Revue COSSI

N° 11 | 2022
Préservation de la mémoire du Web en temps de crise

Archiver le web durant la covid-19 et le premier confinement : organisation, bilan et perspectives

Alexandre FAYE *Conservateur*
Département du dépôt légal
Service du dépôt légal numérique
Bibliothèque Nationale de France

Édition électronique :
URL :
https://revue-cossi.numerev.com/articles/revue-11/30-archiver-le-web-durant-la-covid-19-et-le-premier-confinement-organisation-bilan-et-perspectives
DOI : numerev_1832
Date de publication : 10/01/2022

CertiScience© *Certifié évalué par les pairs*

Cette publication est sous licence **CC BY-NC-ND** (Attribution - No commercial - No derivatives).

Pour **citer cette publication** : FAYE, A. (2022) Archiver le web durant la covid-19 et le premier confinement : organisation, bilan et perspectives. *Revue COSSI*, (11). https://doi.org/10.34745/numerev_1832

Inédite par son ampleur et sa durée, la collecte web de la Covid-19 menée par la BnF a été une entreprise réellement collective déclinée à plusieurs niveaux, international et national. Les archivistes web de la BnF et de son réseau de partenaires territoriaux se sont fortement investis durant le premier confinement pour couvrir tous les aspects de la crise sanitaire et offrir une archive web représentative de cette période. La gestion de la collecte dans la durée ne fut toutefois pas sans difficulté, tout comme sa valorisation et en particulier son indexation. Si des supports de médiation ont rapidement pu être proposés, comme le parcours guidé « L'épidémie de covid-19 et le premier confinement », la BnF et ses équipes ont souhaité associer les chercheurs à des projets d'expérimentation de nouveaux outils plus adaptés à l'exploration des données issues des collectes. L'hyperconnexion des français durant le confinement consolide la place des archives web dans notre mémoire collective, ce qui conduit également à s'interroger sur leur représentativité, les possibilités de comparaison international et les possibilités d'analyse d'un matériau aussi riche. La mise en place de méthodologie permettant d'articuler analyse quantitative et qualitative devient clairement un enjeu pour les années à venir, tout comme le développement d'une communauté de professionnels et de chercheurs autour des archives du web.

Mots-clés :
Dépôt légal, Covid-19, Mémoire collective, Archives du web, Collecte

---

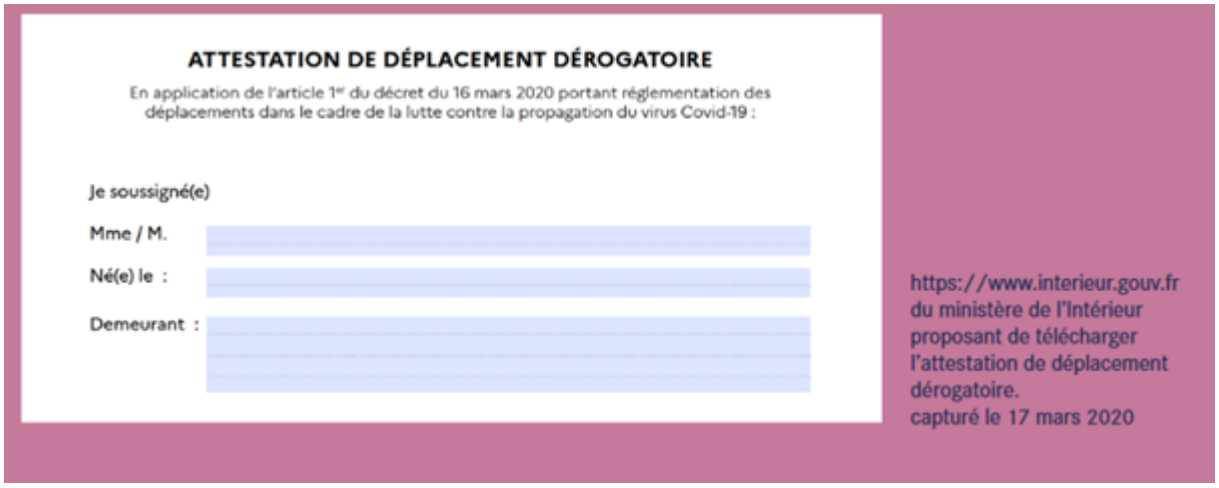
> *Françaises, français, mes chers compatriotes, depuis quelques semaines notre pays connait sa plus grave crise sanitaire, le covid19. Je m'appelle Cédric et je suis coach sportif, passionné, et vous propose tous les jours sur cette page des entrainements de 30 minutes que vous allez pouvoir faire à la maison.[1]*

Cette allocution parodique de lancement d'une chaîne de coaching sportif fait partie des contenus publiés sur le web lors du confinement de mars 2020 et archivés par la BnF dans le cadre de sa collecte Covid-19. Elle nous rappelle la diversité des initiatives et contenus qui furent publiés sur le web au moment où les interactions humaines réelles étaient strictement limitées.

De nombreuses institutions patrimoniales ont ressenti une urgence et se sont fait un

devoir de collecter et conserver les traces de ce premier confinement. Elles ont pris l'initiative de lancer des appels aux dons pour recueillir le témoignage écrit des Français comme le firent les archives départementales des Vosges[2] ou bien des objets marqueurs de cette période (cartes de périmètre de sortie, calendriers de confinement…etc) comme le proposa le Mucem[3]. Le Museum of London[4] ou en France le chercheur Hervé Mazurel[5] se mirent même à collecter les rêves.

Cet article propose de revenir sur la collecte Covid-19 menée par la Bibliothèque nationale de France dans le cadre du dépôt légal de l'internet. Celle-ci s'est déroulée dans un moment où le basculement de la vie sociale vers le numérique s'est fortement accentué du fait du confinement, le web pouvant apparaître plus encore qu'en temps normal comme un objet marqueur d'une époque. Cette présentation rappelle brièvement l'ampleur de ce phénomène, avant de revenir sur l'implication des équipes de la BnF et de son réseau de partenaires dans le processus de collecte, sur les actions de valorisation mises en œuvre à destination du grand public et des chercheurs de conclure par une interrogation sur les apports et conséquences de cette collecte dans les pratiques professionnelles des archivistes de l'internet.


ATTESTATION DE DÉPLACEMENT DÉROGATOIRE

En application de l'article 1er du décret du 16 mars 2020 portant réglementation des déplacements dans le cadre de la lutte contre la propagation du virus Covid-19 :

Je soussigné(e)

Mme / M.

Né(e) le :

Demeurant :

https://www.interieur.gouv.fr du ministère de l'Intérieur proposant de télécharger l'attestation de déplacement dérogatoire.
capturé le 17 mars 2020

