Les archives du Web et la pandémie de Covid-19 : entre logiques institutionnelles et initiatives personnelles

Sarah Gensburger, Louis Gabrysiak, Marta Severo

Les archives du Web et la pandémie de Covid-19 : entre logiques institutionnelles et initiatives personnelles

Louis Gabrysiak

Docteur en sociologie, laboratoire d'excellence « Les passés dans le présent » (Labex PasP)

Sarah Gensburger

Directrice de recherche au CNRS, Sciences Po Paris, Centre de sociologie des organisations (CSO)

Marta Severo

Professeure des universités, Université Paris-Nanterre,
laboratoire « Dispositifs d'information et de communication à l'ère numérique » (Dicen-IdF)

Au cours des dernières décennies, les institutions de documentation et de recherche ont pris conscience de la nécessité de conserver une trace de ce qui se déroule sur Internet et d'archiver ces informations (Musiani, Paloque-Bergès, Schafer et Thierry, 2020). En effet, Internet représente une source précieuse pour documenter l'histoire et enrichir la mémoire collective d'un pays, méritant ainsi d'être intégré dans le cadre du dépôt légal (Brügger, 2019 ; Schafer et Winters, 2021). En France, deux institutions publiques, la Bibliothèque nationale de France (BnF) et l'Institut national de l'audiovisuel (INA), sont légalement responsables de ce qui est communément qualifié d'archivage du Web. Leurs missions sont définies par la loi relative au droit d'auteur et aux droits voisins dans la société de l'information, dite loi DADVSI, de 2006, suivie du décret d'application de 2011 qui 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 »* (article L131-2 du Code du patrimoine).

L'INA est chargé d'étendre sa conservation de l'audiovisuel français en suivant ses mutations, elle archive ainsi le « Web média » constitué des sites Web des chaînes de télé et de radio, des WebTV et WebRadios, des sites liés aux programmes (sites d'émissions comme sites de fans), certains contenus édités sur les plateformes de partage audiovisuelles et, enfin, les flux Twitter liés à ces programmes, à travers des hashtags et des comptes individuels (sites de chaînes, WebTV…) (Mussou, 2012). La BnF, elle, doit s'occuper du reste de l'Internet français (domaines en .fr ou extensions régionales, sites créés en France, presse, etc.). Dès le départ, l'INA a un périmètre plus balisé. Celui de la BnF se définit davantage en négatif : tout ce que ne couvre pas l'INA. Ce découpage est donc le résultat d'un découpage institutionnel, le Web en lui-même n'étant pas divisé de la sorte (Bachimont, 2017). Ces institutions ont élaboré des politiques continues d'archivage (Bermès, 2020 ; Mussou, 2012). Cependant, elles doivent également être en mesure de préserver les événements en période de crise, où les pratiques d'archivage standards risquent de ne pas suffire.

À la différence des précédentes collectes d'urgence (par exemple celle ayant fait suite aux attentats de 2015), la crise du Covid-19 a ceci de particulier que les institutions et leurs agents se trouvent pris dans la crise qu'ils doivent collecter. Il s'agit de vivre la crise et de la collecter simultanément. Aux difficultés que pose la fermeture des institutions (devant malgré tout réagir rapidement) s'ajoute l'hétérogénéité des situations et des conditions de travail des différents archivistes du Web.

Apparue à bas bruit en janvier 2020, la pandémie a fait l'objet d'une décision de confinement brutale annoncée le 18 mars 2020 dont on ignorait alors la durée[1]. Renvoyés chez eux du jour au lendemain, salariés comme élèves ou étudiants ont dû apprendre à travailler « en distanciel » et à organiser une part

1 Ce premier confinement prendra fin le 10 mai 2020. Il sera suivi par so autres confinement du 31 octobre au 14 décembre 2020 puis du 4 avril au 2 mai 2021. D'autres mesures sanitaires comme le port obligatoire du masque ou l'instauration d'un couvre-feu en fonction des régions émailleront également cette période.

importante de leur vie sociale et culturelle à travers le Web. Le passage a été plus ou moins aisé en fonction du métier, du statut, mais aussi de l'équipement et de l'aisance numérique de la population confinée. À ce changement de vie se sont ajoutés l'inquiétude liée à la maladie, pour certains le stress de la solitude, voire les problèmes financiers. Face à l'incertitude, les théories du complot et l'angoisse sont montées en force. Soumis, comme tout le monde, à ces conditions si particulières, les agents publics en charge du dépôt légal du Web et de la collecte des archives d'une histoire en train de se faire ont réagi au mieux afin d'assurer la collecte.

Au sein du projet de recherche Web-mémoires (Labex Les passés dans le présent / INA), nous avons étudié les cadres sociaux et les dynamiques qui sous-tendent ces initiatives de conservation. À partir d'une campagne d'entretiens[2] auprès des acteurs de l'archivage du Web de la BnF et de l'INA, ainsi que d'une enquête sur les archives numériques produites, cet article entend revenir sur les modalités et la mise en œuvre de cette collecte autour de la pandémie et du premier confinement. En nous appuyant sur l'exemple des collectes numériques liées au Covid-19 créées dans le cadre du dépôt légal, nous voulons réfléchir de manière plus générale sur les politiques institutionnelles mises en place pour prévenir les risques d'oubli, et définir de manière préventive l'héritage de certains événements qui sont aujourd'hui de plus en plus vécus comme se déroulant en partie sur le Web et les réseaux, lieux privilégiés de l'expression des opinions et des sentiments, mais aussi supports des actions et des initiatives devenues numériques. Il s'agit pour nous de considérer les archives du Web comme une modalité de cette « pre-emptive memory » que la crise sanitaire a mis en évidence (Mazzucchelli et Panico, 2021). Après avoir décrit la manière dont la collecte du Web lié au coronavirus s'est organisée à la BnF et étudié les spécificités des résultats en fonction des territoires, nous expliquerons en quoi ont consisté les collectes de l'INA par rapport à son périmètre habituel et sur Twitter.

