître à court terme. Elle a pour objectif de pallier les faiblesses de la collecte large.

11 Ce type de démarche a été mise en place pour la collecte liée au Covid-19 à partir du mois de janvier 2020. D'abord par l'archivage des hashtags « #jenesuispasunvirus » et « #coronavirusenfrance » ainsi qu'une page sur le site du Mouvement contre le racisme et pour l'amitié entre les peuples, titrée « Un virus n'a pas d'origine ethnique [3] » (Faye 2020). Rapidement, la collecte est élargie grâce à la participation de l'ensemble des correspondants internes et externes volontaires. Selon la documentation [4], cette collecte exceptionnelle appelée « collecte de la crise sanitaire », mentionne « 3 260 sélections pour la catégorie "sites et pages Web" », « 1 329 sélections pour la catégorie "Websocial" » et « 249 chaînes vidéos (YouTube) », soit un total de 4 838 sélections.

12 La première étape de notre travail a consisté dans la construction d'un corpus pertinent dans le cadre des archives du Web de la BNF, notamment de cette collecte spéciale liée au Covid-19. Les archives du Web permettent d'adopter une approche diachronique, par l'archivage des mêmes pages et sites, de multiples fois au fil du temps. Nous nous sommes fixé un bornage temporel d'un an, du 1er mars 2020 au 28 février 2021, avec comme ambition d'étudier l'évolution du contenu de différents sites, particulièrement du langage employé dans un sous-corpus de sites recueillant des témoignages à travers la réalisation d'analyses textuelles.

13 Sur la base d'une requête sur les mots « mémoire », « journal » (de confinement, littéraire, d'écrivain…), « témoignages » et « archives », une réduction du vaste corpus de la BNF permet d'aboutir à une liste de 425 sites et pages Web [5]. Ces différents sites et pages ont été « crawlés » un nombre très inégal de fois, de 1 à 480 fois, sur des périodes temporelles plus ou moins étendues. Les entrées du corpus le plus souvent collectées sont les pages de réseaux sociaux, les différents hashtags sur Twitter sélectionnés ainsi que des sites dédiés spécifiquement à la pandémie [6]. À l'inverse, les pages des sites de presse ne sont généralement « crawlées » qu'une seule

fois. La page la plus « crawlée » est celle du hashtag « #JournaldeConfinement » sur Twitter [7]. Non seulement celle-ci l'a été 480 fois, mais elle l'a été sans interruption à partir du 18 mars 2020 jusqu'au moins fin février 2021.

Méthodologie

14 Notre objectif était de réaliser une analyse de contenu d'un échantillon de pages de notre corpus à travers des méthodes quantitatives de fouille de texte. Les développements de ces analyses ont une double vocation : étudier l'évolution des pratiques mémorielles numériques liées au Covid-19 et servir de protocole expérimental pour tester des pistes d'exploration des archives du Web de la BNF. En ce sens, nous décrirons les résultats obtenus en réponse à nos questions de recherche et nous pointerons les contraintes techniques et, par conséquent, les difficultés rencontrées tout au long du processus de traitement des données.

15 La question de recherche qui nous anime est relative à l'évolution du langage et des thématiques abordées par les internautes dans les témoignages disponibles dans la collection des archives du Web de la BNF dédiée à la pandémie de Covid-19. Dans ce contexte, nous pouvons formuler une double hypothèse : (i) l'évolution de la pandémie avec ses différentes phases (confinement, déconfinement, port de masque, etc.) a provoqué une évolution du langage caractérisée par l'émergence de nouveaux termes et la disparition d'autres, et par un changement de ton selon la période ; (ii) l'évolution des actualités qui ont animé la période de la pandémie a eu un impact sur les thèmes abordés dans les témoignages archivés.

16 Pour tester cette hypothèse, il nous a fallu d'abord définir un sous-corpus adapté à l'analyse de contenu. Pour ce faire, nous avons identifié un échantillon comprenant :

- Des sites Web destinés à l'expression individuelle ou collective des internautes (par exemple, sites de témoignages, blogs, collectes en ligne, etc.). Tous les sites Web institutionnels et journalistiques ont été retirés de ce sous-corpus.
- Des sites qui sont principalement constitués de texte. Tous les sites ne comportant que peu de textes et beaucoup d'images (celles-ci sont difficilement traitables dans les archives de la BNF) ont été exclus.
- Des sites qui ont été « crawlés » cinq fois au minimum dans la période du 1er mars 2020 au 28 février 2021 et peuvent permettre une analyse diachronique.

17 La taille du corpus obtenu après application de ces critères s'élève à une dizaine de sites Web.

18 La plage temporelle sélectionnée pour cette étude correspond à la première année de la pandémie de Covid-19, pendant laquelle tous les événements sont nouveaux et se bousculent : confinements, déconfinements, port du masque obligatoire, etc. Nous avons découpé cette année en cinq périodes distinctes :

- 1er mars au 11 mai 2020 (période du premier confinement, légèrement étendue de quelques jours en amont de celui-ci)
- 12 mai au 19 juillet 2020 (période de déconfinement avant l'obligation du port de masque)

- 20 juillet au 15 décembre 2020 (de l'obligation du port du masque jusqu'à la fin du deuxième confinement)
- 16 décembre 2020 au 28 février 2021 (période de contrôle après le deuxième confinement)
- 1er mars 2021 au 31 juillet 2023 (période de vérification dans les sites Web de présence ou non de commémorations de cette première année du Covid)

Tableau 1. Nombre d'occurrences de « crawls » par période d'analyse

Nom du site	Mars à mai 2020	Mai à juil. 2020	Juil. à déc. 2020	Déc. 2020 à févr. 2021	Mars 2021 à juil. 2023
Génération Covid	1	0	1	1	10
Reinfo Covid	0	0	1	2	17
Par ma fenêtre	0	2	6	2	2
Le Jour d'après	2	2	2	0	6
Inventons le monde d'après	0	2	1	0	2
C'est la lutte virale	1	2	1	0	2
Notre nouvelle vie	0	2	1	0	0
Corona Maison	0	3	3	0	2

19 La première difficulté rencontrée est celle du peu d'occurrences de « crawls » de certains sites sur les plages qui nous intéressent (tableau 1). Par exemple, pour la première période, seulement trois sites Web sont concernés par notre analyse, car les autres n'apparaissent pas encore dans la collecte. Par ailleurs, comme nous le détaillerons dans le paragraphe suivant, l'analyse de chaque site demande la mise en place d'un *workflow* spécifique. Cela nous a obligés à limiter ultérieurement notre corpus à deux sites Web : *Génération Covid* et *Le Jour d'après*. Ces sites ont été sélectionnés parce qu'ils avaient le plus d'occurrences et une structure de page principalement basée sur des entrées textuelles. Il peut sembler au lecteur qu'un corpus constitué de seulement deux sites ne soit pas représentatif du Web mémoriel du Covid-19 mais il n'en a pas la prétention. En revanche, ces deux sites sont les seuls qui permettent une analyse textuelle diachronique dans le dépôt légal de la BNF.