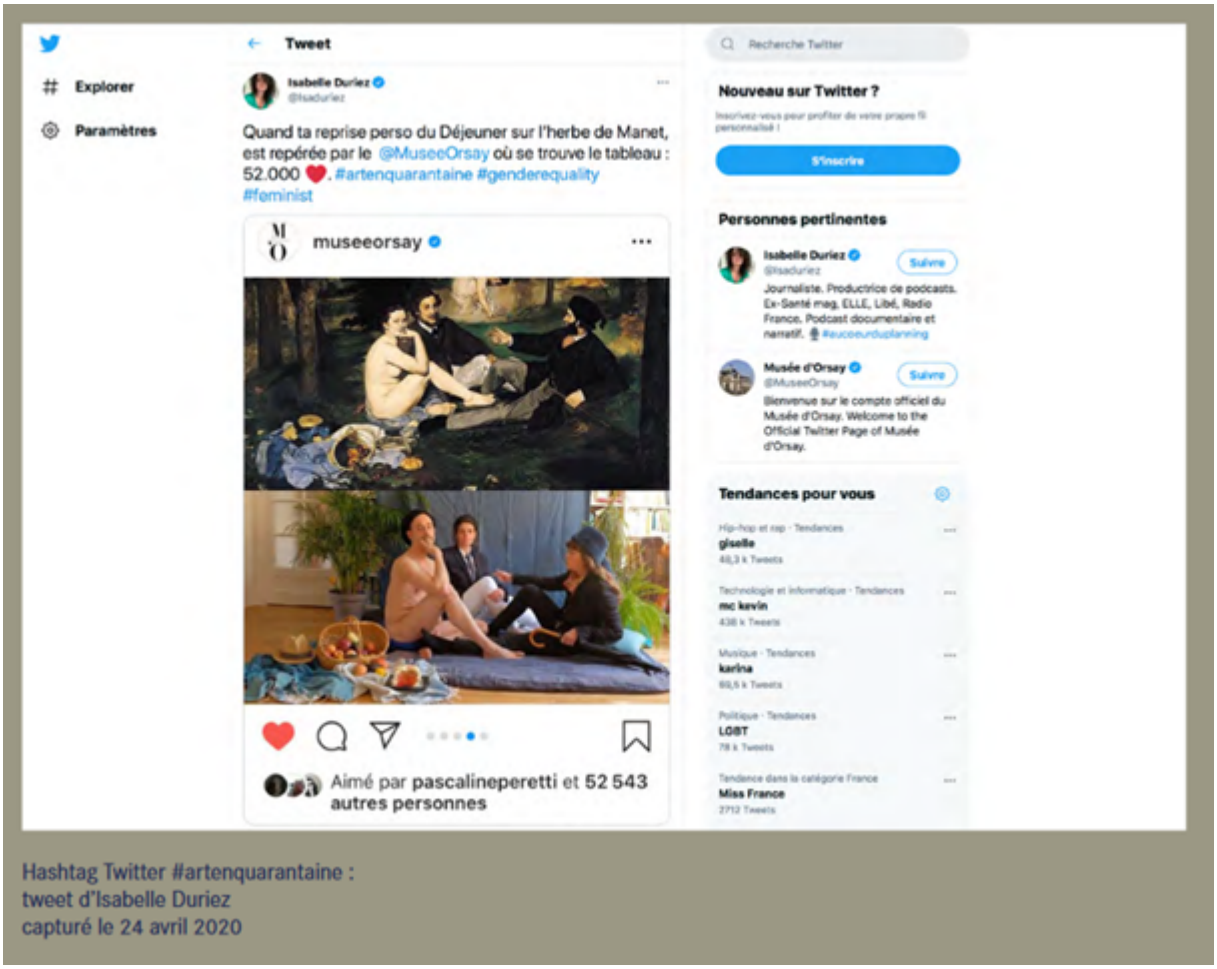
Un temps de pandémie ultraconnecté

La crise sanitaire de la Covid-19 fut marquée dans de nombreux pays par la récurrence de périodes de confinement visant à limiter les contacts directs entre personnes. En France, un confinement national fut annoncé par le président Emmanuel Macron dans son intervention télévisée du 16 mars 2020. Le report d'une grande partie des activités quotidiennes vers le monde numérique fut immédiat, en particulier pour les interactions sociales et les relations de travail : 40% des internautes français déclarent avoir créé un compte sur un réseau social ou une application de messagerie instantanée pendant le confinement[6]. Le nombre de télétravailleurs a atteint 41 % des salariés[7].

La Covid-19 a immédiatement saturé l'espace médiatique traditionnel, comptant pour plus de 80% du temps d'antenne des journaux télévisées[8] et étant mentionnée dans 60% des articles publiés de la presse régionale durant le mois d'avril 2020[9]. Le phénomène s'est étendu à l'ensemble du web et aux réseaux sociaux avec un record

de 1,4 million de tweets comptés par l'INA pour la journée du 17 mars et une augmentation de 15% du temps passé chaque jour sur internet en 2020[10].

Face à cette situation inédite, les Français ont investi le web et les réseaux sociaux pour s'entraider, pour se cultiver, pour s'éduquer mais aussi pour s'exprimer ou exprimer leur solidarité avec les soignants dans le prolongement des applaudissements qui avaient lieu chaque jour à 20h. De nombreux exemples attestent d'un véritable confinement créatif[11] et une multitude d'initiatives furent prises dans un esprit Do-it-Yourself : mise en ligne de tutoriels pour créer son masque, vidéos de coaching sportif, circulation de memes et réalisation de défis créatifs sur les réseaux sociaux comme avec le hashtag #artenquarantaine. Le web se fit également le relais des débats politiques sur la pénurie de masques et de gel hydroalcoolique, ainsi que des controverses sur l'usage de l'hydroxychloroquine prôné par le professeur Didier Raoult. Pour s'adapter à la situation, de nombreux organismes, associations ou encore salles de spectacle ont utilisé le web et les réseaux sociaux pour garder un contact avec leurs adhérents, leur public ou plus largement les internautes. Les hôpitaux, mairies, agences de santé, instituts de recherche ont créé des pages d'information dédiées et parfois de véritables plateformes documentaires comme celle mise en place par l'Assistance publique des hôpitaux de Paris (AP-HP) et ont mis directement en ligne les fiches médico-sanitaires de leurs propres équipes.



Hashtag Twitter #artenquarantaine :
tweet d'Isabelle Duriez
capturé le 24 avril 2020

Lancement d'une collecte internationale

Les archivistes de l'internet se devaient de saisir l'ensemble de ce déferlement « Covid-19 ». Considérant immédiatement l'ampleur mondiale du phénomène, l'IIPC sous l'impulsion du Content Development Group (CDG) et en partenariat avec Archive-It décida de lancer une collecte internationale. Les membres du consortium eurent la possibilité d'alimenter un tableau excel en sélections documentaires dès le 16 février 2020. Les consignes initiales données par le CDG, avant même les premiers confinements en Europe, structurèrent fortement les premières sélections en proposant un ensemble de catégories pour qualifier les sélections. Outre les aspects médicaux et scientifiques, la collecte se proposait de couvrir l'ensemble des aspects que pouvait prendre la crise sanitaire (aspects sociaux, aspects économiques, aspects politiques) et accordait une place importante aux mesures d'endiguement prises localement (sous une étiquette « Regional or local containment efforts »).

La BnF répondit à l'initiative du consortium en apportant trois jeux de données durant l'année 2020, représentant un total 591 sélections dont les archives sont désormais consultables en ligne sur le site d'archive-it. Mais cet ensemble ne représente qu'une petite partie, une sélection dans la sélection, de ce qui allait devenir la collecte Covid-19.

En effet, après l'annonce du confinement de mars 2020 et face à l'ampleur des conséquences de la crise sanitaire, il est apparu évident que la collecte en cours devait et allait prendre une ampleur nouvelle. La présidence de la BnF a de fait rapidement souhaité mettre en place une collecte collaborative d'envergure en mobilisant le réseau des partenaires. Les objectifs documentaires restèrent proches de ceux de la collecte IIPC, tout en étant élargis aux réseaux sociaux (principalement Twitter) et aux chaines vidéos (Youtube), et ils furent repris à travers un « Memento pour la collecte en rapport avec la crise sanitaire du coronavirus (Covid-19) ».

La coopération nationale alimenta en retour les contributions de la BnF au projet d'archivage international Novel Coronavirus (2019-nCoV) outbreak[12] par des sélections de contenus de médias et sites locaux aussi bien pour les régions métropolitaines que pour les Outre-mer.