Archiver la crise sanitaire en temps réel à la BnF

Le Covid-19 : mise en place d'une collecte dans des conditions inédites

Le coronavirus fait son entrée dans les archives du Web de la BnF dès le mois de janvier 2020. En plus d'une collecte annuelle en octobre-novembre de l'ensemble du Web français assurée par le département du Dépôt légal du Web (5,8 millions de sites en 2022), il existe à la BnF des collectes thématiques (au moment des élections par exemple) ainsi qu'une procédure de collecte « d'urgence », qui peut être déclenchée pour archiver rapidement des sites amenés à disparaître à court terme. Cette dernière catégorie a été mise en place suite aux attentats de novembre 2015 à Paris. Elle a conduit à une innovation organisationnelle : un « café de l'actualité éphémère » (« Actualité FMR ») est mis en place tous les vendredis pour discuter, entre membres volontaires, de l'actualité, des événements à venir et se répartir en conséquence les tâches d'archivages avec des petites collectes ciblées.

Ces collectes d'archives sont effectuées par les cinq ou six agents du département du Dépôt légal du Web (DLWeb), par des correspondants thématiques au sein des différents départements de la BnF et par les correspondants régionaux, des bibliothécaires en charge du dépôt légal imprimeur au sein des bibliothèques et centres d'archives habilités (26) et qui ont également une mission au niveau de l'archivage du Web lié à leur territoire. L'ajout et la gestion des URL se font via un outil, BCWeb, accessible à distance. Quiconque a un compte peut s'y connecter, entrer une URL, déterminer une fréquence et une profondeur de crawl (collecte[3]).

Dans le cadre de la « collecte de l'actualité FMR », les 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 ! »[4], sont collectés (Faye, 2020). Rapidement, une activité de collecte liée à l'événement va se développer (Schafer, 2022). Cette collecte du Web, incluant sites Internet, réseaux sociaux et vidéos en ligne, s'est très vite intensifiée pour conserver les réactions et les modes de vie de la société française face à la pandémie et au confinement (Schafer, 2022 ; Gebeil, Schafer, Benoist, Faye, Tanesie, 2020). Le nombre de sélections va s'étendre et la décision va être prise de solliciter la participation de l'ensemble des correspondants internes et externes volontaires.

Dès les premiers jours du confinement, le département du DLWeb rédige et fait parvenir à ses correspondants un « Mémento pour la collecte en rapport

2 Au total, nous avons rencontré et interrogé 12 personnes participant ou ayant participé à l'archivage du Web français, 10 du côté de la BnF, sous forme d'entretiens individuels et d'un entretien collectif (membres du département du Dépôt légal du Web, correspondants internes et externes ayant participé à la collecte) et 2 à l'INA (le responsable de la mission au dépôt légal du Web et la chargée de mission en documentation pour la valorisation des collections Web du dépôt légal).

3 Voir : https://fr.wikipedia.org/wiki/Robot_d%27indexation

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

avec la crise sanitaire du coronavirus (Covid-19) »[5]. Celui-ci sert à assurer la bonne division du travail d'archivage et la répartition des tâches. Il définit en effet le périmètre de la collecte pour chaque département disciplinaire de la BnF ainsi que pour les correspondants régionaux. Le mémento précise qu'il faut éviter les articles de presse, déjà couverts par la collecte « Actualités » (en effet la BnF archive systématiquement la presse et une partie de la presse en ligne, gratuite ainsi que quelques journaux payants avec qui elle a noué un partenariat). Les correspondants sont libres dans leurs sélections, il leur est demandé, *« pour garantir une meilleure qualité de collecte », de « sélectionner les contenus spécifiques au coronavirus (ne pas saisir l'URL de départ d'un Centre hospitalier universitaire - CHU par exemple) ».* Il détaille également les règles à suivre pour l'indexation de ces sites ou pages, en donnant des consignes pour les mots-clés à leur attribuer : chaque fiche créée doit être indexée avec « coronavirus », elle doit également comporter au moins l'un des mots-clés généraux repris de la collecte de l'IIPC[6] : « origines du virus », « propagation du virus », « mesures d'endiguement », « aspects médicaux et scientifiques », « aspects sociaux », « aspects économiques », « aspects politiques ». Il est laissé à la discrétion de l'archiviste la possibilité d'ajouter d'autres mots-clés si cela lui semble utile. Enfin, le mémento répartit des thématiques selon les départements et les bibliothèques dépôt légal imprimeur (BDLI). Le département Littérature et Art se voit par exemple confier les sujets « expression individuelle », « expériences individuelles », « contenus culturels gratuits » alors que celui des Sciences et Techniques a en charge les « aspects scientifiques », la « recherche » et les « aspects sanitaires ». Les BDLI sont chargées de ce qui a trait à leur zone géographique et à l'impact local ou régional sur l'économie, et doivent ajouter à chacune de leur sélection un mot-clé géographique (ville, région…).