20 Ces sites sont encore présents dans le Web vivant et peuvent être retrouvés à leur URL respective :

- *Génération Covid* : http://generationcovid.fr
- *Le Jour d'après* : https://lejourdapres.parlement-ouvert.fr

21 Le site *Génération Covid*, porté par la société Clear Prod, société spécialisée dans le développement de sites Web, a pour vocation de recueillir des témoignages d'expériences vécues pendant la pandémie. Il s'agit de donner un espace de prise de parole aux internautes avec pour objectif de créer une mémoire citoyenne de cette période.

22 Le site *Le Jour d'après* diffère du premier. Plutôt que de collecter des traces d'expériences vécues à la première personne, il se présente comme un espace de discussions et d'échanges sous la forme de parlement ouvert, un peu comme un forum, pour repenser nos sociétés après la pandémie. Le site collecte des idées, témoignages et échanges entre les participants

organisés selon des thèmes : santé, travail, etc. Les consultations ouvertes ne sont souvent pas en lien direct avec le Covid-19. Le projet est porté par des « parlementaires de différentes sensibilités politiques » et appelle « les forces vives de notre pays et les citoyennes et citoyens » à contribuer au projet.

Entre contraintes et défis techniques

23 Les archives du Web de la BNF sont accessibles uniquement dans les salles de recherche sur les différents sites de la BNF, ainsi que dans les établissements partenaires en région. Ils ne sont pas accessibles en ligne ni interrogeables par API pour garantir le respect du droit d'auteur du dépôt légal du Web. Un usager qui consulte les archives dans les salles de la BNF aura accès en lecture seule à un nombre limité d'informations. À travers un moteur de recherche, il pourra lancer des requêtes qui lui permettront d'obtenir une liste (non ordonnée) des sites répondant à sa demande. Ensuite, il pourra naviguer dans chaque site un par un en visualisant seulement les pages qui ont été capturées pour la période interrogée.

24 Pour obtenir des informations plus précises sur un corpus d'archives du Web (par exemple, pour savoir quel type de « crawl » a été effectué sur telle page) ainsi que des extractions des données, il faut s'adresser aux personnels du BNF DataLab qui peuvent accéder aux archives avec d'autres outils que ceux mis à disposition des utilisateurs et fournir des corpus de pages adaptées à l'analyse quantitative. Dans notre cas, nous avons pu obtenir pour les deux sites de notre sous-corpus des documents aux formats HTML et TXT correspondant à toutes les captures de la première période du 1er mars ou 11 mai 2020. Dans le respect du droit d'auteur, ces documents ont été déposés sur le disque d'une machine virtuelle (sous le système d'exploitation Ubuntu 22.04) uniquement accessible dans les locaux du BNF DataLab (site de Tolbiac).

25 Ces fichiers ne sont pas des exactes reproductions des sites Web tels qu'ils ont existé en ligne : il s'agit d'une reconstruction page à page selon le niveau de profondeur auquel accède le robot. Des pages ou éléments d'une page à l'intérieur des sites Web peuvent être manquants s'ils sont à une profondeur à laquelle le robot ne descend pas ou s'ils sont hébergés sous un autre nom de domaine (par exemple, pour certaines images). Un autre paramètre est à prendre en compte lors de l'analyse. Du fait de l'automatisation du processus lors de la reconstruction de chacun des sites, les URL initiales de chacune des pages sont modifiées par le robot. Le processus d'archivage des sites diffère de celui d'objets plus traditionnels puisque les archives qui en résultent sont des reproductions du Web vivant les plus fidèles possibles, mais qui n'ont jamais existé pour autant.

26 Les documents donnés, au format TXT et HTML, contiennent tous les deux du texte brut. La différence fondamentale est que le TXT étant un format sans aucune structuration ou hiérarchisation des contenus, tout le texte est « à plat », alors que pour le format HTML, les différents éléments textuels sont encapsulés dans des balises explicitant formellement leur niveau de structuration. Toutefois, contrairement à un autre format plus rigoureux (et verbeux) comme le XML, le HTML n'impose pas de structuration précise et rigoureuse au-delà des spécifications du World Wide Web

Consortium (W3C). Il en résulte des pratiques très différentes d'emploi de cette technologie dont certaines s'éloignent largement des « bonnes pratiques » recommandées par les communautés du Web. Ce qui est à la fois un avantage, car cela offre beaucoup de possibilités de structuration des contenus (choix personnels), et un inconvénient lorsqu'il s'agit de traiter massivement des corpus de pages Web. Les fichiers TXT ne contiennent plus les balises HTML, toutefois ils contiennent tout ce qui se trouve entre ces balises : tous les boutons, les liens, etc. Ces fichiers sont pratiques pour avoir une idée grossière de ce qui se trouve dans le corpus, mais ne seront pas utilisés pour l'obtention des résultats souhaités, car ils comportent trop de « bruit » qu'il n'est plus possible de nettoyer automatiquement à cause de l'absence des balises.

27 Au contraire, dans les documents HTML, du fait de l'architecture interne des éléments textuels, il est possible de naviguer à l'intérieur des documents et d'en diviser les contenus avec des requêtes basées sur les balises HTML : il s'agit de l'action de « parser » les documents. Or, comme nous venons de le mentionner, les pratiques d'écriture du format HTML sont plutôt libres et il est difficilement possible d'appliquer exactement la même requête d'extraction des contenus pour différents sites Web. Il s'agit là d'une difficulté majeure de l'analyse de texte. Afin d'obtenir des résultats fiables, l'une des opérations les plus importantes consiste à nettoyer les données et à supprimer le bruit à travers cette opération de « parsing » qui devra être réalisée de manière personnalisée pour chaque page.

28 Pour nettoyer et « parser » le corpus, nous avons utilisé le langage de programmation Python et la librairie Beautiful Soup.

```
find_date = soup.find("div", class_= "author-data__extra")
date_element = find_date.find("span").text.strip() if find_date else
"no date"
```

29 Cet exemple de code est tiré du script permettant de « parser » les contenus du site *Le Jour d'après*. L'élément que nous cherchons à isoler est la date qui se trouve encapsulée dans une balise à l'intérieur d'une balise ayant pour classe author-data__extra. De plus, toutes les entrées du corpus n'ayant pas forcément une date qui leur est associée, nous avons ajouté une condition pour indiquer si les contributions contiennent cette information ou non. Cette organisation des informations dans le document HTML est spécifique à ce site Web, nous ne pouvons pas la réutiliser pour « parser » les autres sites.

30 La machine virtuelle où ces documents étaient disponibles comporte deux modes différents : un mode avec un accès à Internet dans lequel les données ne sont pas disponibles, mais dans lequel on peut configurer la machine virtuelle ; et un mode sans accès à Internet avec un disque supplémentaire monté sur la machine virtuelle (/Documents/espace-BnF/) qui contient les archives du Web extraites par la BNF. Au démarrage du traitement des données, il faut donc jouer avec ces deux modes pour configurer et tester les différents traitements à appliquer aux données.