Premières sélections

Les premières sélections menées par les archivistes[13] de la BnF ont été faites fin janvier 2020, alors que la mégapole de Wuhan était mise en quarantaine par les autorités chinoises. Les hashtags #JeNeSuisPasUnVirus et #CORONAVIRUSENFRANCE firent leur apparition sur Twitter pour dénoncer la stigmatisation vécue par les membres de la communauté asiatique en France. Le Mouvement contre le racisme et pour l'amitié entre les peuples (MRAP) publia rapidement une page « un virus n'a pas d'origine ethnique » sur son site web. Ces publications font partie des premiers contenus archivés.

Tout au long de la crise sanitaire, la réactivité a été un élément essentiel de la réussite de cette collecte. Contrairement aux autres collectes projets préparés très en amont, il était impossible pour les archivistes de prévoir une date de début et de fin de collecte, ni de connaître à l'avance les réactions et les mesures sanitaires à chacune des phases de l'épidémie[14]. Cette réactivité est apparue d'autant plus nécessaire que certaines pages web ont pu disparaitre rapidement ou connaitre des mises à jour régulières.

Pour établir leurs sélections, les archivistes utilisent de sélection BCweb développée par la BnF, qui présente l'avantage full web et donc accessible aussi bien sur son lieu de travail que depuis son domicile en télétravail. Cet outil permet aux archivistes de gérer différentes collectes, d'y saisir des sélections sous la forme de fiches documentaires et de définir le meilleur paramètre de collecte en fonction de l'objectif de collecte. Il peut par exemple s'agir de crawler l'ensemble d'un site en utilisant une profondeur dite « domaine » à une fréquence faible d' « une fois par an » ou bien de capturer plusieurs fois par jours les dernières publications liées à un hashtag Twitter.

Le dispositif technique et organisationnel retenu au sein de l'outil de sélection BCweb fut celui de la collecte de « l'Actualité éphémère » mise en place après les attentats de 2015 pour permettre une meilleure réactivité des équipes. Son objectif est de faciliter les sélections documentaires déterminées par les enjeux contemporains et l'actualité immédiate. Il faut noter dès à présent que cette collecte diffère totalement des crawls des collectes « Actualité » et « Presse payante », qui regroupent plus d'une centaine de sites de presse nationale et régionale dont la une, les articles liés et les comptes Twitter sont collectés une fois par jour. Elle a justement pour but de permettre la sélection de contenus en lien avec l'Actualité pour les sites qui ne relèvent du périmètre de la presse traditionnelle. Il peut s'agir de collecter les tweets liés à un hashtag populaire, des communiqués de presse d'association, des nécrologies, des articles de presse professionnelle ou spécialisé, des blogs personnels d'opinion…etc. Elle s'appuie sur la coordination d'un réseau de correspondants, interne à la BnF, référents pour leur discipline.

Une entreprise collective

À la suite de l'annonce du confinement, ces correspondants au réseau interne BnF se sont fortement investis. Ils pouvaient s'appuyer sur leur base de travail, c'est-à-dire la liste des sites web relevant d'une discipline et déjà collectés par ailleurs dans le cadre des collectes thématiques courantes[15], le plus souvent sur un rythme annuel[16], et y rechercher des contenus publiés en lien avec la crise sanitaire pour les sélectionner. Cette approche méthodique fut notamment utilisée par la correspondante du Département Sciences et techniques.

Dans le même temps, la collecte fut ouverte au réseau externe de correspondants régionaux, qui était déjà mobilisé par la collecte des « Élections municipales » de 2020. Ce réseau de partenaires territoriaux est une pièce essentielle du dispositif de coopération nationale. Il est constitué de 26 bibliothèques territoriales[17] et services d'archives qui partagent la mission de dépôt légal de l'internet et qui participent aux

collectes collaboratives comme celles qui touchent l'activité électorale française. Sa participation[18] s'est rapidement avérée essentielle, dans le cadre d'une pandémie qui touchait différemment les territoires, de pouvoir enrichir la collecte par des sélections de niveau régional ou local. Ainsi l'approche régionale et locale fut également au cœur de la dynamique d'archivage avec l'archivage des publications web de nombreux CHU, musées, groupes syndicaux locaux, diocèses, municipalités, académies, sans oublier des blogs et sites d'information locale qui pouvaient échapper aux collectes courantes de la BnF.

Grâce au renfort de 70 correspondants volontaires, les sélections connurent une augmentation spectaculaire entre la mi-mars et le mois de mai. Un point haut fut atteint durant le mois d'avril avec 500 sélections réalisées par semaine. En analysant les mots clés saisis par les correspondants, il apparaissait que la moitié des contenus possédaient une indexation géographique (région, département ou ville). Ainsi, la répartition de la veille et des sélections entre le réseau interne et externe permit d'obtenir une couverture disciplinaire et territoriale jugée relativement complète et donc représentative du web à cet instant.

Dans la pratique, les correspondants pouvaient s'appuyer sur le « Memento » rédigé par l'équipe du service du dépôt légal de l'internet. Dans le cadre du périmètre qui était le leur, couverture plutôt disciplinaire pour le réseau interne et régionale pour le réseau externe, ils sont restés totalement libres dans leurs choix de sélections. En analysant le résultat de la collecte pour cette période de février-juillet 2020, on note toutefois des variations du nombre de sélections d'une région à l'autre. Elle s'explique par le temps consacré par le correspondant régional à cette tâche, mais aussi par la méthodologie adoptée. L'existence d'une collecte régionale déjà en place, ce qui reste un cas minoritaire actuellement, peut faciliter la mobilisation de cette première base de travail pour la veille. L'adoption de techniques de veille méthodiques conduit également à un plus grand nombre de sélections. Ainsi, la correspondante de Bourgogne-Franche-Comté renouvelait régulièrement les mêmes requêtes en recherche avancée sur Google pour pouvoir ajouter systématiquement de nouveaux contenus pertinents pour sa région. Pour plusieurs régions comme la Corse ou le Grand Est et en particulier pour l'Alsace, qui était fortement touchée par la première vague, la BnF a pu pallier les difficultés techniques et humaines locales en réalisant directement des sélections régionales.

Il fut cependant difficile de conserver un haut niveau de mobilisation du réseau de partenaires dans la durée. Néanmoins, l'ensemble des sélections documentaires de la première vague put servir de base de travail lors des vagues épidémiques suivantes et la collecte resta toujours ouverte aux correspondants régionaux. Plusieurs d'entre eux, comme la correspondante de Provence-Alpes-Côte d'Azur, reprirent contact lors de la deuxième vague et relancèrent les sélections pour couvrir les évolutions locales de la fin 2020 et la mise en place des premiers couvre-feux.



Diversité des archives: site web, réseaux sociaux et vidéos

Dans leur sélection, les archivistes sont amenés à considérer la diversité des formats et des plateformes qui constituent le web aujourd'hui. La notion d'espace numérique[19] a ainsi été appliquée. Pour une institution comme l'Inserm, ce sont à la fois le site web principal, la salle de presse (presse.inserm.fr), le compte twitter et la chaine Youtube, qui ont été archivés. Techniquement plus délicate, les collectes des vidéos Youtube et des contenus Instagram ont été réalisé séparément en juillet et décembre 2020. Quel que soit le type d'objet archivé, les outils employés en production restent les mêmes (la BnF utilise le crawler Heritrix) de façon à délivrer des fichiers de conservation répondant tous aux mêmes standards (format WARC).