Lors de l'entretien qu'il nous a accordé, Alexandre Faye, chargé de collection numérique au département DLWeb, nous explique que la réactivité dont a pu faire preuve la BnF provient de la conjonction de deux éléments : d'une part la mise en place, à la suite des attentats de 2015, d'un cadre et de procédures pour lancer des collectes d'urgence, d'autre part la collecte normale des élections municipales en cours au moment du Covid. De ce fait, les listes d'adresses mail étaient à jour et les guides pour l'utilisation de l'outil BCWeb ont pu être rapidement actualisés et diffusés. La collecte de la crise sanitaire a donc été une entreprise collective. Au total, 58 correspondants ont participé à la saisie d'URL. Entre février et juillet 2020, 4469 URL ont été sélectionnées, complétées à partir de juillet par des collectes spécifiques de chaînes YouTube et de comptes Instagram. Les trois types d'acteurs mis au jour précédemment ont été mobilisés (membres du département du Dépôt légal du Web, correspondants internes et correspondants externes).

Ces 4469 URL peuvent être classées en trois catégories en fonction de leur statut par rapport au Covid. La première peut être désignée de « Web covid », elle rassemble l'ensemble des sites créés durant la pandémie et disposant de « covid » dans leur nom de domaine. Leur collecte est plutôt le fait du département DLWeb, les sélections ont été faites à partir de listes de noms de domaines fournies par l'Afnic[7] ou des hébergeurs comme OVH. Ensuite, ce que nous pouvons désigner comme « covid sur le Web » est composé des actualités scientifiques, économiques, sociales en lien avec la pandémie. Son archivage est davantage le fait des correspondants régionaux des BDLI. Enfin, ce qui apparaît comme le « Web pendant le covid » rassemble des sites préexistants qui ont une activité particulière durant le Covid. Ce dernier groupe a été retenu davantage par les correspondants internes des différents départements disciplinaires de la BnF qui, pour cela, sont souvent repartis des listes de leurs collectes courantes.

Il est à noter que ce maillage laisse courir le risque de manques dans la collecte, Alexandre Faye a assuré le suivi et a fait en sorte de boucher certains trous. En effet, il y a des sites qui ne relèvent ni des correspondants internes, ni externes et qui ne sont pas forcément en .fr :

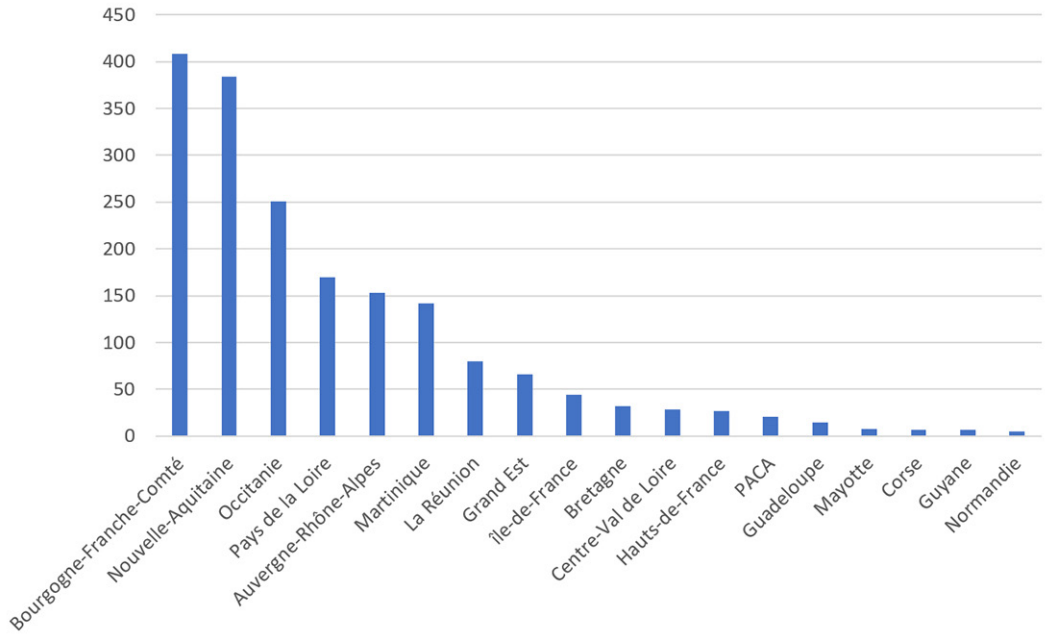
« Même si on sélectionne de la masse, on peut avoir des trous importants, et moi ma démarche c'était plutôt de me dire où est-ce qu'on peut avoir des trous. Donc il y avait toujours un peu cette logique de regarder l'actualité, puisqu'effectivement il y a eu pas mal d'actualités sur la continuité pédagogique, les violences conjugales, voilà, aussi tout ce qui était un peu arnaques, cybersécurité, des choses comme ça, il pouvait y avoir des moments, des choses qui remontaient un peu dans l'actualité, et donc du coup, bon, à ce moment-là, on va se mettre dessus, et l'idée c'est effectivement, si on fait la recherche sur Google, classiquement avec juste les mots-clés, on va voir surtout la presse qui remonte, donc en général il faut quand même faire l'effort aussi d'aller se dire, tiens, il y a peut-être derrière un communiqué de presse, il y a peut-être une association et d'aller chercher aussi des magazines plus spécialisés, ou des fois des magazines professionnels, enfin voilà, si les pompiers de Paris ont un truc en ligne, ça vaut le

5 Nous avons obtenu du DLWeb une version de ce mémento.

6 L'IIPC (International Internet Preservation Consortium) est une organisation internationale regroupant un ensemble de bibliothèques nationales et d'institutions patrimoniales afin d'assurer une bonne coordination et coopération dans l'archivage du Web. Voir : https://netpreserve.org/event/iipc-cdg-collection-novel-coronavirus-outbreak/

7 Afnic, Association française pour le nommage Internet en coopération. Elle gère les noms de domaines en .fr et communique à la BnF l'ensemble des sites en .fr créés.