31 Une subtilité réside dans ce changement de mode. La bascule d'un mode à l'autre affecte le fonctionnement nominalement attendu (par habitude) d'un ordinateur. Lorsque la machine passe en mode sans Internet, une capture (*snapshot*) de la machine avec Internet est effectuée et c'est sur

cette capture que l'utilisateur retournera lorsqu'il repassera en mode Internet. En conséquence, dans le mode sans Internet, les écritures numériques inscrites en dehors de l'espace-BnF ne seront pas sauvegardées.

32 Lors de la configuration de la machine virtuelle, nous avons rencontré des difficultés pour installer ou configurer les logiciels. Nous nous sommes tournés vers des solutions libres et *open source* comme le logiciel Orange Data Mining [8] pour lequel la condition hors connexion a généré des erreurs. Une des étapes fondamentales pour la fouille de texte est le prétraitement du corpus. Il s'agit, comme nous le verrons plus loin, de l'étape de préparation des données nécessaire à l'application des différents algorithmes de traitement du corpus. Parmi toutes les étapes de prétraitement dans le logiciel, la normalisation des données (racinisation ou lemmatisation) fait des appels récurrents à Internet. Or, comme la machine virtuelle utilisée est déconnectée d'Internet, le logiciel devient inaccessible tant que les requêtes ne sont pas résolues. Des « freezes [9] » du logiciel ont persisté malgré la mise en place d'un proxy pour rediriger les requêtes vers une erreur 404.

La mémoire du Covid-19 dans les archives du Web de la BNF

33 En suivant notre double objectif, les résultats de l'analyse seront détaillés en deux parties. Nous décrirons d'abord la méthodologie développée pour faire de l'analyse de contenu dans les archives du Web de la BNF. Dans un deuxième temps, nous présenterons les résultats concernant les représentations mémorielles présentes dans l'archive et l'évolution du langage qui leur est associée. Comme nous l'avons déjà observé, cette analyse ne veut pas être une étude exhaustive, mais vise à proposer une preuve de concept.

Méthodologie de traitement des données : preuve de concept

34 La sélection de deux sites Web permet de réaliser une preuve de concept de fouille de texte des pages contenues dans les archives du Web de la BNF. Comme cela est mentionné précédemment, pour pouvoir traiter les documents, il nous fallait d'abord récupérer les contenus textuels dans chacune des pages.

35 Voici un autre exemple de code :

```
content_element = contribution_element.find("div", class_= "testimo-
nial-content").text.strip()
```

36 Cet extrait de code montre le ciblage des contenus textuels dans les balises avec une classe spécifique *testimonial-content*. De cette manière, seul le texte contenu dans cette balise ayant pour propriété la classe *testimonial-content* sera retourné par la fonction. En revanche, ce ciblage n'est valable que pour une partie du texte du site *Génération Covid*. D'autres scripts sont nécessaires pour les autres contenus et pour les autres sites Web.

37 L'étape suivante, une fois que l'ensemble des éléments à analyser sont ciblés, consiste à isoler ces derniers du reste du bruit. Le texte a été récolté dans un fichier au format CSV. Chaque témoignage fait l'objet d'une entrée dans le document, comportant la date de publication et le texte du témoignage lui-même (figures 1 et 2).

Figure 1. Un témoignage du site Web *Génération Covid*

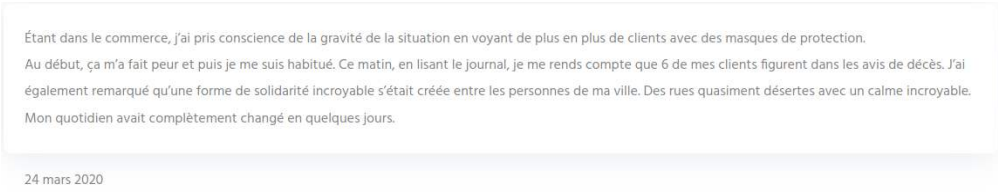

Image produite par les auteurs

Figure 2. Le même témoignage collecté dans le fichier au format CSV

24/03/2020	Étant dans le commerce, j'ai pris conscience de la gravité de la situation en voyant de plus en plus de clients avec des masques de protection. Au début, ça m'a fait peur et puis je me suis habitué. Ce matin, en lisant le journal, je me rends compte que 6 de mes clients figurent dans les avis de décès. J'ai également remarqué qu'une forme de solidarité incroyable s'était créée entre les personnes de ma ville. Des rues quasiment désertes avec un calme incroyable. Mon quotidien avait complètement changé en quelques jours.

Image produite par les auteurs

38 Après ces différentes étapes de construction et de prétraitement du corpus, les données textuelles peuvent être analysées par des algorithmes de fouille de texte avec le logiciel Orange Data Mining.

39 L'étude du corpus est réalisée grâce à deux méthodes intégrées dans le *workflow* d'Orange Data Mining. La première méthode permet de compter les occurrences des mots présents dans le corpus. Pour cela, il faut d'abord appliquer un nouveau prétraitement aux textes afin de séparer les mots entre eux (tokénisation) puis en supprimer certains avec une liste de *stop words* (ou « mots vides »). Cette méthode seule est cependant incomplète. Un nombre d'occurrences élevé d'un même mot ne signifie pas que ce terme est le plus important du corpus : si sa présence est située sur seulement quelques documents, son importance en sera grandement diminuée. Il existe une autre méthode qui permet de compléter ces informations : le regroupement hiérarchique. Grâce à la vectorisation des mots, il devient possible de calculer la distance qui sépare chaque vecteur et de les regrouper en fonction de la distance qui les sépare : les distances les plus courtes étant assemblées entre elles. De cette manière, il est possible d'observer les représentations majoritaires d'un corpus et celles qui sont plus à la marge.

40 Ces méthodes appliquées au corpus de données extraites des archives de la BNF laissent supposer le bon fonctionnement de la chaîne de traitement. Toutefois, la quantité de données, qui au moment de l'écriture de cet article correspond seulement aux données pour la première période temporelle (du 1er mars au 11 mai 2020), ne permet pas de valider ou non notre hypothèse.

41 Le site *Le Jour d'après* comptabilise beaucoup d'entrées, 3 497 exactement, et cela uniquement pour les contributions, sans les commentaires qui ne sont donc pas présents dans les documents, ce qui constitue une limite majeure de ce corpus. Une fois le contenu des documents HTML « parsé » et réindexé dans un fichier au format CSV, il a été possible de traiter les données avec Orange Data Mining.