Pour la collecte Vidéo, la démarche retenue consiste à capturer l'ensemble des vidéos d'une chaine et à ne pas faire de sélection au niveau de la vidéo. Dans la mesure du possible, les équipes techniques de la BnF s'efforcent de capturer l'environnement des pages web des vidéos (onglets À propos, Communauté) ou à défaut des métadonnées permettant de recréer l'environnement artificiellement. Très importante en termes de consommation budgétaire, cette opération requière l'adoption de critères plus stricts pour éviter que les besoins de stockage dépassent les capacités des espaces alloués sur les serveurs de la BnF. Une priorité fut donc donnée aux chaines dont le contenu était

majoritairement en lien avec la crise sanitaire et les chaines vidéo généralistes des grands médias furent écartées. 109 chaines vidéo Youtube furent archivées couvrant aussi bien la vidéo scientifique que la vidéo d'humour confiné. C'est à cette occasion que fut archivée la chaine vidéo du coach sportif de la République, citée en préambule de cet article.

Pour compléter ces sélections, les archivistes de la BnF souhaitèrent intégrer des sites web apparus durant le confinement considérant que ces créations ad-hoc témoignaient également de la réaction du web. Le travail fut mené à partir des listes produites par l'AFNIC [20], ce qui permet de repérer les nouveaux noms de domaine déposés. Dans un premier temps, un filtrage fut effectué pour ne retenir que les noms de domaine comportant les chaînes de caractère "corona", virus" et "covid". Une analyse méthodique des résultats permit de recenser 640 sites actifs répondant aux critères de la collecte. Il s'agit essentiellement de sites commerciaux (vente de visières), d'entraide (réseau de makers proposant des masques aux soignants), d'information locale ou municipale sur le virus...etc.

Bilan de la collecte après le premier confinement

Durant la première phase de l'épidémie, entre janvier et juillet 2020, les archivistes ont ainsi réalisé 5 142 sélections de sites, blogs, réseaux sociaux et vidéos correspondant à 1 014 crawls effectués par les robots de la BnF et une production de 15 504 fichiers WARC. L'ensemble représente 274 784 516 urls collectées et un volume de 15To de données compressées dont 1To de vidéo. À titre de comparaison, la collecte large qui vise à couvrir l'ensemble du domaine .fr et des sites hébergés en France atteint 110/120 To chaque année tandis que le poids des collectes projets[21] (Intelligence artificielle, Enjeux environnementaux, Jeux olympiques) est généralement compris entre 1 et 2 To.

Les mots clé qui ont guidé la sélection donnent également quelques indications de la répartition thématique des contenus. Sans surprise, les Aspects sociaux (1 531) dominent, devant les Aspects économiques (522), les Aspects médicaux et scientifiques (451), les Mesures d'endiguement (374) et enfin les Aspects politiques (146)[22]. Des mots clés plus précis permettent de retrouver plus finement certains types de publications web (journal de confinement, témoignage, éducation...etc).

Les statistiques de production témoignent de la forte mobilisation du réseau des correspondants (internes et externes) et de la maitrise de l'application de sélection BCweb. Les archivistes ont pu continuer à travailler tout en étant confiné(e)s. La collecte Covid-19 représenta d'ailleurs pour certaines équipes une occasion de garder le contact entre collègues durant le confinement et de continuer à travailler ensemble dans un contexte général de déploiement du télétravail[23].

Gestion de la collecte sur un temps long

Du fait de cette dynamique initiale, la masse des sélections généra une charge importante de gestion et de vérification pour les archivistes du service du dépôt légal numérique de la BnF. La prise en compte de l'évolution temporelle des contenus est une difficulté majeure de ce type de collecte dynamique. Lorsqu'un archiviste saisit l'url d'une sélection dans la base de travail, il lui attribue une url et complète la fiche avec certaines informations techniques et descriptives (mot-clé, notes de contenu). Il définit également une fréquence de collecte. De nombreuses sélections visaient à capturer les contenus de pages et de rubriques régulièrement actualisées. Pages de CHU, d'ARS, de municipalités, d'association, de syndicat et presse professionnelle ont relayé massivement les bilans épidémiologiques, les recommandations gouvernementales, des consignes et procédures pour le déconfinement[24] auprès de leurs lecteurs et adhérents. Elles entrent de fait dans cette catégorie.

Dès juin 2020, il devenait essentiel de désactiver une partie des sélections, revoir les fréquences de collecte, mettre à jour les urls. Pour ce dernier cas, les changements sémantiques (/coronavirus /covid_19) et l'actualité (/variant /vaccin) se sont souvent répercutés dans les adresses des pages visées. Ce travail fastidieux de vérification des urls de départ et de contrôle des résultats aussi bien technique que visuel accompagna l'évolution de la pandémie et fut poursuivi de 2020 à 2022 à chaque nouvelle vague épidémique. Grâce à cela, la base de travail créée durant le premier confinement continua de servir de réservoir aux archivistes, qui pouvaient ainsi plus aisément sélectionner et désactiver en masse des urls de départ pour les crawls à lancer. Le nombre total de sélections originales durant les deux années de crise sanitaire avoisina les 6300. Ce bilan démontre que les outils et l'organisation des collectes ont aujourd'hui atteint une certaine maturité.

Parallèlement à la poursuite de la collecte, l'objectif pour les équipes de la BnF était de maintenir l'élan initial pour mettre en œuvre la valorisation des archives produites durant le premier confinement.



Publication d'un parcours guidé de l'archive

Afin de donner une visibilité à ces archives, les équipes de la BnF entreprirent la rédaction d'un parcours guidé intitulé « L'épidémie de Covid-19 et le premier confinement ». Ces publications permettent d'ordonner la collection autour de grands thèmes et proposent un choix de captures pour chacun d'eux. Chaque capture est décrite brièvement et contextualiser. Fonctionnant comme un outil de découverte, les parcours guidés donnent la possibilité de rejouer les sites web archivés mis en avant et de naviguer dans les captures ou bien de faire défiler les captures via une barre chronologique. Ils peuvent être produits par les archivistes pour toucher le grand public ou par des chercheurs partenaires qui publient et documentent leurs sources en histoire ou sociologie du web sous cette forme. Les parcours guidés sont consultables dans leur intégralité à la BnF et au sein du réseau des bibliothèques partenaires[25]. Une version PDF de chaque parcours publié est mise en ligne sur le site de la BnF[26]. Celle de « L'épidémie de Covid-19 et du premier confinement » est accompagnée d'un diaporama réalisé avec l'accord des ayants-droits réalisé pour faciliter les réutilisations pédagogiques. Les illustrations de cet article en sont tirées.

Le parcours guidé consacré à « L'épidémie de Covid-19 et le premier confinement » fut publié en mars 2021. 16 thèmes le composent pour un ensemble de 126 captures.

Organisation thématique du parcours guidé :

1. Les mots de la Covid-19
2. Une pandémie hors de contrôle.
3. Confinement et restriction des libertés publiques
4. Le système de santé.
5. Chercher, informer, vulgariser
6. Je veux aider.
7. Une mondialisation à l’arrêt, une économie en panne.
8. Au théâtre chez vous ! Spectacles en confinement
9. Un confinement créatif
10. Le quotidien du confinement
11. Ma commune durant le confinement.
12. Une attention aux plus fragiles et vulnérables.
13. Un monde du travail bouleversé.
14. L’enseignement à distance.
15. Le déconfinement.
16. Interpréter l’évènement, se réinventer.

Le PDF et le diaporama sont accessibles sur le site de la BnF.



Une indéxation nécessairement large

En parallèle, les archives constituées furent indexée pour permettre la recherche par mot et offrir ce service aux chercheurs. Néanmoins, l'indexation de cette collection posa des problématiques de définition du périmètre de la collection. Les collections indexées précédemment correspondaient à des ensembles plus aisés à définir avec une correspondance plus évidente entre les fichiers de conservation produits et le contenu documentaire. Ainsi, les crawls des collectes « Actualité » et « Presse payante » regroupent plus d'une centaine de sites de presse nationale et régionale dont la une, les articles liés et les comptes Twitter sont collectés une fois par jour. Les fichiers produits sont clairement séparés et identifiables. Au final, l'indexation de la collecte « Actualités 2010 – 2021 »[27] est une opération essentiellement technique qui ne pose pas de difficulté documentaire.