Figure 1. Répartition géographique des mots-clés de la collecte Covid (par région)



coup d'aller voir ce qu'ils ont sur leur site, et du coup on est dans cette démarche de compléter, c'est plutôt qu'il faut se mettre une limite en termes de temps, combien de temps on va y passer, voilà, et après il faut réfléchir aux paramètres, à la configuration, donc est-ce que ça, c'est plutôt quelque chose à prendre en mensuel, en hebdo » (Alexandre Faye). Ce qui s'applique aussi aux sites complotistes ou aux *fake news*, qu'il a ciblés volontairement.

Des disparités territoriales

Le nombre élevé de participants et participantes à ce travail collectif recouvre néanmoins des engagements et des investissements différenciés, notamment en fonction des territoires. La BnF fonctionnant en partenariat avec les BDLI, une attention particulière est portée de longue date à la question de la territorialité du Web et/ou des « territoires » sur le Web. Cet enjeu géographique est important dans le cadre de la pandémie et du confinement. Bien que l'événement soit sans aucun doute global, l'échelon local a joué un rôle déterminant dans la gestion du quotidien, l'organisation de mesures de solidarité et des mobilisations sociales. La progression de l'épidémie, sa temporalité et son impact sur la situation des hôpitaux ont été variables selon les lieux. Certaines régions ou certains départements sont très présents dans la collecte Covid, d'autres sont presque totalement absents. Le graphique ci-dessous (figure 1), a été obtenu à partir du nombre d'occurrences de chaque région, département et ville présents dans les mots-clés de la liste de la collecte Covid[8]. Nous avons uniformisé ces différents mots-clés géographiques en recodant chaque ville et département par leur région d'appartenance.

Ce graphique donne à voir la très forte variabilité de l'engagement des différents correspondants locaux. La région Bourgogne-Franche-Comté est par exemple très fortement sur-représentée (eu égard à sa population), d'autres régions comme le Grand Est, l'Île-de-France ou la Normandie sont, elles, sous-représentées.

Ces différences s'expliquent notamment par la diversité des situations (et des contraintes) vécues par les correspondants et les correspondantes. Derrière la « collecte Covid » se trouvent en réalité autant de collectes locales, comme autant de manières de composer avec la crise. Parmi les régions peu représentées, le Grand Est fait habituellement figure de « *bon élève* » (selon la correspondante locale) lors des opérations de collectes. Sa faible présence ici provient d'un défaut d'équipement informatique lors du confinement, empêchant la participation. La forte présence de la Martinique tient pour partie à l'investissement personnel intense de la correspondante. L'activité de collecte coïncidait avec son propre besoin de lire tout ce qui pouvait être produit sur le Covid et la situation en Martinique, comme moyen de s'extraire, de « *se détacher, prendre du recul* », alors que « *des gens mouraient à la pelle* ». La proximité physique de l'archiviste avec les malades et les morts a conféré à sa tâche une importance particulière, celle-ci est vue

8 Ce graphique repose sur l'ensemble des adresses URL de la collecte consacrée à l'épidémie de Covid-19 de la BnF. La liste des URL et des mots-clés qui leur sont associés est disponible ici : https://api.bnf.fr/fr/node/176

comme une véritable « *mission de service public, pour l'histoire de demain sur la Martinique* ».

Les méthodes de recherche mises en place afin de trouver et sélectionner de nouveaux sites sont également variables. Elles dépendent de l'existence ou non de corpus préétablis pour de précédentes collectes sur lesquels s'appuyer (par exemple, le fait d'avoir une base constituée de « sites locaux ») ainsi que des manières de faire de l'archiviste. La correspondante de Bourgogne Franche-Comté, région très présente dans la collecte, a construit un ensemble de requêtes[9] pour moteurs de recherche, réitérées régulièrement afin d'avoir une vue assez exhaustive de sa région sur le Web. La correspondante de Martinique a mêlé connaissance préalable des blogs et sites d'opinions importants et des requêtes plus simples, mais reprises chaque jour en visant l'exhaustivité de la recherche (l'ensemble des pages Google).

Enfin, il faut noter que la faible présence de l'Île-de-France montre que celle-ci n'est pas pensée comme une « région » par l'institution (il n'y a pas de « correspondant régional » en Île-de-France), mais qu'au contraire, ce qui se passe à Paris est considéré comme ayant d'emblée un rayonnement national. Ainsi, dans la constitution de la mémoire du confinement, les différentes régions ne seront pas présentes de la même manière non pas en raison des événements eux-mêmes, du nombre de documents potentiellement à archiver ou selon un choix d'échantillonnage, mais en fonction des conditions de travail et des pratiques antérieures des archivistes.

Le Covid et les confinements comme « événements médiatiques » dans les archives de l'INA

Covid et collections courantes

À l'INA, il n'existe pas de département DLWeb, mais un responsable de mission, un chargé de la collecte et une chargée de mission et de valorisation des collections Web, ainsi que trois développeurs. Leur périmètre d'archivage étant davantage défini que celui de la BnF, leur collecte courante totalise un peu plus de 16 000 sites (dont environ 14 000 encore actifs aujourd'hui). Est inclus tout ce qui peut prolonger la mission habituelle de l'INA, soit la conservation du patrimoine audiovisuel, en dehors de ses supports historiques, la radio et la télévision. En plus de la collecte des sites « médias », une composante importante de l'archivage d'Internet par l'INA concerne Twitter, comme nous allons le voir[10].