42 Pour cela, nous avons appliqué au corpus plusieurs réglages des paramètres : une transformation de toutes les lettres en bas-de-casse ; une tokénisation du texte en mots au moyen de l'expression régulière \w+ ; plusieurs filtres pour supprimer une liste de *stop words* prédéfinis, les nombres, la ponctuation, en conservant les 300 *tokens* (éléments) les plus fréquents ; une normalisation des *tokens* par lemmatisation avec UDPipe [10] qui utilise un modèle de données préentraîné dans ce but.

43 Sans les filtres, nous obtenons 21 612 types différents pour 356 944 *tokens* dans ce corpus. Avec les filtres, nous réduisons cette quantité à 238 types pour 44 683 *tokens*.

44 Comme première méthode d'exploration, nous nous sommes appuyés sur des nuages des mots obtenus grâce au *widget* Word Cloud. Ces représentations permettent d'observer nos 300 *tokens* les plus fréquents ainsi que le poids (nombre d'occurrences) de chaque type dans le corpus. Cependant, il est important de rappeler que les proximités entre mots sont arbitraires et ne correspondent pas à une proximité sémantique ou contextuelle.

45 Dans un deuxième temps, nous avons procédé à une analyse par *clustering* qui devrait permettre d'identifier de manière plus précise les champs sémantiques présents et leur évolution dans le temps.

46 Pour ce qui concerne *Génération Covid*, nous n'avons sur la première période qu'une vingtaine de témoignages, ce qui n'est pas un échantillon suffisant pour valider des résultats. Pour cette raison, nous avons décidé de reproduire l'analyse également sur les autres périodes en nous appuyant sur le Web vivant.

La mémoire du Covid-19 sur le site *Le Jour d'après*

47 Si nous considérons le corpus *Le Jour d'après* de la première période, nous pouvons voir que les premiers types avec le plus d'occurrences sont liés au travail et à des préoccupations économiques (figure 3). Le site *Le Jour d'après*, en tant que plateforme pour penser le monde après la pandémie, accorde beaucoup d'espace au monde de l'entreprise au début de la crise sanitaire. Même si les mots « santé » et « social » arrivent en deuxième position, d'autres termes comme « jeunes », « solidarité » ou « commun » (160 occurrences) viennent en fin de liste avec un écart de près de 750 occurrences par rapport à « travail ».

48 Après les champs sémantiques liés à l'économie et à la santé (« santé » et « vie » autour de 580 occurrences), le troisième champ qui apparaît est lié aux « enfants » et à l'« éducation » (entre 375 et 460 occurrences). Remarquons qu'avant les mots concernant la solidarité et les communs, apparaissent à nouveau des mots appartenant au champ de l'économie, notamment des « entreprises ». Quelques termes liés à l'« écologie » et à l'« environnement » sont également présents avec des fréquences moins importantes, entre 100 et 250 occurrences.

49 Ce nuage de mots nous informe également sur le nombre de contributions concernées par chaque *token*, ce qui permet de se forger une première idée de la répartition des *tokens* dans le corpus. À titre d'exemple, le terme « travail », pour lequel il y a 806 occurrences, est présent dans 486 contributions, alors que « jeune », avec ses 159 occurrences, est présent dans 126 contributions sur les 3 497 du corpus.

Figure 3. Nuage de mots de l'ensemble du corpus du site Web *Le Jour d'après* dans Orange Data Mining

Image produite par les auteurs

50 La conclusion préliminaire que nous pouvons tirer de cette analyse est que le corpus tiré du site Web *Le Jour d'après* possède un champ lexical imposant qui concerne le travail, la production et la consommation, c'est-à-dire les domaines où l'impact de la pandémie a été le plus évident et immédiat dans la période temporelle analysée.

51 Au moment du *clustering* (figure 4), nous avons arbitrairement décidé de créer 10 clusters pour répartir les contributions. Sans surprise, C4, le cluster qui contient le plus de contributions (2 598) est principalement relatif à la « vie », au « travail », à la « production », alors que les termes « enfant », « local », « social » et « santé » y sont plus secondaires même s'ils restent présents malgré tout. Les clusters C2 et C3 sont également liés au « travail » mais s'en distinguent en intégrant une dimension temporelle que C4 ne contient pas. Pour C2 (13 contributions), le mot « travail », dont le poids est six fois supérieur au deuxième et au troisième terme du groupe, laisse ensuite la place aux mots « revenir », « temps », « vie » et « société ». Alors que pour C3 (89 contributions), le terme « revenir » disparaît. Toujours dans la dynamique de production, le cluster C5 (211 contributions) est, quant à lui, dédié au champ lexical de la « production » et de la « consommation ». Le premier cluster (C1) regroupe 185 contributions pour lesquelles les termes « activité », « social », « secteur », « public » et « société » ont un poids conséquent par rapport aux autres mots. Ces thématiques sont déclinées dans les autres clusters de C6 à C10. Le cluster C7 (222 contributions) est axé sur le champ lexical de l'entrepreneuriat et intègre une dimension spatiale centrée sur la « France » et les « Français ». Le cluster C8, qui est très similaire au cluster précédent (224 contributions), voit disparaître le champ lexical du travail au profit d'un champ lexical géographique étendu à l'échelle de l'« Europe ». Les clusters C9 (231 contributions) et C10 (238 contributions) sont plus axés sur les termes « santé », « politique » et « public ». Ils se distinguent par la réapparition du mot « revenir » qui n'était plus très présent depuis le deuxième cluster.

52 C6 est un des plus petits clusters avec seulement 65 contributions. Il se détache de tous les autres, car la notion de « travail » y est secondaire, au profit de l'« école », des « enfants » et de l'« éducation ». Cette très faible représentation de la génération suivante est très révélatrice des enjeux qui préoccupent les contributeurs du jour d'après la pandémie.

Figure 4. Répartition des contributions du site Web *Le Jour d'après* en clusters dans Orange Data Mining