Il est important de rappeler à ce niveau que la collecte Covid-19 est avant tout un travail de sélection documentaire. Elle ne recoupe pas l'organisation technique des collectes et surtout des crawls réalisés par les robots de la BnF. Techniquement, elle relève des collectes courantes qui sont organisées selon les paramètres de fréquence et de profondeur de collecte, ce qui signifie que les contenus conservés au sein d'un fichier d'archive sont mélangés et proviennent de différentes collectes documentaires courantes (collecte de l'Actualité éphémère, collecte Littérature et Arts, collecte Alsatiques...etc). Pour définir une collection, il était donc nécessaire de s'appuyer sur un critère chronologique. L'ensemble indexé correspond concrètement aux fichiers WARC produits entre janvier et juillet 2020 dans le cadre des collectes Actualité, Presse payante et des collectes courantes[28]. En parallèle, les chaînes vidéos collectées furent intégrées au parcours guidé « Vidéo » et la liste des sélections sur la Covid-19 fut publiée sur le site API et Jeux de données de la BnF[29].

Expérimentation de nouveaux outils: datavisualisation et appréhension des données

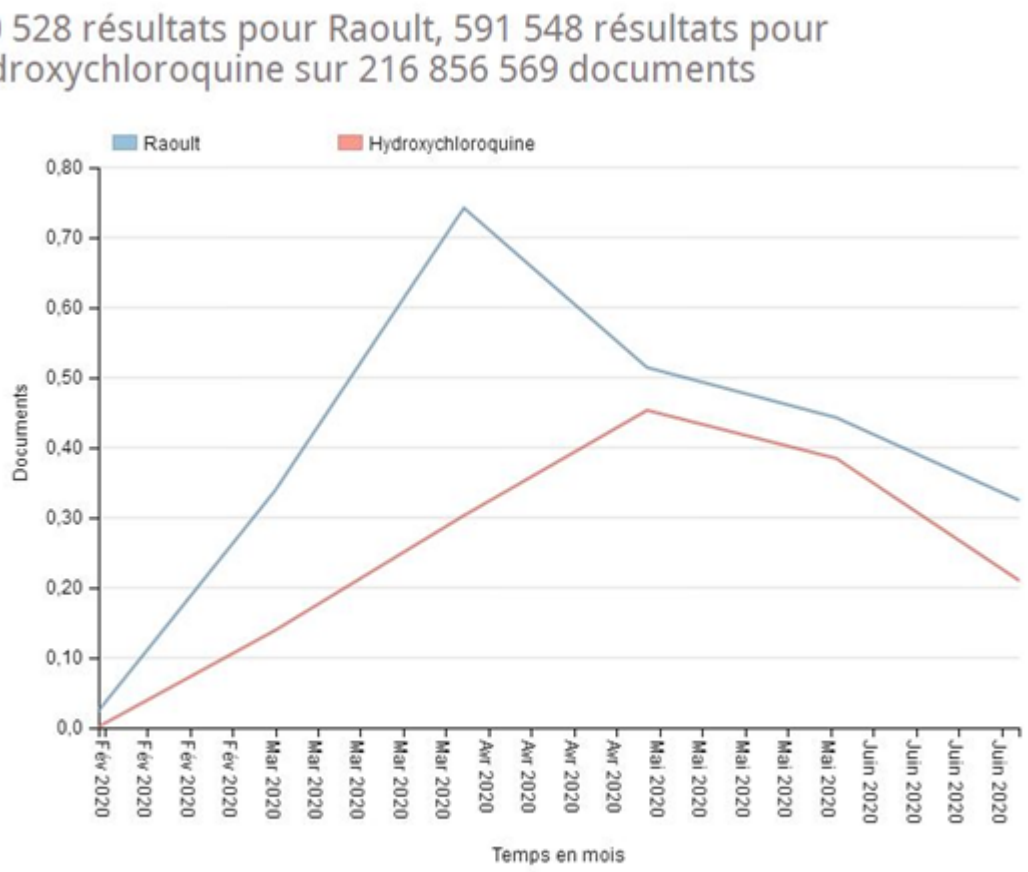
La BnF prit également l'initiative de réunir un groupe de chercheurs sous la forme de deux ateliers qui eurent lieu à la fin de l'année 2020[30]. L'objectif de ces ateliers était

de mesurer leur intérêt pour la collecte et de les associer à des réflexions sur l'évolution des outils. Ils firent en outre ressortir la nécessité de renforcer le dispositif d'accueil, ce qui conduisit à proposer les archives web de la Covid-19 comme objet de recherche[31] pour la publication du premier appel à projet du BnF DataLab.

Conscientes de la difficulté que pouvait représenter l'appréhension et l'analyse de la masse considérable de données que représentent les archives web de la covid-19, les équipes de la BnF souhaitèrent tester de nouveaux outils de datavisualisation et recueillir l'avis des chercheurs sur ces derniers. Une recherche n-grams fut développée pour permettre de comparer les occurrences de mots ou d'expressions et visualiser leur évolution temporelle. Cette fonctionnalité, même si elle peut présenter des biais d'interprétation[32], fut intégrée en septembre 2021 et fait aujourd'hui partie de l'offre standard proposée pour les collections indexées (Actualités, Attentats, Incunables du web et désormais Covid-19). L'installation de l'outil SolrWayback développé par la Bibliothèque royale du Danemark permit également de tester cette application, qui intègre de nombreuses fonctionnalités de datavisualistion et facilite les recherches d'images. Un petit échantillon correspondant à 10% de la collecte Covid-19 fut indexé avec cet outil. Jugé très concluant, il est aujourd'hui utilisé sur d'autres projets de recherche lorsque la taille du corpus le permet[33].



Nuage de mots générés avec l'outil SolrWayback.
Analyse du dns mediterranee-infection.com sur un échantillon de 10% de la collecte Covid-19



La recherche n-gram permet de visualiser l'évolution de la fréquence de mots ou d'expressions à travers le temps dans les documents archivés.

Une mémoire immédiate: l'archive web au regard de la sensibilité

Durant le confinement et la période suivante du déconfinement, plusieurs articles de presse se sont intéressés à l'archivage du web et à la collecte Covid-19 menée par la BnF. Ils témoignent d'une curiosité pour la mémoire de l'évènement et ce nouveau type d'archives encore peu familières du grand public. La collecte web y est associée collectes aux autres collectes physiques organisées par les musées et les services d'archives (collecte d'artefacts ou recueil de journaux de confinement).

La constitution d'une mémoire immédiate de l'évènement a semblé évidente tant nos vies étaient bouleversées et l'intention mémorielle peut expliquer le rapprochement avec les collectes « sensibles » évoquées dans l'introduction de cet article. L'usage des mêmes termes « collecte »et « archives » entretient d'ailleurs une certaine confusion entre des projets qui ne reposent pas sur les mêmes méthodes de constitution des collections. Le travail de l'archiviste web s'apparente plus à un travail de documentaliste, reposant sur la sélection et l'organisation des ressources en vue de leur archivage et ultérieurement de leur valorisation. Les modalités du crawl donne de fait un caractère massif à aux collectes web, puisque le robot suit les liens trouvés sur le web sans être guidé. De plus pour une seule page web, il copie également une grande quantité de fichiers de toute nature qui entrent dans la composition de la page (images,

scripts, fichiers de mise en page).