Enfin, l'INA archive un nombre important de vidéos, issues de différentes plateformes (YouTube, Dailymotion, feu Vine et bientôt TikTok). Les diffusions en *live* ne sont pour l'instant pas couvertes. Les diffusions sur Twitch ne sont pas archivées, mais lorsque des chaînes laissent publiquement leurs rediffusions, celles-ci peuvent l'être. L'archivage de vidéos avait pour premier objectif de conserver les vidéos intégrées aux sites Web. Face à la montée en importance d'un écosystème propre à YouTube (qui n'est plus qu'une simple plateforme d'hébergement), des comptes de créateurs de contenus ont été progressivement ajoutés à la collecte. Comme pour Twitter, le choix de ces comptes est à l'appréciation des membres de l'équipe du dépôt légal, il s'agit de faire une sélection comprenant les principaux vidéastes français, des chaînes en lien avec la production audiovisuelle et des chaînes jugées intéressantes vis-à-vis du reste du corpus et des objectifs de l'INA.

Aucun site spécifique au coronavirus n'a donc été ajouté à la base de sites archivés par l'INA, ce qui aurait conduit l'institution à sortir de son périmètre (Schafer, Thièvre et Blanckemane, 2020) à l'inverse de ce qui a été fait à la BnF. Cela ne signifie pas que l'on ne trouve aucune trace du Covid et de cette période dans les archives Internet de l'INA. Celui-ci est présent d'une part au sein du contenu archivé par les collectes courantes durant la pandémie, et d'autre part grâce à la collecte Twitter.

Les collectes courantes ont archivé un volume très important de contenus traitant du Covid. Une interrogation à partir des mots-clés « coronavirus », « covid » et « confinement » en prenant comme bornes temporelles 1er janvier 2020 – 31 juillet 2020 dans la base « Web média » des sites archivés par l'INA renvoie par exemple, pour « coronavirus » 2 754 445 résultats, pour « covid » 2 418 096 et, enfin, pour « confinement » 2 543 795[11]. La période du Covid a ceci de particulier que la pandémie a fait l'objet d'un temps d'antenne dédié aux informations exceptionnellement élevé (Bayet et Hervé, 2020). Et au sein de ces informations, la pandémie a occupé, lors du premier semestre de 2020, 60 % de l'offre d'information globale en nombre de sujets (Poels et Lefort, 2020). En collectant à une fréquence élevée les sites des différents médias, l'INA a donc collecté un volume très important de pages en lien avec la pandémie. À titre d'exemple, on compte pour le seul site *francetvinfo.fr* 45 257 pages collectées contenant le mot « coronavirus » entre janvier et juillet 2020.

Les archives Web de l'INA couvrent donc la façon dont les 16 000 sites présents dans la collecte courante de l'institution traitent (ou non) de la pandémie. En plus des sites d'informations, on trouve des contenus plus marginaux, comme certains blogs personnels ou professionnels couverts par la collecte habituelle qui

9 Voir https://Webcorpora.hypotheses.org/953

10 Sur l'archivage du Web à l'INA, notamment dans ses aspects techniques, voir Boukadida, Blanckemane et Thièvre, 2023.

11 Et, si l'on étend les bornes temporelles jusqu'à fin 2023, ces résultats dépassent les 8 millions pour « coronavirus », 15 millions pour « covid » et 12 millions pour « confinement ».

ont publié des articles en lien avec la pandémie[12]. La base de l'INA offre peu d'information sur ces pages archivées. Nous avons le contenu des balises HTML « title » et « description » qui permettent d'avoir des informations basiques sur leur contenu. Les pages ne sont pas catégorisées par rapport à leur thématique, mais le nom de domaine permet généralement de comprendre qui est l'acteur énonciateur. L'absence d'une collecte Web *ad hoc* fait naître des enjeux quant à l'accès à ce contenu archivé. En l'absence d'une liste de sites ou de pages ayant fait l'objet d'une curation, il faut pouvoir s'orienter parmi ces millions de résultats. Les archives du Web de l'INA sont accessibles via une interface dédiée. Ses deux principaux onglets, « Navigation et « Recherche », proposent deux modes d'entrée distincts dans ces archives. « Navigation » donne accès à une base de données documentaire, contrôlée et documentée par les documentalistes. On y trouve les listes de sites, de hashtags ou de chaînes vidéo qui ont fait l'objet d'une sélection de la part des documentalistes, ainsi que des informations sommaires sur ces sites, hashtags et chaînes vidéo. L'onglet « Recherche », lui, est l'index des contenus eux-mêmes. Tout ce que crawle l'INA est entièrement et systématiquement indexé. Cela relève d'une décision prise dès l'origine de l'archivage et rendue possible encore aujourd'hui par le fait que le corpus ne comporte pas un nombre trop important de sites.