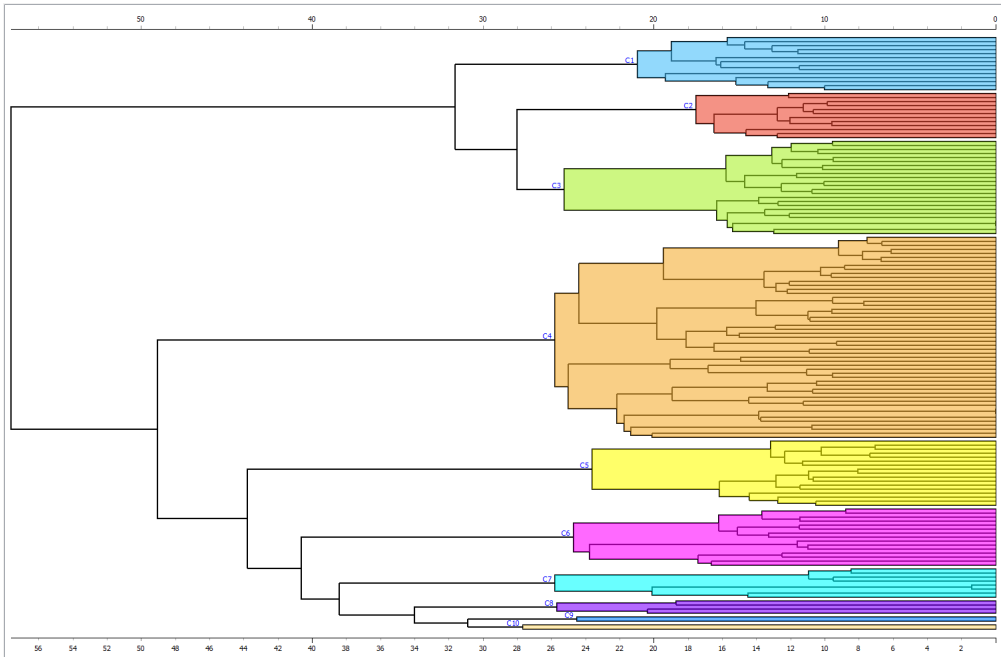

Image produite par les auteurs

53 Le champ lexical repéré dans le premier nuage de mots est très présent dans la plupart des clusters créés par le regroupement hiérarchique. Cela peut s'expliquer par le fait que ce champ lexical est présent dans la plupart des contributions. Toutefois, nous remarquons quelques finesses dans la répartition des groupes. Même si le « travail » domine la plupart des groupes, on remarque des différences dans les mots secondaires qui s'agrègent au thème principal. Chaque groupe contient quelques termes qui définissent le thème principal du groupe et plusieurs dizaines de mots dont le poids dans le texte est moins conséquent, mais qui, ensemble, peuvent définir un second thème adossé au premier.

54 En conclusion, malgré la forte présence d'un champ lexical qui écrase presque tous les autres dans ce corpus, le travail de fouille de texte nous permet de traiter et d'observer les sujets des contributions de l'ensemble du corpus pendant toute la période du premier confinement. Au moment où la situation sanitaire était diffusée en continu sur les médias de masse, les contributeurs du site Web *Le Jour d'après* font le choix de ne pas convoquer le confinement comme thème majeur dans leur perspective du monde après la pandémie. Ils préfèrent se concentrer sur la gestion du travail comme manière de se projeter vers un retour à la normalité.

La mémoire du Covid-19 sur le site *Génération Covid*

55 Comme nous l'avons observé, les captures du site *Génération Covid* pour la première période n'étaient pas suffisantes pour réaliser une analyse quantitative. Nous avons alors décidé d'élargir l'analyse jusqu'en juillet 2023. Cependant, à cause de certaines contraintes logistiques, les personnels du BNF DataLab n'étaient pas en mesure de nous fournir ces données.

56 Compte tenu de cette situation, afin de réaliser une comparaison sur un corpus couvrant les périodes ciblées pour l'analyse et valider l'approche méthodologique, nous avons construit un corpus pour le site *Génération Covid* basé sur le Web vivant et non sur les archives du Web de la BNF. Pour

rester cohérents avec notre démarche, nous nous sommes appuyés sur le site dans son intégralité puisqu'il est toujours accessible en ligne. La récupération des pages a été réalisée avec la librairie Python Requests, et leurs contenus « parsés » avec la librairie Beautiful Soup. Même si les objets du Web vivant et des archives Web de la BNF sont différents, leur structuration reste similaire. Nous avons réutilisé les scripts permettant de « parser » les archives pour ne récupérer que le contenu qui nous intéresse sur le site Web *Génération Covid*.

57 Nous avons obtenu un nombre total de 421 témoignages s'étalant de début mars 2020 jusqu'à juillet 2023 (tableau 2).

Tableau 2. Découpage des témoignages par période

Période	Nombre de témoignages
1er mars au 11 mai 2020	128
12 mai au 19 juillet 2020	41
20 juillet au 15 décembre 2020	27
16 décembre 2020 au 28 février 2021	28
1er mars 2021 au 20 juillet 2023	197

58 Dans la partie précédente, nous avons établi une méthodologie propre à nos questions de recherche et adaptée à nos données textuelles pour analyser le site *Le Jour d'après*. Pour effectuer l'analyse diachronique de *Génération Covid*, nous allons répéter cette méthode sur chacun des segments temporels décrits ci-dessus.

59 Comme précédemment, la première méthode d'analyse consiste à compter les occurrences des 300 *tokens* les plus fréquents dans les corpus. Le terme ayant le plus d'occurrences, quel que soit le corpus, est assez inattendu : il s'agit du mot « jour » dans sa forme lemmatisée. Par exemple, lors de la première période de mars à mai 2020, ce terme à un poids de 216 occurrences pour une présence dans 89 témoignages sur les 128 qui composent ce sous-corpus. Le ratio est similaire dans les autres sous-corpus. Il y a donc une façon temporelle de témoigner d'une expérience liée au Covid qui passe par l'emploi du mot « jour ». Si nous lisons de manière attentive quelques témoignages, ce terme réfère à deux significations : la première pour compter des jours (le nombre de jours de convalescence ou de la persistance des symptômes post-convalescence) et la deuxième pour distinguer le jour de la nuit. D'autres mots, également présents dans tous les sous-corpus, sont principalement liés à la maladie, comme « Covid », « douleurs », « mal » ou encore « symptôme ».

60 Un autre terme central est « peur ». Sa présence oscille fortement en fonction des périodes. Il apparaît dans 54 témoignages lors de la période du premier confinement à raison de 127 occurrences. Alors que dans la deuxième période, il devient quasiment invisible avec seulement 8 occurrences présentes dans 7 contributions.

61 Dans la troisième période, la plus longue et celle qui recueille le moins de témoignages, la « peur » revient en deuxième position dans la liste des *tokens* ayant le plus d'occurrences, avec un poids de 45 occurrences et une présence dans 13 témoignages. Ensuite, la présence de cette « peur » décroît dans les deux dernières périodes : on observe 8 occurrences dans 5 contributions pour la quatrième période pour disparaître complètement de notre liste dans la dernière période jusqu'à aujourd'hui.

62 Le *token* suivant auquel nous allons nous intéresser est « confinement ». On peut observer une très forte présence, même si elle n'est pas majeure, dans le premier sous-corpus de mars à mai 2020. Il est en 13e position, avec un nombre d'occurrences s'élevant à 71 dans 46 instances textuelles sur les 128 de ce sous-corpus. Par la suite, ce nombre reste dans une fourchette constante entre 10 et 20 par période jusqu'à février 2021. Après cette date, le « confinement » n'apparaît plus dans le corpus.