Les collectes des musées et des services d'archives reposent le plus souvent sur des appels aux dons et permettent de récolter des objets marqueurs de l'époque (masques, attestations de sortie), ainsi que des créations ordinaires personnelles (journaux de confinement, calendrier de confinement). Ces artefacts sont porteurs d'une certaine émotion et se font l'écho d'un rapport personnel à l'évènement. Néanmoins, derrière ces premières différences, des rapprochements naturels peuvent être faits entre les archives physiques et les archives web. En effet, des équivalents aux objets physiques collectés ont circulé sur le web et ont été archivés : tutoriel en ligne pour créer son propre masque, site web générant une attestation de sortie, journaux de confinement, témoignages sur les réseaux sociaux...etc. Tous relèvent du quotidien et nous sont familiers. Ces traces de nos vécus intimes de l'évènement participent aussi bien l'exercice d'une mémoire qu'à la constitution d'un patrimoine au sens proprement documentaire[34]. En ce sens, la collecte covid-19 représente une étape importante dans le renforcement de la légitimité patrimoniale des archives du web[35].



De la représentativite

L'acquisition d'une légitimité patrimoniale n'est pourtant pas universelle, même dans le

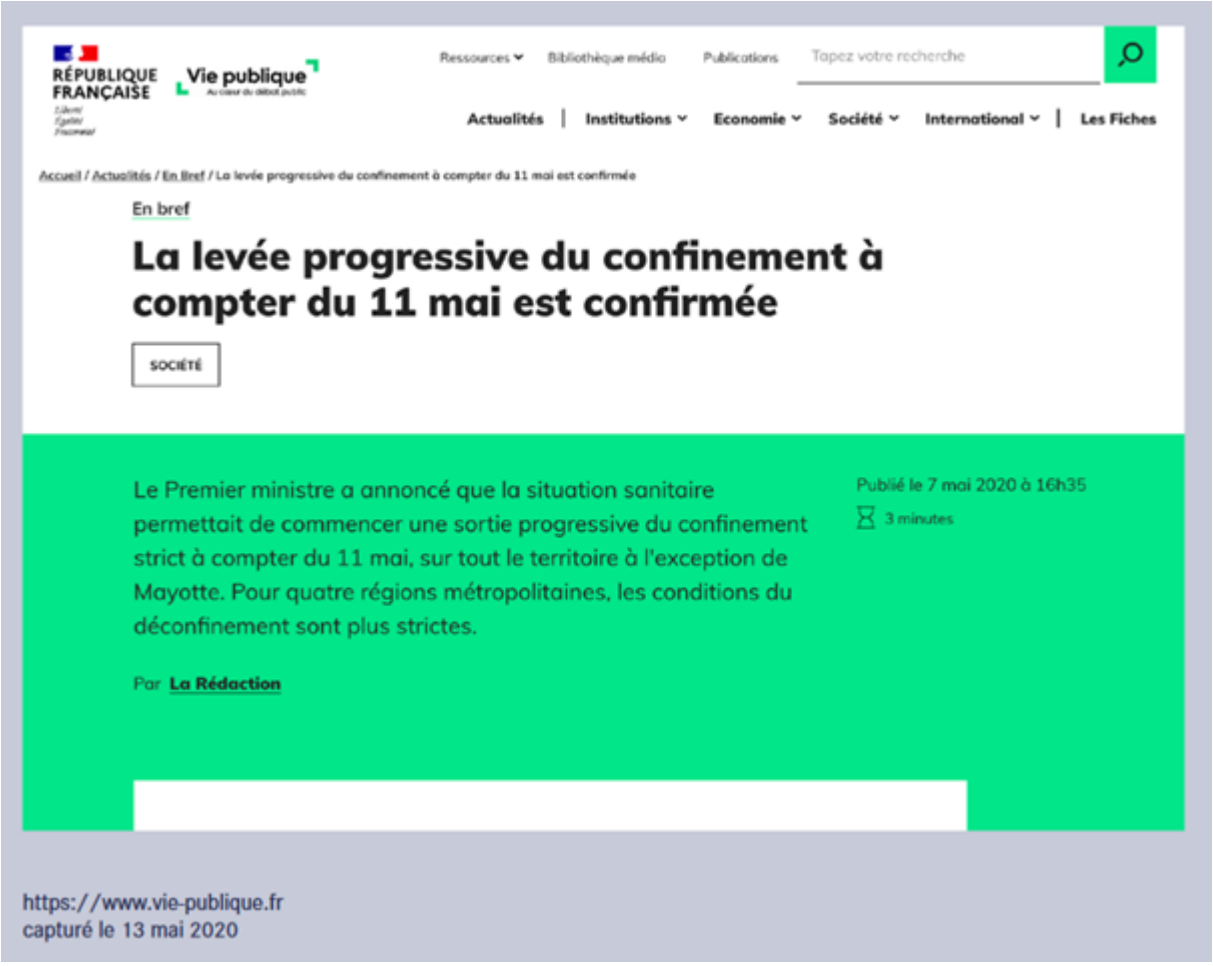
cas d'un évènement planétaire comme celui que nous venons de traverser. En changeant d'échelle, on constate que cette forme de patrimonialisation immédiate concerne essentiellement l'Amérique du Nord et l'Europe. Pour l'archivage du web, en utilisant comme référence la collecte collaborative internationale portée par l'IIPC, la carte s'élargit toutefois à l'Amérique du Sud et à plusieurs pays d'Asie. Une réelle fracture géographique persiste avec les pays les moins développés et plus encore les régimes autoritaires. Les collectes Covid-19 peuvent ainsi être considérées comme des marqueurs culturels et politiques de nos sociétés. Le plus souvent, elles invitent les citoyens à participer promouvant l'idée que l'histoire de chacun est importante. Cette ouverture à tous est néanmoins questionnable dans la mesure où elle ne saurait générer d'elle-même une parfaite égalité de représentation[36]. Le même questionnement peut être porté sur les archives du web, d'autant que la construction des politiques documentaires d'archivage pose la représentativité comme principe de directeur. Que signifie cette idée appliquée au web, espace transnational composé d'une multitude de plateformes et de formats ? La communication d'une documentation de la collecte la plus complète possible apporte de nombreux éléments de réponse à cette question. Documenter une collecte est un élément essentiel pour en apprécier les lacunes et les pertinences, permettre l'exercice de la critique. Critères et listes de sélection d'une part, statistiques et données techniques d'autre part (nombre de fichiers WARC produits, paramètres de collecte, code réponse des sites) sont les informations essentielles pour que les chercheurs puissent exercer une critique des matériaux. Il est indéniable que l'amélioration des outils de consultation et de recherche facilite l'exploitation de ces grands corpus archivistiques issus du web. Pour autant, ils ne doivent pas masquer le matériau ou conduire à déroger à cette première prise de connaissance de la source étudiée.

Comparaisons internationales: le projet warcnet

L'analyse et la comparaison des collectes web est un exercice d'autant plus essentiel dans le cadre de la pandémie, que les dispositifs nationaux de collecte mis en place par les bibliothèques nationales varient aussi bien au niveau technique qu'organisationnel. Les formats et les outils, même s'ils sont largement partagés par la communauté des archivistes, ne sont pas toujours employés avec les mêmes configurations. Dans le cadre du projet Warcnet, la chercheuse Valérie Schafer a ainsi mené une série d'entretiens pour éclairer ces collectes avec l'objectif de pouvoir entamer des comparaisons entre pays européens.