La collecte Twitter

En plus de la collecte des sites « médias », une composante importante de l'archivage d'Internet par l'INA concerne donc Twitter, archivage commencé en 2014. Outre des comptes liés aux chaînes de télévision et d'émissions ou émanant de journaliste, l'INA récupère des hashtags, sélectionnés manuellement par Tiffany Anton[13], parfois épaulée par Jérôme Thièvre et Boris Blanckemane. Dans la liste se retrouvent des hashtags liés à une émission ou une chaîne (par exemple #FortBoyard, #TopChef), Twitter est alors vu comme un prolongement de l'émission, utilisé dans une logique de « second écran ». À l'opposé d'une conception du spectateur comme passif, l'institution se place dans le paradigme de la « participation culturelle » (Fortin, 2022). La réception d'une émission est vue comme une partie intégrante de celle-ci. L'INA conserve également les hashtags liés à des événements d'actualité, considérant qu'ils sont inextricablement liés à leur représentation médiatique. La première collecte d'actualité a eu lieu suite aux attentats de novembre 2015. Elle a consisté en l'archivage des tweets contenant les hashtags « #PrayForParis », « #JeSuisParis », « #Bataclan » … (Truc, 2020). D'autres collectes ont eu lieu pour les élections présidentielles, le mouvement BalanceTonPorc, et la pandémie. Les critères de sélection sont de plusieurs ordres. D'abord, T. Anton assure un monitoring des sujets en « tendances », à l'aide d'un site externe qui les recense. Les principaux « *trending topics* » France sont alors sélectionnés. Ensuite, selon les hashtags, un outil est utilisé afin d'estimer le pourcentage de tweets en français. Cette étape a par exemple conduit à renoncer à l'archivage du #MeToo, utilisé dans de nombreux pays, dont la France, mais dont la majorité des tweets n'était pas en français. Malgré un processus de sélection formalisé, l'« *appréciation de l'archiviste* », selon les mots de T. Anton, reste cruciale. Pour éviter une dépendance totale à l'algorithme de Twitter, des hashtags moins visibles, mais pertinents, sont parfois sélectionnés, tandis que ceux artificiellement populaires (boostés par une communauté politique) sont écartés.

En plus de ces collectes courantes, la période de la crise sanitaire pouvant être qualifiée d'événement d'actualité fortement médiatisé et commenté, une collecte Twitter a été mise en place. Par le travail de veille de T. Anton, le premier hashtag à avoir été archivé l'a été précocement (bien avant l'annonce des divers confinements). En effet, #Coronavirus a commencé à être collecté le 22 janvier 2020. S'ensuivent les #Viruschinois (23/01/20, jusqu'au 30/06/2020), #CoronavirusEnFrance (27/01/20-28/04/20), #JeNeSuisPasUnVirus (29/01/20-29/02/20). À partir du 16 mars, date de l'annonce du premier confinement par Emmanuel Macron[14], de nombreux mots-clés sont ajoutés à la collecte, parfois en lien avec des émissions de radios ou de télés (elles-mêmes archivées par la collecte courante), parfois non (figure 2 page suivante).

De nombreux autres ont été ajoutés progressivement, en fonction des « *trending topics* » de Twitter ou des événements médiatiques spécifiques. Beaucoup de chaînes de télé et de radio ont en effet diffusé des émissions spéciales. Lorsque celles-ci ont donné naissance à un hashtag singulier, il a également été conservé (par exemple « #LCIVousDonneLaParole » ou « #ParolesDeSoignants »).

En tout et pour tout, 149 hashtags liés à la pandémie ont fait l'objet d'une collecte entre janvier et juillet 2020. Ils ont été répartis par des membres de l'INA, a posteriori, en cinq grandes catégories : Confinement, Déconfinement, Épidémie, Soignants, Hors-Périmètre[15]. Les tweets qui leur sont attachés sont variés dans leurs contenus. Outre des

12 Par exemple le blog de l'œnologue Yohan Castaing, https://www.anthocyanes.fr/ qui a publié des articles sur le report de la semaine des primeurs en 2020 à cause de la pandémie, ainsi que des podcasts avec des professionnels devant faire face au confinement.

13 Documentaliste multimédia durant la pandémie, devenue chargée de mission en documentation pour la valorisation des collections Web du dépôt légal depuis.

14 https://www.elysee.fr/emmanuel-macron/2020/03/16/adresse-aux-francais-covid19

15 La liste complète des hashtags peut être trouvée ici : https://f-origin.hypotheses.org/wp-content/blogs.dir/3864/files/2020/10/INADlWeb-Etude-Twitter-Coronavirus.pdf

Figure 2. Tableau des dates de début et de fin des premiers hashtags collectés en lien avec le coronavirus

Hashtag	Date de début de collecte	Date de fin de collecte
#CoronairusFR	13/03/2020	16/03/2020
#RestezChezVous	16/03/2020	28/04/2021
#CarlaBruni	16/03/2020	28/04/2021
#Irresponsables	16/03/2020	15/09/2021
#JeNiraiPasVoter	16/03/2020	26/04/2022
#DistanciationSociale	16/03/2020	16/03/2022
#ResterChezSoi	16/03/2020	30/01/2022
#OnVousRépond	16/03/2020	30/01/2020
#ConfinementTotal	16/03/2020	24/04/2021
#RESTERCHEZVOUS	16/03/2020	26/07/2021
#StayTheFHome	16/03/2020	05/03/2021
#Stade3	16/03/2020	26/07/2021
#Coronapocalypse	16/03/2020	02/02/2022
#StayHome	16/03/2020	26/07/2021
#RestezChezSoi	16/03/2020	26/07/2021
#Restezàlamaison	16/03/2020	26/07/2021

témoignages individuels sur sa situation ou celle de ses proches, ces hashtags sont également le lieu de conflits politiques ; la collecte Twitter apparaît donc comme complémentaire des collectes courantes en permettant davantage de restituer la conflictualité de cette période, et des différentes polémiques qui se sont fait jour.