63 Enfin, un dernier terme auquel nous avons accordé notre attention est « masque », objet emblématique ayant marqué la pandémie. Sans trop de surprise, ce mot n'apparaît pas dans les deux premiers sous-corpus, de mars à juillet 2020. Cependant, à partir du 20 juillet et jusqu'à mi-décembre 2020, il entre en scène dans un peu plus d'un tiers des témoignages, pour 12 occurrences au total. À partir de décembre 2020, la présence du « masque » s'intensifie, avec une vingtaine d'occurrences dans 9 contributions. Enfin, dans la dernière période, le nombre d'occurrences du terme est de 42 pour 22 témoignages, ce qui reste assez faible par rapport aux 197 contributions de ce sous-corpus.

64 Cette première analyse permet d'emblée une observation diachronique des différents *tokens* qui peuvent signaler, ou non, une évolution du langage dans le corpus. Certains termes comme « masque » ou « confinement » pourraient intuitivement servir de repères. On s'attend à les voir apparaître à tel moment ou tel autre dans la chronologie des événements de cette pandémie. Pourtant, ce sont des mots qui restent plus discrets comme « peur » qui surgissent violemment au début de la pandémie pour ensuite s'effacer progressivement de notre corpus. D'un autre côté, nous remarquons qu'il y a aussi des repères constants, à la fois dans la mobilisation du champ lexical du temps (« jour », « semaine », etc.) et celui des « symptômes » de la « maladie ».

65 La deuxième analyse a porté sur le regroupement hiérarchique des contributions du corpus *Génération Covid*, qui est plus petit que celui du site Web *Le Jour d'après* dont nous avons parlé précédemment. En conséquence, le nombre de clusters trouvé par l'algorithme est plus faible (figure 5).

Figure 5. Clusters du sous-corpus de témoignages de la période allant de mars à mai 2020, réalisé dans Orange Data Mining

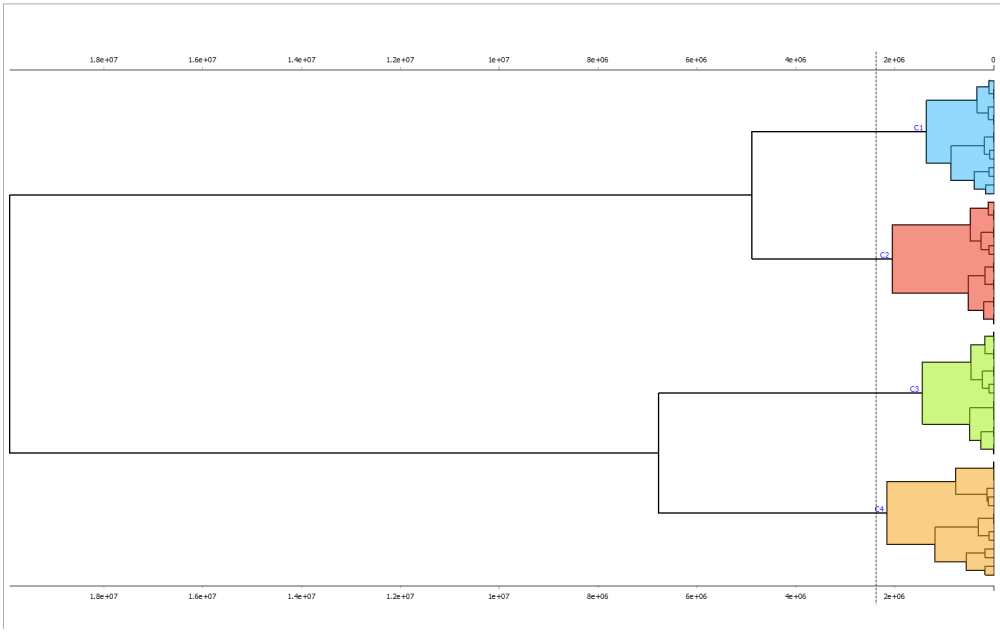

Image produite par les auteurs

66 Dans ce sous-corpus, nous voyons quatre clusters distincts. Ils ont en commun d'avoir le *token* « jour » comme élément le plus représentatif. Le premier cluster (C1) concerne principalement les « symptômes » tels que la « fièvre », les « douleurs » ou encore la « fatigue ». D'autres mots liés au « corps » forment un ensemble organique secondaire très important dans ce cluster : « gorge », « poumon », « sang », « poitrine », etc. La « peur » est un élément dont le poids est très faible, alors qu'elle est très présente dans les trois autres clusters (C2, C3 et C4). Le deuxième cluster donne beaucoup d'importance au « médecin » ainsi qu'à la « semaine » en sus des éléments nommés précédemment. C3, quant à lui, est largement centré autour du « confinement » alors que C4 associe les termes « temps » (« jour », « mars », « avril ») à des « symptômes » comme la « fièvre ».

67 Durant la deuxième période, de mai à juillet 2020, les différents clusters (de C1 à C4) se distinguent les uns des autres par les différents « symptômes » mobilisés. Toutefois, le langage évolue grâce à l'apparition dans tous les clusters du mot « test » correspondant au dépistage du virus.

68 Ensuite, pour le sous-corpus couvrant la période de juillet à décembre 2020, découpée en 5 clusters, le terme « test » ainsi que les « symptômes » restent bien présents. Néanmoins, on observe un retour de la « peur » qui redevient centrale dans les différents clusters. Le « masque » s'impose dans le cluster C1 alors qu'il reste minoritaire dans les quatre autres (C2, C3, C4 et C5). Nous remarquons également l'apparition d'un champ lexical lié à la « famille » (C5) qui renvoie principalement aux figures féminines qui la composent. « Fille » est l'un des *tokens* les plus présents dans les clusters C1 et C5 alors qu'il s'efface dans le cluster C2 pour atteindre le même niveau qu'« enfant », au profit de « maman » et « mère » qui prennent plus de place. Le « mari » et le « père », quant à eux, bien que représentés, sont plus discrets.

69 La période suivante, de décembre 2020 à février 2021 est découpée en quatre clusters. Ils ne présentent rien de nouveau, ce sont les champs lexicaux des « symptômes » et du « temps » qui recouvrent les *tokens* les plus lourds de ce sous-corpus. Nous pouvons toutefois noter la présence de « confinement » dans le top 5 du cluster C3 et l'apparition du *token* secondaire « histoire » dans le cluster C4. Enfin, à partir de février 2021, les clusters de cette période se distinguent par deux *tokens* : « travail », notamment dans le cluster C5 et « positif » dans les cinq clusters.

70 Cette deuxième analyse permet de mettre en évidence deux éléments. Tout d'abord, nous observons une constance dans l'emploi de certains termes tout au long des périodes observées. Cette constance se caractérise par les champs lexicaux de la maladie, des symptômes et du corps. Ce phénomène pourrait être expliqué par le fait que les témoignages relatent des expériences vécues par les contributeurs pendant la pandémie. Ensuite, nous pouvons observer des variations de vocabulaire en fonction des différentes périodes comme c'est le cas pour le terme « peur ». D'autres termes comme « masque » ou « test » apparaissent plus tardivement dans le corpus, ce qui laisse supposer une corrélation entre les événements marquants de la pandémie et les témoignages publiés sur le site Web *Génération Covid*. En contraste avec ces occurrences, nous aurions pu attendre de certaines qu'elles soient des marqueurs, comme « confinement », mais, finalement, elles restent marginales, que ce soit dans la première analyse avec le faible nombre d'occurrences ou dans la deuxième,

où ce terme n'est quasiment jamais un élément majeur d'un cluster, excepté dans les clusters C3 et C4 de la première période de confinement (mars à mai 2020) où sa présence est moyenne.