Cette série d'interviews illustrent la diversité des paysages juridiques nationaux et des choix techniques et documentaires[37]. Elle aide les archivistes à prendre du recul sur leurs pratiques de veille. La collecte de chaines vidéo Youtube ressort comme une spécificité forte de la BnF, même si l'Ina a également collecté une douzaine de chaines en lien avec la pandémie[38]. L'approche des réseaux sociaux diffère d'une institution à l'autre. En France, l'Ina s'appuie sur les API Twitter et récupère des millions de données bien formatées propres à des analyses massives[39]. La BnF a opté pour une approche plus « photographique » permettant de restituer la page à la date de la capture et donc

le contexte de publication des tweets à un instant t. Cependant seuls les derniers tweets sont capturés ce qui ne permet pas d'avoir une archive exhaustive. Ces différences ont de profondes implications pour les chercheurs, qui peuvent dès lors s'orienter vers une collection plutôt qu'une autre, tandis que les archivistes peuvent considérer que ces approches s'avèrent complémentaires. Suite à cette prise de conscience, la BnF a souhaité poursuivre ce type d'échange méthodologique et comparatif avec ses partenaires territoriaux en organisant le 9 novembre 2021 un webinaire consacré aux collectes régionales du web[40].



Temporalité appliquée à la collection

L'appréhension d'une archive produite sur une durée aussi longue reste toutefois difficile. Par nature une archive renvoie à une multitude de temporalités. La date qui prévaut lorsqu'on réalise une recherche est la date d'archivage, qui est nécessairement postérieure à la date de publication. Il est de fait difficile de reconstituer les temporalités originales de publication et de circulation des contenus. La navigation en elle-même au sein de l'archive web conduit à des sauts temporels, minimes ou au contraire importants, lorsqu'on passe d'une page à une autre en fonction de l'écart temporel entre les deux captures[41]. Dans le cadre de la collecte Covid-19, l'évolution des sélections dans le temps complexifie encore l'appréhension de la temporalité des

contenus. Une page web peut avoir connu de multiples mises à jour durant le confinement sans que toutes ses versions aient été capturées et à l'inverse une page peut avoir donné lieu à des captures régulières sans modification des contenus. Surtout, les sélections se sont poursuivies après l'été 2020 et des sélections rétrospectives peuvent avoir été faites longtemps après la fin du premier confinement. Les contenus ne sont alors pas inclus dans l'index de recherche par mot-clé. Dans le même ordre d'idée, il semble difficile d'envisager une indexation de l'ensemble des archives produites entre février 2020 et la vague omicron de janvier 2022. La difficulté est à la fois technique (capacité machine) et conceptuelle (pourrait-on encore parler de collecte Covid-19 si le bruit généré par d'autres contenus s'accroit ?).



Analyse quantitative et analyse qualitative

De fait, la collecte covid-19 de par sa masse résume à elle seule la problématique que rencontre désormais la communauté des archivistes et des chercheurs. Pour reprendre les termes de l'historien Ian Milligan : We can collecte all this data, but what happens when it comes time to analyze it ?[42] Le BnF DataLab entend répondre à ce défi. Inauguré le 18 octobre 2021, cet espace propose un service d'assistance et d'accompagnement à la recherche en partenariat avec la Très grande infrastructure de recherche Huma-Num (CNRS). Pour sa première année d'existence, il accueille 5 projets pour une durée d'un an, dont un consacré à la viralité sur le web porté par Valérie Schafer. L'offre de service concernant les archives du web a été pensée à partir de l'expérience acquise depuis 2002 et est structurée autour de trois besoins : la collecte à

la demande, l'aide à la fouille de données et l'extraction de métadonnées et d'archives. C'est dans ce cadre que de nouvelles approches sont explorées pour articuler analyse quantitative et qualitative grâce à des méthodes d'échantillonage, de cartographie exploratoire ou de traitement en masse des urls.

À côté de ces travaux qui relèvent des humanités numériques, les approches plus classiques reposent elles aussi sur une méthodologie de plus en plus formalisée. Dans son ouvrage Web site Stories, la chercheuse Sophie Gebeil[43] expose ainsi sa démarche. Le travail critique des archives web est enrichi par la réalisation d'enquêtes auprès des producteurs et l'établissement de comparaisons avec les autres médias audiovisuels (TV, films documentaires, militants) et apportent une profondeur historique à l'analyse. Le projet Mémoires du Covid-19 et archives du Web porté par Sarah Gensburger devrait aussi continuer à enrichir ses réflexions en interrogeant la relation entre histoire et mémoire numérique de l'évènement vécu.

Nouvelles perspectives: le projet ResPadon

La collecte Covid-19 a aussi suscité de nombreuses questions de la part des professionnels de l'information et du monde académique éveillant un intérêt nouveau pour ce type de sources encore méconnues. Un des objectifs des prochaines années est la valorisation de cette offre documentaire au sein des universités. C'est le sens du projet ResPadon (Réseau pour développer et diversifier les usages des archives du web) lancé conjointement par la BnF, l'Université de Lille, le campus Condorcet et Sciences Po[44]. Le projet repose sur l'expérimentation d'un accès distant expérimental aux archives du web au sein bibliothèques de l'Université de Lille. Si les enjeux juridiques et organisationnels sont encore nombreux, cette avancée permet déjà de rassembler une communauté de professionnels et de chercheurs autour des archives du web et dessiner ensemble, patiemment, les collections et les outils de demain.

> *Vive la République, vive la France et vive le sport[45].*

Pour suivre l'actualité des Archives du web à la BnF : @DLwebBnF

Notes

[1] Training Plus, Allocution du coach sportif de la République https://youtu.be/mWF9iI9XEd8

[2] Clémence Jost, Elles ont archivé le confinement : zoom sur 4 collectes d'archives pour l'Histoire. Archimag, 24/03/2021 pour la version en ligne. https://www.archimag.com/archives-patrimoine/2021/03/24/archives-confinement-4-collectes-histoire

[3] MUCEM, Vivre au temps du confinement.
https://www.mucem.org/vivre-au-temps-du-confinement-la-collection

[4] Guy Boyer, Recueillir les rêves du confinement. Connaissance des arts, 30/11/2020 pour la version en ligne.
https://www.connaissancedesarts.com/arts-expositions/art-contemporain/recueillir-les-reves-du-confinement-11149478/

[5] Elizabeth Serin, Hervé Mazurel, Arianna Cecconi, Tuia Cherici, Colloque « Rêves de confins ». Enregistrement vidéo. BnF, 22/05/2021. Consultable en ligne :
https://www.youtube.com/watch?v=xxGereF1kwY

[6] Médiamétrie. L'année internet 2020.
https://www.mediametrie.fr/fr/lannee-internet-2020

[7] Céline Boulay-Espéronnier, Cécile Cukierman, Stéphane Sautarel, 8 questions sur l'avenir du télétravail, vers une révolution un travail à distance ? Rapport d'information du Sénat n°89, enregistré le 22 octobre 2021.

[8] Géraldine Poels et Véronique Lefort, « Covid-19 dans les JT : un niveau de médiatisation inédit pour une pandémie », La Revue des médias, 1er octobre 2020. Les journaux télévisés sont ceux de TF1, France 2, France 3, France 5, Arte et M6.
https://larevuedesmedias.ina.fr/pandemie-covid-19-coronavirus-journal-televise

[9] Claude Grasland, « La pandémie de Covid-19 a bouleversé les contenus de la presse régionale », La Revue des médias, 20 août 2020.
https://larevuedesmedias.ina.fr/la-pandemie-de-covid-19-bouleverse-les-contenus-de-la-presse-regionale

[10] Médiamétrie. L'année internet 2020.
https://www.mediametrie.fr/fr/lannee-internet-2020

[11] Ce confinement créatif, qui par certains côtés relevait d'une forme d'injonction sociale, a pu également être considéré comme une reconnaissance acquise par la créativité ordinaire en un siècle. Voir : Jules Zimmermann, Fallait-il être créatif pour réussir son confinementâ□¯? Usbek et Rica, 25 mai 2020 [Retiré du site]. Consultable sur les archives de l'internet :
http://archivesinternet.bnf.fr/20200921123328/http://usbeketrica.com/fr/article/fallait-il-etre-creatif-pour-reussir-son-confinement

[12] Novel Coronavirus (2019-nCoV) outbreak : la collection internationale est accessible en ligne à l'adresse : https://archive-it.org/collections/13529

[13] Le terme « archiviste » est entendu ici au sens générique recouvrant différents statuts : archiviste et bibliothécaire.