Ce nombre élevé de hashtags illustre la particularité de la période Covid sur Twitter. Pour les événements précédemment mentionnés, les collectes se contentaient d'un petit nombre de hashtags (moins d'une dizaine pour les attentats, mais très fortement utilisés ; le mouvement « #balancetonporc » lui, s'est structuré autour de ce hashtag qui joue le rôle d'un mot d'ordre). Les hashtags liés au Covid et aux mesures de confinement se renouvellent et changent très rapidement. La difficulté pour l'archivage est alors qu'il faut suivre en permanence les outils de monitoring et être très réactif. Il y a un risque d'être toujours un peu en retard, de manquer une partie des premiers tweets, bref, de courir en permanence après des sujets en tendances éphémères. Des mots-clés ont probablement été manqués ou collectés très en retard, lorsque ceux-ci sont apparus un jour de week-end, lorsque la documentaliste ne travaille pas. La multiplication des mots-clés à cette période conduit également à interroger les méthodes habituelles de sélection. En effet, l'intensification des échanges numériques fait que la quasi-totalité des hashtags inventés et/ou utilisés durant les premiers mois de confinement font référence ou sont liés à cette période. Tous n'ont pas atteint pour autant les « *trending topics* ». Là encore, l'archiviste joue donc un rôle essentiel de détection et de sélection. Un exemple de hashtags mineur en termes de nombre de tweets, mais sur lequel nous avons voulu travailler (et archivé par l'INA) : le 18 mars 2020, le compte des Archives des Vosges poste un tweet indiquant *« Nous vivons un épisode exceptionnel, qui est déjà l'Histoire. Participez à la collecte #memoiredeconfinement ! Envoyez par mél à vosges-archives@vosges.fr vos témoignages, récits, photos (pdf et jpg 200ko max) ou vidéos (20 Mo), nous les conserverons pour l'éternité ! »*. Le hashtag ne générera qu'un petit nombre de tweets et de retweets (autour de 600), mais l'initiative aura un retentissement important pour le milieu professionnel des archivistes. L'avoir archivé offre donc une source précieuse pour les chercheurs s'intéressant à ces questions.

« Ce n'est pas possible de savoir ce qui va faire le buzz. Ce n'est pas possible de savoir s'il y a un sujet, quels sont les hashtags qui vont être utilisés pour un sujet. Donc c'est souvent un travail qu'on fait, qu'il faut faire rapidement. Et certainement, on a dû collecter des choses qui étaient, avec du recul, moins pertinentes que d'autres. […] Il y a une part de subjectivité qui est évidente. Malheureusement, on ne peut pas être complètement objectif. Et être exhaustif » (Jérôme Thièvre).

Qu'il s'agisse de l'archivage de sites comme de l'archivage de tweets via l'API de Twitter, l'institution se place dans une logique des archives Web comme données (Schafer et Gebeil, 2023). Le volume important de ces données, le nombre très élevé de contenus en lien avec la pandémie et les confinements signifient que, pour devenir et faire mémoire, ces données doivent faire l'objet d'un important travail de sélection et de mise en corpus *a posteriori*. Plus que le risque de manques, le problème tient à ce qui relève du « Covid ». Certains *trending topics* ne sont pas dans la collection « Covid », mais lui sont liés. Le confinement est un fait social total, donc tout ce qui s'est passé peut s'y rattacher.

Conclusion

Le Covid-19 a marqué une accélération de la documentalité contemporaine (Ferraris, 2021). Elle a entraîné le développement massif d'initiatives de collecte *ad hoc*, en temps réel, et ce dans toutes

les parties du monde. Ces initiatives ont émané de citoyens, d'universitaires, de conservateurs de musée, d'acteurs locaux et, bien sûr, d'archivistes. Si ce que d'aucuns ont qualifié de « *Covid memory boom* » a déjà fait l'objet de nombreuses recherches (Fridman et Gensburger, 2024), les archives du Web du Covid-19 ont, elles, finalement peu fait l'objet de travaux comme si elles étaient le résultat d'un processus fluide et continu de mise en collection de l'activité digitale dans son ensemble. L'exemple des archives du Web liées à la pandémie illustre non seulement la réactivité des institutions comme la BnF et l'INA face à une crise inédite, mais a aussi montré que, comme pour toute archive, futures données et sources, la fabrication des archives du Web du Covid-19 est le produit d'un contexte institutionnel et de dynamiques organisationnelles spécifiques qu'il est indispensable de documenter pour comprendre les cadres sociaux de cette mémoire du futur que constituent ces collections numériques singulières.

En documentant les modalités de construction des mémoires lors des collectes numériques liées au Covid-19 créées dans le cadre du dépôt légal, cet article peut également constituer une piste de réflexion sur les politiques institutionnelles mises en place pour prévenir les risques d'oubli et définir de manière préventive l'héritage de certains événements qui sont aujourd'hui de plus en plus vécus comme se déroulant en partie sur le Web et les réseaux, lieux privilégiés de l'expression des opinions et des sentiments, mais aussi supports des actions et des initiatives devenues numériques. La question qui demeure est celle de l'accès et des modalités d'utilisation de ces archives. En effet, à la masse importante qu'elles représentent s'ajoute l'hétérogénéité des contenus, mêlant textes, images, vidéos… Il est donc impératif de développer des outils et des applications capables de faciliter la navigation et l'exploitation de ces archives, afin que cette richesse documentaire puisse véritablement servir la recherche (Delannay, Severo et Gabrysiak, 2024) et la constitution d'une « mémoire collective » de l'événement. ◉