71 Les analyses des contributions du site Web *Génération Covid* permettent d'observer des fluctuations du langage dans les témoignages des contributeurs au fil des événements. La première méthode employée circonscrit les vocabulaires utilisés dans le corpus et de se projeter vers la deuxième méthode ; elle permet également de vérifier l'étape de prétraitement du corpus par itération. Si le corpus n'est pas bien préparé, des *tokens* sans aucun sens apparaîtront dans les calculs du nombre d'occurrences. Ensuite, la deuxième méthode offre un autre type de visualisation du corpus, complémentaire du premier, en répartissant les données textuelles en groupes sémantiques. De cette manière, nous avons pu faire émerger des tendances dans les sujets abordés et leur importance au sein du corpus.

72 Cette recherche dans les archives du Web de la BNF et l'analyse conjointe du contenu de ces deux sites nous permet d'observer une forme de division spontanée du travail de mémoire et de recueil de témoignages. En effet, les analyses textuelles présentées mettent au jour des clusters dont les thèmes ne se recoupent pas d'un site à l'autre. La pandémie peut être vue comme un « fait social total » (Gaille et Terral 2021) qui a concerné, à des degrés divers, l'ensemble des aspects des existences. Si toutes ces initiatives s'inscrivent dans un mouvement global de « devoir de documentation » de la pandémie (Kosciejew 2022), chaque site Web analysé poursuit un objectif qui lui est propre et dont dépendent la forme et le contenu des témoignages recueillis. Là où *Génération Covid* s'est, dès le départ, positionné comme une collecte de témoignages sur la maladie, le deuil et la rémission, *Le Jour d'après* se présente comme une initiative plus ouvertement politique, ayant pour ambition de tirer des leçons de la pandémie pour apporter des idées nouvelles. Ces deux sites s'inscrivent par ailleurs, comme nous l'avons rappelé au début de cet article, dans une constellation plus vaste de sites créés au moment de la pandémie et visant à recueillir et diffuser diverses formes de témoignages sur cette période. En archivant ces différents sites Web, la collecte de la pandémie faite par la BNF constitue déjà une forme de métacollecte. Pour que les chercheurs et historiens de demain puissent se saisir de l'ensemble de ces collectes et témoignages recueillis ainsi que les traiter, il reste à établir des protocoles permettant la mise en données de ces archives, dont nous avons ici posé les premiers jalons.

Conclusion

73 Les analyses approfondies menées sur les sites Web *Le Jour d'après* et *Génération Covid* ont fourni des preuves de concept significatives. Bien que les conclusions tirées soient spécifiques à chaque site et que leur portée ne puisse être généralisée à l'ensemble de la collecte Covid-19 des archives du Web de la BNF, ces preuves de concept jouent un rôle crucial dans la validation des approches méthodologiques appliquées au corpus et dans la confirmation de la robustesse de la chaîne de traitement des données. En particulier, les variations de langage observées entre les périodes déterminées attestent de la pertinence de la périodisation choisie pour l'étude et montrent l'intérêt de construire des corpus qui combinent et comparent des données des archives du Web avec des données du Web vivant. De même, le choix des environnements technologiques, des logiciels et des algorithmes, bien que perfectibles, demeure cohérent avec la problématique et offre des résultats interprétables conformes à nos attentes.

74 Au-delà de ces considérations, l'exploration du corpus des archives du Web de la Bibliothèque nationale de France nous a confrontés à la réalité tangible des archives numériques, englobant des infrastructures institutionnelles, logicielles et matérielles, entre autres. Cette convergence de dynamiques crée un corpus d'une richesse informative considérable mais la navigation à travers celui-ci peut sembler initialement complexe en raison de la délicatesse de la chaîne de traitement. Nous ne voulons pas cacher l'aspect « bricolage » de notre *workflow* mais, au contraire, en montrer sa richesse et sa nécessité. En effet, il découle de deux réalités incontournables : d'une part, la BNF ne propose actuellement aucun *workflow* complet pour le traitement de ce type de données et, d'autre part, le développement de nos propres scripts, même pour des tâches telles que le nettoyage des données, offre une immersion unique dans le corpus d'archives, révélant une fine granularité dans la structuration des textes à l'intérieur de chaque document HTML.

75 Notre approche s'est efforcée de prendre en compte les divers niveaux de structuration des informations dans l'écosystème de la BNF, s'étendant du macro au micro, de l'infrastructure globale aux nuances des balises HTML au sein de chaque document. En résumé, cette tentative de compréhension exhaustive souligne l'importance de s'adapter aux particularités de la matérialité des archives du Web, démontrant ainsi la nécessité d'approches flexibles et contextualisées pour explorer ces vastes gisements d'informations.

Bibliographie

Adams, Tracy et Sara Kopelman. 2022. « Remembering Covid-19 : Memory, Crisis, and Social Media ». *Media, Culture & Society* 44 (2) : 266-285. https://doi.org/10.1177/01634437211048377.

Bachimont, Bruno. 2023. « L'archive du Web : une nouvelle herméneutique de la trace ? » *Web Corpora* (blog). https://webcorpora.hypotheses.org/288.

Benoist, David, Alexandre Faye, Pascal Tanesie, Sophie Gebeil et Valérie Schafer. 2020. « Exploring Special Web Archive Collections Related to Covid-19 : the Case of the French National Library (BNF) ». *WARCnet Papers*, décembre. https://cc.au.dk/fileadmin/user_upload/WARCnet/Gebeil_et_al_COVID-19_BnF.pdf.

Brügger, Niels. 2012. « L'historiographie de sites Web : quelques enjeux fondamentaux ». *Le Temps des médias* 18 (1) : 159-169. https://doi.org/10.3917/tdm.018.0159.

Ducas, Sylvie, Rossana De Angelis et Agathe Cormier, éd. 2022. *Les Écritures confinées. Créer, afficher, diffuser*. Paris : Hermann.

Faye, Alexandre. 2020. « Les archives Web du coronavirus : une entreprise collective ». *Web Corpora* (blog). https://webcorpora.hypotheses.org/856.

Gaille, Marie et Philippe Terral, éd. 2021. *Pandémie. Un fait social total*. Paris : CNRS Éditions.

Gebeil, Sophie. 2017. « Les vidéos humoristiques de l'Internet à l'assaut du traitement médiatique de l'"Arabe" (2005-2013) ». *Le Temps des médias* 28 (1) : 123-143. https://doi.org/10.3917/tdm.028.0123.