[14] Alexandre Faye, Les archives web du coronavirus : une entreprise collective. Web

Corpora, 22 avril 2020. https://webcorpora.hypotheses.org/856

[15] Au sein de la BnF, chaque département organise sa propre collecte pour que des sites web jugés importants et relevant de son périmètre documentaire puissent être régulièrement archivés.

[16] Véronique Tranchant, Chantal Puech, Sophie Gebeil, Valérie Schafer et Alexandre Faye, Dans les coulisses de la collecte Covid-19. Entretien sur les pratiques des correspondants du DLWeb. Web Corpora, 16 novembre 2020. https://webcorpora.hypotheses.org/953

[17] Ange Aniesa, Ariane Bouchard. Constituer un réseau d'accès aux archives de l'internet : l'exemple français. IFLA WLIC 2017, IFLA, Aug 2017, Wroclaw, Pologne. ffhal-01703421 https://hal-bnf.archives-ouvertes.fr/hal-01703421/ Voir la liste : https://www.bnf.fr/fr/selection-partagee-et-acces-en-region-aux-archives-de-linternet

[18] 11 bibliothèques du réseau territorial ont accepté de participer à l'effort de collecte.

[19] La notion est utilisée ici pour désigner l'ensemble des canaux de communication numérique mobilisés par des candidats : sites web, réseaux sociaux, chaîne vidéo

[20] L'Association française pour le nommage Internet en coopération (AFNIC) est une association loi de 1901. Elle a pour mission de gérer les domaines Internet nationaux de premier niveau de la France comme le .fr ou le .re. L'agence est aussi prestataire le .bzh, le .alsace, le .corsica…etc. La coopération entre l'association et la BnF améliore la qualité des collectes.

[21] Les collectes projets couvrent des sujets d'intérêt définis au niveau international par le consortium IIPC : https://netpreserve.org/projects/collaborative-collections/. La BnF contribue à ces collectes collaboratives transnationales dont le résultat est accessible en ligne sur https://archive-it.org/explore?show=Collections. Elle mène et reconduit également ses propres collectes projets sur le sujet, souvent avec des sélections et des paramètres de collecte plus larges.

[22] Ces grandes catégories qui structurent la collecte ont été reprises de la collecte collaborative « Novel Coronavirus (2019-nCoV) outbreak » lancée par consortium International Internet Preservation Consortium (IIPC) en association avec Internet Archive. Il a été demandé aux correspondants de les utiliser. D'autres mots clés ont pu être ajoutés plus librement pour qualifier plus finement le contenu de la sélection.

[23] David Benoist, Alexandre Faye, Pascal Tanesie, Sophie Gebeil, Valérie Schafer. Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF). WARCnet Papers, 2020. hal-03066879 : https://hal.archives-ouvertes.fr/hal-03066879/

[24] À titre d'exemples : chu-lyon.fr/fr/covid-19 et mplusinfo.fr/tag/coronavirus (rubrique

du site M+ votre Mulhouse)

[25] Voir la carte mise à jour des déploiements en région et Outre-mer : https://umap.openstreetmap.fr/fr/map/ou-consulter-les-archives-de-linternet-de-la-bnf_73737#6/46.362/3.450

[26] L'ensemble des versions PDF des parcours guidés sont disponibles sur le site de la BnF : https://www.bnf.fr/fr/parcours-guides-archives-de-linternet

[27] Cette collection indexée est mise à jour chaque année.

[28] Toutefois, les fréquences de collecte qui ne sont pas proposées dans le cadre de la collecte « Actualité éphémère » furent écartées pour éviter de générer du bruit. Les fichiers WARC retenus correspondent aux crawls de fréquence « mensuelle », « hebdomadaire » et « plusieurs fois par jour » (paramètre de fréquence utilisé pour les réseaux sociaux) pour la période définie.

[29] Après un travail de vérification et nettoyage, la liste propose 4600 urls de départ utilisés pour les crawls menés entre le 1er février et le 31 juillet 2020. Elle est accessible au format CSV : https://api.bnf.fr/fr/node/176

[30] En visioconférence évidemment !

[31] Le BnF DataLab accueille des équipes de recherche porteuse de projets en analyse et traitement des collections numériques. https://www.bnf.fr/fr/actualites/bnf-data-lab-appel-projets-2021

[32] Le résultat des recherches à partir des index ne peut s'interpréter comme celui que donnerait une base de données structurée.

[33] L'outil a été déployé par la BNUS dans le cadre du projet Bodycapital pour indexer une collecte « Corpus de recherche » réalisé par la BnF pour les chercheurs. Le corpus porte sur les questions d'alimentation et de santé.

[34] Nicolas Truong et alter, Les penseurs de l'intime. Editions de l'Aube, 2021.

[35] Sur la question de la légitimité conférée au patrimoine numérique, voir : Emmanuelle Bermès, Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019). Histoire. Paris, Ecole nationale des chartes, 2020. NNT : 2020ENCP0001. HAL Id : tel-02475991, version 1 https://tel.archives-ouvertes.fr/tel-02475991/document

[36] Daniella Kostroun, Collecter les histoires de la Covid-19 pour les générations futures. In : Shigehisa Kuriyama, Ota de Leonardis, Carlos Sonnenschein, et Ibrahima Thioub. Covid-19. Tour du monde. Editions du Manucius, 2021. pp124-125.

[37] Voir la publication des WarcNet Papers - Exploring special web archives collections

related to COVID-19 : https://cc.au.dk/en/warcnet/warcnet-papers/

[38] Jérôme Thièvre, Boris Blanckemane, Valérie Schafer. Exploring special web archive collections related to COVID-19: The case of INA. WARCnet Papers, Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster and Michael Kurzmeier, 2020. https://cc.au.dk/fileadmin/user_upload/WARCnet/Schafer_et_al_Exploring_special_web_archives.pdf

[39] Sur les applications possibles à la recherche de ces données, on peut citer à titre d'exemple les travaux de Frédéric Clavert sur le centenaire de la Première guerre mondiale. Frédéric Clavert, Temporalités du Centenaire de la Grande Guerre sur Twitter. In : Valérie Schafer (sous la dir.), Temps et temporalités du web. Presse universitaire de Nanterre, 2018.

[40] Lucie Behr, Francisco Calderon, Eve Charras, Anaïs Crinière-Boizet, Arnaud Dhermy et Catherine Soulé-Sandic, Les collectes régionales du web. Web Corpora, 29 décembre 2021. https://webcorpora.hypotheses.org/1004

[41] Niels Brügger, A brief Outline of Temporalities of the Web. In : Valérie Schafer (sous la dir.), Temps et temporalités du web. Presse universitaire de Nanterre, 2018.

[42] Ian Milligan, You shouldn't need to be a web historian to use web archives. WARCnet Web Archive Studie, Youtube 2020. https://www.youtube.com/watch?v=Ki-QGu9AGiQ

[43] Sophie Gebeil, Website Story. INA, 2021

[44] https://www.collexpersee.eu/projet/respadon/ et sur Twitter @Respadon_Projet

[45] Training Plus, ibid.