Bibliographie

- Bachimont B., 2017, *Patrimoine et numérique. Technique et politique de la mémoire*, Bry-sur-Marne, INA.
- Bayet A. et Hervé N., 2020, « Étude. Information à la télé et coronavirus : l'INA a mesuré le temps d'antenne historique consacré au Covid-19 », *La revue des médias*. En ligne : https://larevuedesmedias.ina.fr/etude-coronavirus-covid19-temps-antenne-information
- Bermès E., 2020, L*e numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)*, thèse de doctorat, Paris, École nationale des chartes.
- Boukadida H., Blanckemane B. et Thièvre J., 2023, « Le dépôt légal du Web à l'Institut National de l'Audiovisuel », *Les Cahiers du numérique*, vol. 19, nº 1, p. 35-58.
- Brügger N., 2012, « L'historiographie de sites Web : quelques enjeux fondamentaux », *Le Temps des médias*, vol. 1, nº 18, p. 159-169.
- Brügger N., 2019, « Understanding the Archived Web as a Historical Source », dans N. Brügger et I. Milligan (dir.), *The SAGE Handbook of Web History*, SAGE Publications, p. 16-29.
- Delannay R., Severo M. et Gabrysiak L., 2024, « Mémoires du Covid-19 et archives du Web : preuve de concept pour une analyse quantitative du dépôt légal de la BNF », *Humanités numériques*, nº 9. En ligne : https://journals.openedition.org/revuehn/3955
- Faye A., 2020, « Les archives Web du Coronavirus : une entreprise collective », *Web Corpora*. En ligne : https://Webcorpora.hypotheses.org/856
- Ferraris M., 2021, *Documentalité : pourquoi il est nécessaire de laisser des traces*, Paris, Cerf (coll. Passages).
- Fortin A., 2022, « Sur le concept de pratiques culturelles », dans M.-C. Lapointe et G. Pronovost (dir.), *Les enquêtes sur les pratiques culturelles. Mesures de la culture au Québec et ailleurs dans le monde*, Québec, Presses de l'Université de Québec, p. 147-162.
- Fridman O. et Gensburger S., 2024, *The COVID-19 Pandemic and Memory. Remembrance, commemoration, and archiving in crisis*, New York, Palgrave Macmillan.
- Gabrysiak L. et Gensburger S., 2022, « Who are the Witnesses of the Covid-19 Lockdown? The Case of France », *Witnessing memory and crisis*, Amsterdam University Press, vol. 1, p. 11-19.
- Gebeil S., Schafer V., Benoist D., Faye A. et Tanesie P., 2020, « Exploring special Web archive collections related to COVID-19 : The case of the French National Library (BnF) », *WARCnet Papers*.
- Geeraert F. et Bingham N., 2020, « Exploring special Web archives collections related to COVID-19: The case of the IIPC Collaborative collection », *WARCnet Papers*.
- Genin C., 2018, « 20 ans d'archivage du Web », *Biens Symboliques / Symbolic Goods*, nº 2. En ligne : https://journals.openedition.org/bssg/271
- Halbwachs M., 1994 [1925], *Les cadres sociaux de la mémoire*, Paris, Albin Michel (coll. Bibliothèque de l'Évolution de l'Humanité).
- Holownia O., Geeraert F., Grotke A., Harbster J. et Nagashybayeva G., 2022, « Exploring special

Web archives collections related to COVID-19: The case of the Library of Congress », *WARCnet Papers*.

- Mazzucchelli F. et Panico M., 2021, « Pre-emptive memories: Anticipating narratives of Covid-19 in practices of commemoration », *Memory Studies*, vol. 14, nº 6, p. 1414-1430.
- Musiani F., Paloque-Bergès C., Schafer V. et Thierry, B., 2020, *Qu'est-ce qu'une archive du Web ?*, Marseille, OpenEdition Press (coll. Encyclopédie numérique). En ligne : https://books.openedition.org/oep/8713?lang=fr
- Mussou C., 2012. « Et le Web devint archive : enjeux et défis », *Le Temps des médias*, vol. 2, nº 19, p. 259-266.
- Poels G. et Lefort V., 2020, « Étude INA. Covid-19 dans les JT : un niveau de médiatisation inédit pour une pandémie », *La revue des médias*. En ligne : https://larevuedesmedias.ina.fr/pandemie-covid-19-coronavirus-journal-televise
- Schafer V., 2022, « Préserve-moi ! Des journaux intimes à ceux de confinement dans les archives du Web », *Le Temps des médias*, vol. 1, nº 38, p. 175-194.
- Schafer V., Thievre J. et Blanckemane B., 2020, « Exploring special Web archives collections related to COVID-19: The case of INA », *WARCnet Papers*.
- Schafer V. et Gebeil S., 2023, « Des archives du Web aux données », *Balisages*, nº 6. En ligne : https://journals.openedition.org/balisages/1066
- Schafer V. et Winters J., 2021, « The values of Web archives », *International Journal of Digital Humanities*, vol. 2, p. 129-144. En ligne : https://doi.org/10.1007/s42803-021-00037-0
- Severo M. et Gensburger S., 2024, « Collecting traces of the outside world: an alternative collective memory of the lockdown », *International Journal of Heritage Studies*, vol. 30, nº 4, p. 385-403.
- Truc G., 2020, « Le 13-Novembre sur Twitter : de l'information à la compassion », *La revue des médias*. En ligne : https://larevuedesmedias.ina.fr/13-novembre-attentats-twitter-hashtags-reaction