Gensburger, Sarah et Marta Severo. 2021. « L'espace public du confinement. Archives, participation et inclusion sociale ». *Revue d'histoire culturelle* 3. https://doi.org/10.4000/rhc.662.

Hervé, Nicolas. 2022. « Étude quantitative de l'intensité médiatique des six premiers mois de la pandémie du Covid-19 ». *Les Cahiers du journalisme – Recherches* 2 (8-9) : R13-R30. https://cahiersdujournalisme.org/V2N8/CaJ-2.8-R013.html.

Kosciejew, Marc. 2022. « Remembering Covid-19, a Duty to Document the Coronavirus Pandemic ». *IFLA Journal* 48 (1) : 20-32. https://doi.org/10.1177/03400352211023786.

Maemura, Emily. 2023. « All WARC and No Playback : the Materialities of Data-Centered Web Archives Research ». *Big Data & Society* 10 (1) : 20539517231163172. https://doi.org/10.1177/20539517231163172.

Schafer, Valérie. 2021. « Web Archives of the COVID Crisis : Digital Voices, Preservation and Loss ». Table ronde au *workshop Pandemic Grief (1) : Grief in Public, Vulnerability, and the Political « We »*, Munich, 16 septembre. https://orbilu.uni.lu/handle/10993/48056.

Schafer, Valérie et Jane Winters. 2021. « The Values of Web Archives ». *International Journal of Digital Humanities* 2 (1-3) : 129-144. https://doi.org/10.1007/s42803-021-00037-0.

Segault, Antonin et Marta Severo. 2023. « Les archives audiovisuelles de l'INA, une ressource d'expérimentation entre pédagogie et recherche ». *Les Cahiers du numérique* 19 (1-4) : 93-118. https://www.cairn.info/revue-les-cahiers-du-numerique-2023-1-page-93.htm.

Notes

1 Le projet *Web-mémoires* a été développé en collaboration avec le DataLab de la BNF et l'INA Lab. Il a été soutenu par le Labex *Les passés dans le présent* (Réf. ANR-11-LABX-0026-01) et l'INA.

2 Sur ces différences entre éléments, pages, sites, voir Brügger 2012.

3 https://mrap.fr/un-virus-n-a-pas-d-origine-ethnique.html.

4 La liste des sites collectés durant cette collecte exceptionnelle et la documentation associée peuvent être trouvées ici : https://www.data.gouv.fr/fr/datasets/collecte-du-web-consacree-a-lepidemie-de-covid-19/.

5 Cette requête a été menée au sein de l'ensemble des sites de la collection Covid-19, non seulement parmi les mots-clés associés aux sites archivés par les documentalistes de la BNF mais également en recherche sur du texte brut. L'enjeu était de construire un sous-corpus constitué par les sites recueillant des témoignages et récits liés à la pandémie. En effet, il est important de noter qu'une particularité du corpus Covid-19 était d'être indexé en texte brut. Cependant, le moteur de recherche qui permet de chercher dans les archives du Web ne s'appuie pas sur un algorithme de classement. Par conséquent, la construction d'un sous-corpus est une action entièrement qualitative qui implique de regarder les résultats un par un. D'autres projets, comme *AWAC2* (*Analysing Web Archives of the COVID Crisis*) ont été confrontés aux mêmes difficultés ; voir à ce propos https://netpreserveblog.wordpress.com/2022/12/20/studying-women-and-the-covid-19-crisis-through-the-iipc-coronavirus-collection/.

6 Par exemple, http://luttevirale.fr.

7 https://twitter.com/hashtag/JournalDeConfinement/.

8 Orange Data Mining est un logiciel libre, développé par le laboratoire de bio-informatique à la faculté d'informatique et des sciences de l'information de l'université de Llubljana en Slovénie. Nous levons la confusion possible avec la firme française Orange : Orange Data Mining n'a aucun lien avec cette entreprise de services en télécommunication. Le logiciel est adossé à une licence GNU General Public License 3.0 et les documentations et documents additionnels sont tous sous licence Creative Commons Attribution-ShareAlike. La base logicielle est développée dans le langage de programmation C++, et les *widgets* sont développés avec le langage de programmation Python. Orange Data Mining propose une interface de programmation visuelle pour construire des chaînes de traitement des données. Chaque composant de cette chaîne est nommé *widget*. De la même façon que dans les logiciels Pure Data ou Max/MSP, il suffit de positionner les *widgets* dans un espace de travail et de les relier entre eux pour produire la chaîne de traitement des données. Il existe une pléthore de plugins additionnels que les utilisateurs peuvent installer en supplément de la structure de base pour compléter le logiciel, comme c'est le cas pour le plugin *text* dédié à la fouille de texte.

9 Un « freeze » fait référence à l'état d'un système informatique qui cesse de répondre aux commandes de l'utilisateur.

10 Voir le site Web https://ufal.mff.cuni.cz/udpipe/1/ pour plus de détails.

Auteurs

Roch Delannay

Chaire de recherche du Canada sur les écritures numériques, université de Montréal, Montréal, Canada et EA 7339 DICEN-IDF, université Paris-Nanterre, Nanterre, France
Roch Delannay est doctorant en humanités numériques à l'université Paris-Nanterre et à l'université de Montréal. Sous la direction de Marta Severo, Marcello Vitali-Rosati et Emmanuel Château-Dutier, il mène sa recherche au croisement de la théorie des médias et des sciences de l'information et de la communication sur les processus de construction et de manifestation de l'intimité des chercheurs et chercheuses dans les publications scientifiques. Il coordonne le projet de recherche *Stylo* développé à la Chaire de recherche du Canada sur les écritures numériques.
ORCID 0000-0002-3519-4365
roch.delannay@umontreal.ca

Marta Severo

EA 7339 DICEN-IDF, université Paris-Nanterre, Nanterre, France
Marta Severo est professeure des universités en sciences de l'information et de la communication à l'université Paris-Nanterre. Elle est directrice adjointe du laboratoire DICEN-IDF. Les pratiques participatives et contributives en ligne figurent parmi les thématiques de recherche qu'elle développe. Elle a notamment porté le projet ANR *Collabora* sur les plateformes contributives culturelles et obtenu le statut de chercheur junior de l'Institut universitaire de France. En 2019, elle a rejoint le conseil scientifique d'OpenEdition.
ORCID 0000-0001-6901-7001
msevero@parisnanterre.fr

Louis Gabrysiak

UMR 7220 ISP, université Paris-Nanterre, Nanterre, France
Louis Gabrysiak est docteur en sociologie. Après une thèse à l'EHESS sur les styles de vie et la vocation des universitaires, il mène des recherches sur les collectes mémorielles mises en place durant la pandémie de Covid-19. Ses travaux s'intéressent aux professionnels mettant en place ces collectes ainsi qu'à leurs contributeurs.
ORCID 0000-0002-1071-8679
louis.gabrysiak@gmail.com

Droits d'auteur